Lecture Notes on Data Engineering and Communications Technologies 1

Fatos Xhafa
Leonard Barolli
Flora Amato *Editors*

Advances on P2P, Parallel, Grid, Cloud and Internet Computing

Proceedings of the 11th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC–2016) November 5–7, 2016, Soonchunhyang University, Asan, Korea

Springer

Lecture Notes on Data Engineering and Communications Technologies

Volume 1

Series editor

Fatos Xhafa, Technical University of Catalonia, Barcelona, Spain
e-mail: fatos@cs.upc.edu

The aim of the book series is to present cutting edge engineering approaches to data technologies and communications. It publishes latest advances on the engineering task of building and deploying distributed, scalable and reliable data infrastructures and communication systems.
The series has a prominent applied focus on data technologies and communications with aim to promote the bridging from fundamental research on data science and networking to data engineering and communications that lead to industry products, business knowledge and standardisation.

More information about this series at http://www.springer.com/series/15362

Fatos Xhafa · Leonard Barolli
Flora Amato
Editors

Advances on P2P, Parallel, Grid, Cloud and Internet Computing

Proceedings of the 11th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC-2016) November 5–7, 2016, Soonchunhyang University, Asan, Korea

Springer

Editors
Fatos Xhafa
Technical University of Catalonia
Barcelona
Spain

Flora Amato
University of Naples Federico II
Naples
Italy

Leonard Barolli
Fukuoka Institute of Technology
Fukuoka
Japan

ISSN 2367-4512 ISSN 2367-4520 (electronic)
Lecture Notes on Data Engineering and Communications Technologies
ISBN 978-3-319-49108-0 ISBN 978-3-319-49109-7 (eBook)
DOI 10.1007/978-3-319-49109-7

Library of Congress Control Number: 2016956191

© Springer International Publishing AG 2017
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Welcome Message from the 3PGCIC-2016 Organizing Committee

Welcome to the 11th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC-2016), which will be held in conjunction with BWCCA-2016 International Conference, November 5-7, 2016, Soonchunhyang (SCH) University, Asan, Korea.

P2P, Grid, Cloud and Internet computing technologies have been very fast established as breakthrough paradigms for solving complex problems by enabling large-scale aggregation and sharing of computational, data and other geographically distributed computational resources.

Grid Computing originated as a paradigm for high performance computing, as an alternative to expensive supercomputers. Since late 80's, Grid computing domain has been extended to embrace different forms of computing, including Semantic and Service-oriented Grid, Pervasive Grid, Data Grid, Enterprise Grid, Autonomic Grid, Knowledge and Economy Grid, etc.

P2P Computing appeared as the new paradigm after client-server and web-based computing. These systems are evolving beyond file sharing towards a platform for large scale distributed applications. P2P systems have as well inspired the emergence and development of social networking, B2B (Business to Business), B2C (Business to Consumer), B2G (Business to Government), B2E (Business to Employee), and so on.

Cloud Computing has been defined as a "computing paradigm where the boundaries of computing are determined by economic rationale rather than technical limits". Cloud computing is a multi-purpose paradigm that enables efficient management of data centres, timesharing, and virtualization of resources with a special emphasis on business model. Cloud Computing has fast become the computing paradigm with applications in all application domains and providing utility computing at large scale.

Finally, *Internet Computing* is the basis of any large-scale distributed computing paradigms; it has very fast developed into a vast area of flourishing field with enormous impact on today's information societies. Internet-based computing serves thus as a universal platform comprising a large variety of computing forms.

The aim of the 3PGCIC conference is to provide a research forum for presenting innovative research results, methods and development techniques from both theoretical and practical perspectives related to P2P, Grid, Cloud and Internet computing.

Many people have helped and worked hard to produce a successful 3PGCIC-2016 technical program and conference proceedings. First, we would like to thank all the authors for submitting their papers, the PC members, and the reviewers who carried out the most difficult work by carefully evaluating the submitted papers. Based on the reviewers' reports, the Program Committee selected 44 papers (29% acceptance rate) for presentation in the conference and publication in the Springer Lecture Notes on Data Engineering and Communication Technologies. The General Chairs of the conference would like to thank the PC Co-Chairs Flora Amato, University of Naples, Italy, Tomoki Yoshihisa, Osaka University, Japan, Jonghyuk Lee, Sangmyung University, Korea for their great efforts in organizing a successful conference and an interesting conference programme. We would like to appreciate the work of the Workshop Co-Chairs Xu An Wang, Engineering University of CAPF, China, Hyobum Ahn, Kongju University, Korea and Marek R. Ogiela, AGH, Poland for supporting the workshop organizers. Our appreciations also go to all workshops organizers for their hard work in successfully organizing these workshops.

We thank Shinji Sakamoto, Donald Elmazi and Yi Liu, FIT, Japan, for their excellent work and support with the Web Submission and Management System of conference.

We are grateful to Prof. Kyoil Suh, Soonchunhyang University, Korea and Prof. Makoto Takizawa, Hosei University, Japan, Honorary Co-Chairs for their support and encouragment.

Our special thanks to Prof. Nobuo Funabiki, Okayama University, Japan for delivering an inspiring keynote at the conference.

Finally, we would like to thank the Local Arrangement at Soonchunhyang University, for making excellent local arrangement for the conference.

We hope you will enjoy the conference and have a great time in Soonchunhyang University, Asan, Korea!

3PGCIC-2016 General Co-Chairs

Fatos Xhafa, Technical University of Catalonia, Spain
Leonard Barolli, Fukuoka Institute of Technology, Japan
Kangbin Yim, Soonchunhyang University, Korea

Message from the 3PGCIC-2016 Workshops Chairs

Welcome to the Workshops of the 10th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC 2016), held November 5-7, 2016, Soonchunhyang (SCH) University, Asan, Korea. The objective of the workshops was to present research results, work on progress and thus complement the main themes of 3PGCIC 2016 with specific topics of Grid, P2P, Cloud and Internet Computing.

The workshops cover research on Simulation and Modelling of Emergent Computational Systems, Multimedia, Web, Streaming Media Delivery, Middleware of Large Scale Distributed Systems, Network Convergence, Pervasive Computing and Distributed Systems and Security.

The held workshops are as following:

1. The 9th International Workshop on Simulation and Modelling of Emergent Computational Systems (SMECS-2016)
2. The 7th InternationalWorkshop on Streaming Media Delivery and Management Systems (SMDMS-2016)
3. The 6th International Workshop on Multimedia, Web and Virtual Reality Technologies and Applications (MWVRTA-2016)
4. The 4th International Workshop on Cloud and Distributed System Applications (CADSA-2016)
5. The 3rd International Workshop on Distributed Embedded Systems (DEM-2016)
6. International Workshop on Business Intelligence and Distributed Systems (BIDS-2016)
7. International Workshop on Signal Processing and Machine Learning (SiPML-2016)
8. International Workshop On Analytics & Awareness Learning Services (A2LS-2016)

We would like to thank all workshop organizers for their hard work in organizing these workshops and selecting high quality papers for presentation at workshops, the interesting programs and for the arrangements of the workshop during the conference days.

We hope you will enjoy the conference and have a great time in Asan, Korea!

3PGCIC-2016 Workshops Chairs

Xu An Wang, Engineering University of CAPF, China
Hyobum Ahn, Kongju University, Korea
Marek R. Ogiela, AGH, Poland

3PGCIC-2016 Organizing Committee

Honorary Chairs

Makoto Takizawa, Hosei University, Japan
Kyoil Suh, Soonchunhyang University, Korea

General Co-Chairs

Fatos Xhafa, Universitat Politècnica de Catalunya, Spain
Leonard Barolli, Fukuoka Institute of Technology, Japan
Kangbin Yim, Soonchunhyang University, Korea

Program Committee Co-Chairs

Flora Amato, University of Naples, Italy
Tomoki Yoshihisa, Osaka University, Japan
Jonghyuk Lee, Sangmyung University, Korea

Workshop Co-Chairs

Xu An Wang, Engineering University of CAPF, China
Hyobum Ahn, Kongju University, Korea
Marek R. Ogiela, AGH, Poland

Finance Chairs

Makoto Ikeda, Fukuoka Institute of Technology, Japan

Web Administrator Chairs

Shinji Sakamoto, Fukuoka Institute of Technology, Japan
Donald Elmazi, Fukuoka Institute of Technology, Japan
Yi Liu, Fukuoka Institute of Technology, Japan

Local Organizing Co-Chairs

Sunyoung Lee, Soonchunhyang University, Korea
Hwamin Lee, Soonchunhyang University, Korea
Yunyoung Nam, Soonchunhyang University, Korea

Track Areas

Data intensive computing, data mining, semantic web and information retrieval

Chairs:

Lars Braubach, Hamburg University, Germany
Giuseppe Di Fatta, University of Reading, UK

PC Members:

Costin Badica, University of Craiova, Romania
David Camacho, Universidad Autonoma de Madrid, Spain
Mario Cannataro, University Magna Græcia of Catanzaro, Italy
Mehmet Cudi Okur, Yasar University, Turkey
Giancarlo Fortino, University of Calabria, Italy
Sule Gunduz Oguducu, Istanbul Technical University, Turkey
Franziska Klügl, Örebro University, Sweden
Marco Lützenberger, DAI Labor Berlin, Germany
Mohamed Medhat Gaber, Robert Gordon University, UK
Paulo Novais, University of Minho, Portugal
Alexander Pokahr, University of Hamburg, Germany
Daniel Rodríguez, University of Alcalá, Spain
Domenico Talia, University of Calabria, Italy
Can Wang, Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australia
Ran Wolff, Yahoo Labs, Israel
Giuseppe Fenza, University of Salerno, Italy

Data Storage in Distributed Computation and Cloud Systems

Chairs:

Douglas Macedo, Federal University of Santa Catarina (UFSC), Brazil
Bogdan Nicolae, IBM, Ireland

Mario Dantas, Federal University of Santa Catarina, Brazil
Michael Bauer, University of Western Ontario, Canada
Rodrigo Righi, University of Vale do Rio dos Sinos, Brazil
Edward Moreno, Federal University of Sergipe, Brazil
Mathias Steinbauer, Johannes Kepler University, Austria
Diego Kreutz, University of Luxembourg, Luxembourg
Fabrizio Messina, University of Catania, Italy
Francieli Zanon, Federal University of Santa Catarina, Brazil
Marcio Castro, Federal University of Santa Catarina, Brazil

Secure Technology for Distributed Computation and Sensor Networks

Chairs:

Zheng Gong, South China Normal University, China
Giancarlo Fortino, University of Calabria, Italy

PC Members:

Weidong Qiu, Shanghai Jiaotong University
Jian Weng, Jinan University
Changshe Ma, South China Normal University
Shaohua Tang, South China University of Technology
Jin Li, Guangzhou University
Qiong Huang, South China Agriculture University

High Performance and Scalable Computing

Chairs:

Jose Luis Vazquez-Poletti, Universidad Complutense de Madrid, Spain
Corrado Santoro, University of Catania, Italy

PC Members:

Rafael Tolosana-Calasanz, Universidad de Zaragoza, Spain
Rui Han, Institute of Computing Technology, Chinese Academy of Sciences, China
Volodymyr Turchenko, Research Institute for Intelligent Computer Systems Ternopil National Economic University, Ukraine
Mario Cannataro, University of Catanzaro, Italy
Dana Petcu, West University of Timisoara, Romania
Mehdi Sheikhalishahi, CREATE-NET Research Center, Italy
Domenico Talia, Università della Calabria, Italy

Florin Pop, University Politehnica of Bucharest, Romania
Patrick Martin, Queen's University, Canada
Marcin Paprzycki, Polish Academia of Science, Poland
Maria Ganzha, Polish Academia of Science, Poland
Agostino Poggi, University of Parma, Italy
Costin Badica, University of Craiova, Romania
Fabrizio Messina, University of Catania, Italy
Giovanni Morana, C3DNA, Italy
Daniele Zito, C3DNA, Italy

Distributed Algorithms and Models for P2P, Grid, Cloud and Internet Computing

Chairs:

Osamu Tatebe, University of Tsukuba, Japan
Francesco Moscato, Second University of Naples, Italy

PC Members:

Sergio Di Martino, University of Naples "Federico II", Italy
Sara Romano, University of Naples "Federico II", Italy
Salvatore Cuomo, University of Naples "Federico II", Italy
Francesco Piccialli, University of Naples "Federico II", Italy
Crescenzo Diomiaiuta, Institute for High Performance Computing and Networking – National Research Council of Italy (ICAR-CNR), Italy
Fabio Persia Free University of Bozen, Italy
Daniela D'auria, Free University of Bozen, Italy

P2P, Ad-Hoc and Mobile Networking

Chairs:

Majed Haddad, Professor, University of Avignon, France
Jie Li, Professor, University of Tsukuba, Japan

Virtual Organizations and Federations, Cloud provisioning, management and programming

Chairs:

Salvatore Venticinque, Second University of Naples, Italy
Beniamino Di Martino, Second University of Naples, Italy

Bio-Inspired Computing and Pattern recognition

Chairs:

Lidia Ogiela, AGH University of Science and Technology, Poland
Ugo Fiore, Information Services Center of University of Naples - Federico II, Italy

PC Members:

Hoon Ko, J.E. Purkinje University, Czech Republic
Goreti Marreiros, ISEP/IPP, Portugal
Maria Victoria Moreno-Cano, University of Murcia, Spain
Giovanni Acampora, Nottingham Trent University, UK
Ji-Jian Chin, Multimedia University, Malaysia
Ki-tae Bae, Korean German Institute of Technology, Korea
Jongsun Choi, Soongsil University, Korea
Libor Mesicek, J.E. Purkinje University, Czech Republic

e-Health Technologies for Patient Monitoring

Chairs:

Massimo Esposito Institute for High Performance Computing and Networking – National
Research Council of Italy (ICAR-CNR)
George Tadros, Warwick University, UK

PC Members:

Marco Pota Institute for High Performance Computing and Networking – National Research Council of Italy (ICAR-CNR) IT, Italy
Aniello Minutolo Institute for High Performance Computing and Networking – National Research Council of Italy (ICAR-CNR), Italy
Antonio Picariello, Professor, University of Naples “Federico II”, Italy
Vincenzo Moscato, Professor, University of Naples “Federico II” , Italy
Giancarlo Sperlì, University of Naples “Federico II”, Italy
Giovanni Cozzolino, University of Naples “Federico II”, Italy

Adaptive Web-based, distributed and ubiquitous eLearning systems

Chairs:

Nicola Capuano, University of Salerno, Italy
Santi Caballé, Open University of Catalonia, Spain

PC Members:

Jordi Conesa, Open University of Catalonia, Spain
Thanasis Daradoumis, University of the Aegean, Greece
Darina Dicheva, Winston-Salem State University, USA
Michalis Feidakis, University of the Aegean, Greece
Angelo Gaeta, University of Salerno, Italy
David Gañán, Open University of Catalonia, Spain
Agathe Merceron, Beuth University of Applied Sciences, Germany
Jorge Miguel, Open University of Catalonia, Spain
Nestor Mora, Open University of Catalonia, Spain
Francesco Orciuoli, University of Salerno, Italy

Big Data, Data Management and Analytics

Chairs:

Corrado Aaron Visaggio, University of Sannio, Italy
Mario Piattini, Univeristy of Castilla-la-Mancha, Spain

PC Members:

Danilo Caivano, University of Bari, Italy
Nikolaos Georgantas, Inria, France
Lech Madeyski, Wroclaw University of Technology, Poland
Eric Medvet, University of Trieste, Italy
Francesco Mercaldo, University of Sannio, Italy
Jelena Milosevic, Università della Svizzera italiana, Switzerland
Ilia Petrov, Technische Universitat Darmstadt, Germany
Dimitris Sacharidis, Vienna University of Technology, Austria
Alberto Sillitti, Free University of Bozen, Italy
Juan carlos Trujillo, Alicante University, Spain
Giorgio Ventre, Università di Napoli Federico II, Italy
Martin Shepperd, Brunel University, UK

Next Generation Systems for Mobile and Cloud Computing

Chairs:

Yunfei Cao, CETC30, China
Yong Ding, Guilin University of Electronic Technology, China

PC Members:

Hui Li, Xidian University, China
Dianhua Tang, CETC30, China
Junwei Zhang, Xidian University, China
Jian Shen, Nanjing University of Information and Technology, China
Ximeng Liu, Singapore Management University, Singapore
Changlu Lin, Fujian Normal University, China
Jinbo xiong, Fujian Normal University, China

Heterogeneous High-Performance Architectures and Systems

Chairs:

Alessandro Cilardo, University of Naples "Federico II", Italy
José Flich, Polythechnic University of Valencia, Spain

PC Members:

Mario Barbareschi, University of Naples "Federico II", Italy
Rafael Tornero Gavila, Polythechnic University of Valencia, Spain
Edoardo Fusella, University of Naples "Federico II", Italy
Innocenzo Mungiello, Centro Regionale Information Communication Technology SCRL, Italy
Mirko Gagliardi, University of Naples "Federico II", Italy

Sustainable Computing

Chairs:

Ciprian Dobre, University Politehnica of Bucharest, Romania
Constandinos X. Mavromoustakis, University of Nicosia, Cyprus
Kuan-Ching Li, Providence University, Taiwan
Song Guo, University of Aizu, Japan

PC Members:

Nik Bessis, Edge Hill University, UK

Mauro Migliardi, University of Padua, Italy
Florin Pop, University Politehnica of Bucharest, Romania
Ioan Salomie, Technical University of Cluj-Napoca, Romania
George Suciu, BEIA Consult Int., Romania
Nicole Tapus, University Politehnica of Bucharest, Romania

Sergio L. Toral Marín, University of Seville, Spain
Radu Tudoran, European Research Center, Huawei Technologies Duesseldorf GmbH, Germany
Mario Donato Marino, Leeds Beckett University, UK
Athina Bourdena, University of Nicosia, Cyprus and Univ. of the Aegean, Greece
Joel Rodrigues, University of Beira Interior, Portugal
Periklis Chatzimisios, Alexander TEI of Thessaloniki, Greece
Muneer Masadeh Bani Yassein, University of Science and Technology, Jordan
Evangelos Pallis, Technological Educational Institute of Crete, Greece
Angelos K. Marnerides, Liverpool John Moores University, UK
Konstantinos Katzis, European University, Cyprus
Pedro Assuncao, Instituto Politecnico de Leiria/Instituto de Telecomunicações, Portugal
Evangelos Markakis, University of the Aegean, Greece
Carl Debono, University of Malta, Malta
Jordi Mongay Batalla, Warsaw, University of Technology, Poland
George Mastorakis, Technological Educational Institute of Crete, Greece
Nikolaos Zotos, University of the Aegean, Greece
Christos Politis, WMN Research Group, Kingston University London, UK
Peng Li, University of Aziu, Japan
Deze Zeng, China University of Geosciences, China
Lin Gu, Huazhong University of Science and Technology, China
Zhou Su, Waseda University, Japan
Fen Zhou, University of Avignon, France
Qiang Duan, The Pennsylvania State University, U.S.A

3PGCIC-2016 Reviewers

Aleksy Markus
Amato Flora
Barolli Admir
Barolli Leonard
Caballé Santi
Capuano Nicola
Castiglione Aniello
Chen Xiaofeng
Cristea Valentin
Cui Baojiang
Di Martino Beniamino
Dobre Ciprian
Doug Young Suh
Enokido Tomoya
Fenza Giuseppe
Ficco Massimo
Fiore Ugo
Fun Li Kin
Ganzha Maria
Gentile Antonio
Gotoh Yusuke
Hachaj Tomasz
He Debiao
Hellinckx Peter
Hussain Farookh
Hussain Omar
Ikeda Makoto
Kikuchi Hiroaki
Kolici Vladi
Koyama Akio
Kromer Pavel
Kulla Elis
Loia Vincenzo

Ma Kun
Matsuo Keita
Messina Fabrizio
Morana Giovanni
Kryvinska Natalia
Natwichai, Juggapong
Nishino Hiroaki
Ogiela Lidia
Ogiela Marek
Pichappan Pit
Palmieri Francesco
Paruchuri Vamsi Krishna
Platos Jan
Pop Florin
Rahayu Wenny
Rawat Danda
Rein Wen
Rodriguez Jorge
Santoro Corrado
Shibata Yoshitaka
Snasel Vaclav
Spaho Evjola
Suganuma Takuo
Sugita Kaoru
Takizawa Makoto
Tapus Nicolae
Terzo Olivier
Uchida Noriki
Venticinque Salvatore
Wang Xu An
Xhafa Fatos
Yim Kangbin
Yoshihisa Tomoki
Zhang Mingwu
Zomaya Albert

Welcome Message from the 9th SMECS-2016 Workshop Organizers

On the behalf of the organizing committee of 9th International Workshop on Simulation and Modelling of Engineering & Computational Systems, we would like to warmly welcome you for this workshop, which is held in conjunction with the 11th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC-2016) from November 5-7, 2016, Soonchunhyang (SCH) University, Asan, Korea.

Modelling and Simulation have become the de facto approach for studying the behaviour of complex engineering, enteprise information & communication systems before deployment in a real setting. The workshop is devoted to the advances in modelling and simulation techniques in fields of emergent computational systems in complex biological and engineering systems, and real life applications.

Modelling and simulation are greatly benefiting from the fast development in information technologies. The use of mathematical techniques in the development of computational analysis together with the ever greater computational processing power is making possible the simulation of very large complex dynamic systems. This workshop seeks relevant contributions to the modelling and simulation driven by computational technology.

The papers were reviewed and give a new insight into latest innovations in the different modelling and simulation techniques for emergent computational systems in computing, networking, engineering systems and real life applications. Special attention is paid to modelling techniques for information security, encryption, privacy, authentication, etc.

We hope that you will find the workshop as an interesting forum for discussion, research cooperation, contacts and valuable resource of new ideas for your research and academic activities.

Workshop Organizers

Fatos Xhafa, Technical University of Catalonia, Spain
Leonard Barolli, Fukuoka Institute of Technology, Japan

9th SMECS-2016 Program Committee

Workshop Organizers

Fatos Xhafa, Technical University of Catalonia, Spain
Leonard Barolli, Fukuoka Institute of Technology, Japan

PC Chair

Wang Xu An, Engineering University of CAPF, China

PC Members

Markus Aleksy, ABB, Germany
Xiaofeng Chen, Xidian University, China
Ciprian Dobre, University Politehnica of Bucharest, Romania
Antonio Gentile, University of Palermo, Italy
Makoto Ikeda, Fukuoka Institute of Technology, Japan
Kin Fun Li, University of Victoria, Canada
Hiroaki Nishino, University of Oita, Japan
Claudiu V. Suciu, Fukuoka Institute of Technology, Japan
Makoto Takizawa, Hosei University, Japan
Jin Li, Guangzhou University, China
Natalia Kryvinska, University of Vienna, Austria
Flora Amato, University of Naples "Federico II" , Italy

Welcome Message from the 7th SMDMS-2016 Workshop Organizers

It is my great pleasure to welcome you to the 2016 International Workshop on Streaming Media Delivery and Management Systems (SMDMS-2016). We hold this seventh workshop in conjunction with the 11th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC-2016), at Soonchunhyang University, Korea from November 5 to 7, 2016.

The tremendous advances in communication and computing technologies have created large academic and industry fields for streaming media. Streaming media have an interesting feature that the data stream continuously. They include many types of data like sensor data, video/audio data, stock data, and so on. It is obvious that with the accelerating trends towards streaming media, information and communication techniques will play an important role in future network. In order to accelerate this trend, further progresses of the researches on streaming media delivery and management systems are necessary. The aim of this workshop is to bring together practitioners and researchers from both academia and industry in order to have a forum for discussion and technical presentations on the current researches and future research directions related to this hot research area.

SMDMS-2016 contains high quality research papers. We selected 5 papers to be presented during the conference, which are divided in some sessions. I hope you find these sessions interesting, useful and enjoyable.

Many people contributed to the success of SMDMS-2016 organization. I would like to express my gratitude to the authors of the submitted papers for their excellent papers. I am very thankful to the program committee members who devoted their time for preparing and supporting the workshop. Without their help, this workshop would never be successful. A list of all of them is given in the program as well as the workshop website. I would like to also thank to 3PGCIC-2016 organizing committee members for their tremendous support for organizing.

Finally, I wish to thank all SMDMS-2016 attendees for supporting this workshop. I hope that you have a memorable experience you will never forget.

SMDMS-2016 International Workshop Chair

Tomoki Yoshihisa, Osaka University, Japan

7th SMDMS-2016 Program Committee

Workshop Chair

Tomoki Yoshihisa, Osaka University, Japan

International Liaison Chair

Lei Shu, Guangdong University of Petrochemical Technology, China

Program Committee Members

Akimitu Kanzaki, Shimane University, Japan
Hiroshi Yamamoto, Ritsumeikan University, Japan
Katsuhiro Naito, Aichi Institute of Technology, Japan
Kazuya Tsukamoto, Kyushu Institute of Technology, Japan
Mithun Mukherjee, Guangdong University of Petrochemical Technology, China
Shohei Yokoyama, Shizuoka University, Japan
NG Susumu Takeuchi, NTT Labsbr Takeshi Ishihara, Toshiba Corporation
Takeshi Usui, KDDI Labs, Japan
Toshiro Nunome, Nagoya Institute of Technology, Japan
Yasuo Ebara, Kyoto University, Japan
Yousuke Watanabe, Nagoya University, Japan
Yusuke Gotoh, Okayama University, Japan
Yusuke Hirota, Osaka University, Japan

Welcome Message from the 6th MWVRTA-2016 Workshop Organizers

Welcome to the 6th International Workshop on Multimedia, Web and Virtual Reality Technologies and Applications (MWVRTA 2016), which will be held in conjunction with the 11h International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC 2016), November 5-7, 2016, Soonchunhyang (SCH) University, Asan, Korea.

With appearance of multimedia, web and virtual reality technologies, different types of networks, paradigms and platforms of distributed computation are emerging as new forms of the computation in the new millennium. Among these paradigms and technologies, Web computing, multimodal communication, and tele-immersion software are most important. From the scientific perspective, one of the main targets behind these technologies and paradigms is to enable the solution of very complex problems such as e-Science problems that arise in different branches of science, engineering and industry. The aim of this workshop is to present innovative research and technologies as well methods and techniques related to new concept, service and application software in Emergent Computational Systems, Multimedia, Web and Virtual Reality. It provides a forum for sharing ideas and research work in all areas of multimedia technologies and applications.

It is impossible to organize an International Workshop without the help of many individuals. We would like to express our appreciation to the authors of the submitted papers, and to the program committee members, who provided timely and significant review.

We hope that all of you will enjoy MWVRTA 2016 and find this a productive opportunity to exchange ideas and research work with many researches.

MWVRTA 2016 Workshop Organizers

MWVRTA 2016 Workshop Co-Chairs

Leonard Barolli, Fukuoka Institute of Technology, Japan
Yoshitaka Shibata, Iwate Prefectural University, Japan

MWVRTA 2016 Workshop PC Chair

Kaoru Sugita, Fukuoka Institute of Technology, Japan

6th MWRTA-2016 Program Committee

Workshop Co-Chairs
Yoshitaka Shibata, Iwate Prefectural University, Japan
Leonard Barolli, Fukuoka Institute of Technology, Japan

Workshop PC Chair
Kaoru Sugita, Fukuoka Institute of Tehnology, Japan

Program Committee Members
Tetsuro Ogi, Keio University, Japan
Yasuo Ebara, Osaka University, Japan
Nobuyoshi Satou, Iwate Prefectural University, Japan
Makio ISHIHARA, Fukuoka Institute of Technology, Jaspan
Akihiro Miyakawa, Information Policy Division of Nanao-city, Ishikawa, Japan
Akio Koyama, Yamagata University, Japan
Keita Matsuo, Fukuoka Institute of Technology, Japan
Fatos Xhafa, Technical University of Catalonia, Spain
Vladi Kolici, Polytechnic University of Tirana, Albania
Joan Arnedo-Moreno, Open University of Catalonia, Spain
Hiroaki Nishino, Oita University, Japan
Farookh Hussain, Sydney University of Technology, Australia

Welcome Message from the 4th CADSA-2016 Workshop Organizers

Welcome to the Fourth International Workshop on Cloud and Distributed System Applications (CADSA), which is held in conjunction with the 11th Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC-2016), November 5-7, 2016, Soonchunhyang (SCH) University, Asan, Korea.

This International Workshop on Cloud and Distributed System Applications brings together scientists, engineers and students for sharing experiences, ideas, and research results about Domain Specific Applications relying on Cloud Computing or Distributed Systems.

This workshop provides an international forum for researchers and participants to share and exchange their experiences, discuss challenges and present original ideas in all aspects related to the Cloud and Distributed Systems applications design and development.

We have encouraged innovative contributions about Cloud and Distributed Computing, like:

- Distributed Computing Applications
- Cloud Computing Applications
- Collaborative Platforms
- Topologies for distributed Computing
- Semantic Technologies for Cloud
- Modeling and Simulation of Cloud Computing
- Modeling and Simulation of Distributed System
- Distributed Knowledge Management
- Distributed Computing for Smart Cities
- Distributed Computing for E-Health
- Quality Evaluation of Distributed Services

Many people contributed to the success of CADSA-2016. First, I would like to thank the Organizing Committee of 3PGCIC-2016 International Conference for giving us the opportunity to organize the workshop. Second, I would like to thank our Program Committee Members and, of course, I would like to thank all the Authors of the Workshop for submitting their research works and for their participation. Finally, I would like to thank the Local Arrangement Chairs for the 3PGCIC-2016 conference.

I hope you will enjoy CADSA workshop and 3PGCIC International Conference, find this a productive opportunity for sharing experiences, ideas, and research results with many researchers and have a great time in Asan, Korea.

CADSA-2016 Workshop Chair
Flora Amato, University of Naples "Federico II", Italy

4th CADSA-2016 Program Committee

Workshop Chair

Flora Amato, University of Naples "Federico II", Italy

Program Committee Members

Antonino Mazzeo, University of Naples "Federico II" , Italy
Nicola Mazzocca, University of Naples "Federico II", Italy
Carlo Sansone, University of Naples "Federico II", Italy
Beniamino di Martino, Second University of Naples, Italy
Antonio Picariello, University of Naples "Federico II", Italy
Valeria Vittorini, University of Naples "Federico II", Italy
Anna Rita Fasolino, University of Naples "Federico II", Italy
Umberto Villano, Università degli Studi del Sannio, Italy
Kami Makki, Lamar University, Beaumont (Texas), USA
Valentina Casola, University of Naples "Federico II", Italy
Stefano Marrone, Second University of Naples, Italy
Alessandro Cilardo, University of Naples "Federico II", Italy
Vincenzo Moscato, University of Naples "Federico II", Italy
Porfirio Tramontana, University of Naples "Federico II", Italy
Francesco Moscato, Second University of Naples, Italy
Salvatore Venticinque, Second University of Naples, Italy
Emanuela Marasco, West Virginia University, USA
Massimiliano Albanese, George Mason University, USA
Domenico Amalfitano, University of Naples "Federico II", Italy
Massimo Esposito, Institute for High Performance Computing and Networking (ICAR), Italy
Alessandra de Benedictis, University of Naples "Federico II", Italy
Roberto Nardone, University of Naples "Federico II", Italy
Mario Barbareschi, University of Naples "Federico II", Italy
Ermanno Battista, University of Naples "Federico II", Italy
Mario Sicuranza, Institute for High Performance Computing and Networking (ICAR), Italy
Natalia Kryvinska, University of Vienna, Austria
Moujabbir Mohammed, Université Hassan II Mohammedia-Casablanca, Maroco

Welcome Message from the 3rd DEM -2016 Workshop Organizers

Welcome to the Third International Workshop on Distributed Embedded systems (DEM-2016), which is held in conjunction with the 11th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC 2016), November 5-7, 2016, Soonchunhyang (SCH) University, Asan, Korea.

The tremendous advances in communication technologies and embedded systems have created an entirely new research field in both academia and industry for distributed embedded software development. This field introduces constrained systems into distributed software development. The implementation of limitations like real-time requirements, power limitations, memory constraints, etc. within a distributed environment require the introduction of new software development processes, software development techniques and software architectures. It is obvious that these new methodologies will play an key role in future networked embedded systems. In order to facilitate these processes, further progress of the research and engineering on distributed embedded systems is mandatory.

The international workshop on distributed embedded systems (DEM) aims to bring together practitioners and researchers from both academia and industry in order to have a forum for discussion and technical presentations on the current research and future research directions related to this hot scientific area. Topics include (but are not limited to) Virtualization on embedded systems, Model based embedded software development, Real-time in the cloud, Internet of Things, Distributed safety concepts, Embedded software for (mechatronics, automotive, healthcare, energy, telecom, etc.), Sensor fusion, Embedded multicore software, Distributed localisation, Distributed embedded software development and Testing.This workshop provides an international forum for researchers and participants to share and exchange their experiences, discuss challenges and present original ideas in all aspect of Distributed and/or Embedded Systems.

Many people contributed to the success of DEM-2016. First, I would like to thank the organising committee of the 3PGCIC 2016 International Conference for giving us the opportunity to organise the workshop. Second, I would like to thank our program committee members. And of course, I would like to thank all the authors of the workshop for submitting their research works and for their participation.

Finally, I would like to thank the Local Arrangement Chairs for the 3PGCIC conference.

I hope you will enjoy DEM workshop and the 3PGCIC International Conference, and have a great time in Asan, Korea.

DEM 2016 Workshop Chair
Peter Hellinckx, University of Antwerp, Belgium

3rd DEM-2016 Program Committee

Workshop chair

Peter Hellinckx, University of Antwerp

Program Committee Members

Paul Demeulenaere, University of Antwerp, Belgium
Marijn Temmerman, University of Antwerp, Belgium
Joachim Denil, McGill University, Canada
Maarten Weyn, University of Antwerp, Belgium

Welcome Message from the BIDS-2016 Workshop

Welcome to the 2016 International Workshop on Business Intelligence and Distributed Systems (BIDS-2016) at Soonchunhyang (SCH) University, Asan, Korea. BIDS-2016 is held in conjunction with the 11th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC-2016), November 5-7, 2016.

As many large-scale enterprise information systems start to utilize P2P networks, parallel, grid, cloud and Internet computing, they have become a major source of business information. Techniques and methodologies to extract quality information in distributed systems are of paramount importance for many applications and users in the business community. Data mining and knowledge discovery play key roles in many of today's prominent business intelligence applications to uncover relevant information of competitors, consumers, markets, and products, so that appropriate marketing and product development strategies can be devised. In addition, formal methods and architectural infrastructures for related issues in distributed systems, such as e-commerce and computer security, are being explored and investigated by many researchers.

The international BIDS workshop aims to bring together scientists, engineers, and practitioners to discuss, exchange ideas, and present their research findings on business intelligence applications, techniques and methodologies in distributed systems. We are pleased to have four high-quality papers selected for presentation at the workshop and publication in the proceedings.

We would like to express our sincere gratitude to the members of the Program Committee for their efforts and the 11th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing for co-hosting BIDS-2016. Most importantly, we thank all the authors for their submission and contribution to the workshop.

BIDS-2016 International Workshop Co-Chairs

Markus Aleksy, ABB Corporate Research, Germany
Kin Fun Li, University of Victoria, Canada

BIDS-2016 Program Committee

Workshop Co- Chairs

Kin Fun Li, University of Victoria, Canaday
Doug Young Suh, Kyunghee University, Korea

Welcome Message from the SiPML-2016 Workshop Organizers

Welcome to the International Workshop on Signal Processing and Machine Learning (SiPML-2016), which is held in conjunction with the 11th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC 2016), November 5-7, 2016, Soonchunhyang (SCH) University, Asan, Korea.

The workshop brings together engineers, students, practitioners, and researchers from the fields of machine learning (ML) and signal processing (SP). The aim of the workshop is to contribute to the cross-fertilization between the research on ML methods and their application to SP to initiate collaboration between these areas. ML usually plays an important role in the transition from data storage to decision systems based on large databases of signals such as the obtained from sensor networks, internet services, or communication systems. These systems imply developing both computational solutions and novel models. Signals from real-world systems are usually complex such as speech, music, bio-medical, multimedia, among others. Thus, SP techniques are very useful for these type of systems to automate processing and analysis techniques to retrieve information from data storage. Topics of the workshop range from foundations for real-world systems, and processing, such as speech, language analysis, biomedicine, convergence and complexity analysis, machine learning, social networks, sparse representations, visual analytics, robust statistical methods.

We would like to thank the 3PGCIC 2016 International Conference for giving us the opportunity to organise the workshop,. Second, the PC membres of the workshop and the authors of the workshop for submitting their research works and for their participation.

We wish you enjoy the workshop at the 3PGCIC-2016 Conference and have a pleasant stay in Asan, Korea.

SiMPL-2016 International Workshop Co-Chairs

Ricardo Rodriguez Jorge (Autonomous University of Ciudad Juarez, Mexico)
Osslan Osiris Vergara (Autonomous University of Ciudad Juarez, Mexico)
Vianey Gpe. Cruz (Autonomous University of Ciudad Juarez, Mexico)

SiPML-2016 Program Committee

Workshop Chair

Ricardo Rodriguez Jorge, Autonomous University of Ciudad Juarez, Mexico
Osslan Osiris Vergara, Autonomous University of Ciudad Juarez, Mexico
Vianey Gpe. Cruz, Autonomous University of Ciudad Juarez, Mexico

Program Committee

Ezendu Ariwa, University of Bedfordshire, United Kindom
Humberto Ochoa, Autonomous University of Ciudad Juarez ,Mexico
Jiri Bila, Czech Technical University in Prague, Czech Republic
Ke Liao, University of Kansas, USA
Mohamed Elgendi, University of British Columbia, Canada
Nghien N. B., Hanoi University of Industry, Vietnam
Pit Pichappan, Al Imam University, Saudi Arabia
Vicente Garcia, Autonomous University of Ciudad Juarez, Mexico
Yao-Liang Chung, National Taipei University, Taiwan

Welcome Message from the A2LS-2016 Workshop Organizers

Welcome to the First International Workshop On Analytics & Awareness Learning Services (A2LS-2016), which is held in conjunction with the 11th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC 2016), November 5-7, 2016, Soonchunhyang (SCH) University, Asan, Korea.

Data Analysis is a cornerstone of online learning environments. Since the first conception of e-learning and collaborative systems to support learning and teaching, data analysis has been employed to support learners, teachers, researchers, managers and policy makers with useful information on learning activities and learning design. While data analysis originally employed mainly statistical techniques due to the modest amounts and varieties of data being gathered, with the rapid development of internet technologies and increasingly sophisticated online learning environments, increasing volumes and varieties of data are being generated and data analysis has moved to more complex analysis techniques such as educational data mining and, most recently, learning analytics. Now powered by cloud technologies, online learning environments are capable of gathering and storing massive amounts of data of various formats, and tracking user-system and user-user interactions as well as rich contextual information in such systems. This has led to the need to address the definition, modelling, development and deployment of sophisticated learning services offering analytics and context awareness information to all participants and stakeholders in online learning. This workshop seeks original research contributions in analytics and context awareness in learning systems, driven by service-based architectures and cloud technologies.

A2LS-2016 International Workshop Organizers

Santi Caballé, Open University of Catalonia, Spain
Jordi Conesa, Open University of Catalonia, Spain

A2LS-2016 Program Committee

Workshop Organizers

Santi Caballé, Open University of Catalonia, Spain
Jordi Conesa, Open University of Catalonia, Spain

PC Members

Nicola Capuano, University of Salerno, Italy
Francesco Palmieri, University of Salerno, Italy
Miguel Bote, University of Valladolid, Spain
Angel Hernández, Technical University of Madrid, Spain
Stavros Demetriadis, Aristotle University of Thessaloniki, Greece
Thrasyvoulos Tsiatsos, Aristotle University of Thessaloniki, Greece
Fatos Xhafa, Technical University of Catalonia, Spain
Isabel Guitart, Open University of Catalonia, Spain
Néstor Mora, Open University of Catalonia, Spain
David Gañán, Open University of Catalonia, Spain
Jorge Miguel, University San Jorge, Spain

3PGCIC-2016 Keynote Talk

Prof. Nobuo Funabiki, Okayama University, Japan

Java Programming Learning Assistant System: JPLAS

Abstract

As a useful and practical object-oriented programming language, Java has been used in many practical systems including enterprise servers, smart phones, and embedded systems, due to its high safety and portability. Then, a lot of educational institutes have offered Java programming courses to foster Java engineers. We have proposed and implemented a Web-based Java Programming Learning Assistant System called JPLAS, to assist such Java programming educations. JPLAS supports three types of problems that have different difficulties to cover a variety of students: 1) element fill-in-blank problem, 2) statement fill-in-blank problem, and 3) code writing problem. For 1), we have proposed a graph-theory based algorithm to automatically generate element fill-in-blank problems that have unique correct answers. For 2) and 3), we have adopted the test-driven development (TDD) method so that the answer codes from students can be automatically verified using test codes for their self-studies. In this talk, we introduce outlines of JPLAS and its application results to the Java programming course in our department. Besides, we introduce some new features of JPLAS including the offline answering function and the coding rule learning function.

Contents

Part I 11th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC-2016)

A Configurable Shared Scratchpad Memory for GPU-like Processors 3
Alessandro Cilardo, Mirko Gagliardi and Ciro Donnarumma

Research of Double Threshold Collaborative Spectrum Sensing Based on OFDM 15
Ruilian Tan

Research on particle swarm optimization of variable parameter 25
Zhe Li, Ruilian Tan and Baoxiang Ren

An Access Control Architecture for Protecting Health Information Systems 35
Angelo Esposito, Mario Sicuranza and Mario Ciampi

Intelligent Management System for Small Gardens Based on Wireless Sensor Network 49
Xiao-hui Zeng, Man-sheng Long, Qing Liu, Xu-an Wang and Wen-lang Luo

An AHP Based Study of Coal-Mine Zero Harm Safety Culture Evaluation 57
Hongxia Li, Hongxi Di and Xu An Wang

Analysis of Interval-Valued Reliability of Multi-State System in Consideration of Epistemic Uncertainty 69
Gang Pan, Chao-xuan Shang, Yu-ying Liang, Jin-yan Cai and Dan-yang Li

Toward Construction of Efficient Privacy Preserving Reusable Garbled Circuits 81
Xu An Wang

A Heuristically Optimized Partitioning Strategy on Elias-Fano Index 93
Xingshen Song, Kun Jiang and Yuexiang Yang

Smart Underground: Enhancing Cultural Heritage Information Access and Management through Proximity-Based Interaction 105
Giuseppe Caggianese and Luigi Gallo

Ciphertext-Policy Attribute Based Encryption with Large Attribute Universe 115
Siyu Xiao, Aijun Ge, Fushan Wei and Chuangui Ma

Asymmetric Searchable Encryption from Inner Product Encryption 123
Siyu Xiao, Aijun Ge, Jie Zhang, Chuangui Ma and Xu'an Wang

Design of a Reconfigurable Parallel Nonlinear Boolean Function Targeted at Stream Cipher 133
Su Yang

Temporally Adaptive Co-operation Schemes 145
Jakub Nalepa and Miroslaw Blocho

Discovering Syndrome Regularities in Traditional Chinese Medicine Clinical by Topic Model 157
Jialin Ma and Zhijian Wang

Fuzzy on FHIR: a Decision Support service for Healthcare Applications . . 163
Aniello Minutolo, Massimo Esposito and Giuseppe De Pietro

Electric Mobility in Green and Smart Cities 173
Adrian-Gabriel Morosan, Florin Pop and Aurel-Florin Arion

SR-KNN: An Real-time Approach of Processing *k*-NN Queries over Moving Objects 185
Ziqiang Yu, Yuehui Chen and Kun Ma

Intrusion Detection for WSN Based on Kernel Fisher Discriminant and SVM 197
Zhipeng Hu, Jing Zhang and Xu An Wang

Automatic Verification of Security of OpenID Connect Protocol with ProVerif 209
Jintian Lu, Jinli Zhang, Jing Li, Zhongyu Wan and Bo Meng

Low Power Computing and Communication System for Critical Environments 221
Luca Pilosu, Lorenzo Mossucca, Alberto Scionti, Giorgio Giordanengo, Flavio Renga, Pietro Ruiu, Olivier Terzo, Simone Ciccia and Giuseppe Vecchi

Risk Management Framework to Avoid SLA Violation in Cloud from a Provider's Perspective 233
Walayat Hussain, Farookh Khadeer Hussain and Omar Khadeer Hussain

Enhancing Video Streaming Services by Studying Distance Impact on the QoS in Cloud Computing Environments 243
Amirah Alomari and Heba Kurdi

Application of Personalized Cryptography in Cloud Environment 253
Marek R. Ogiela and Lidia Ogiela

Optimizing Machine Learning based Large Scale Android Malwares Detection by Feature Disposal 259
Lingling Zeng, Min Lei, Wei Ren and Shiqi Deng

Research on Decisive Mechanism of Internet Financial Interest Rate 269
Shengdong Mu, Yixiang Tian, Li Li and Xu An Wang

Toward Construction of Encryption with Decryption Awareness Ability for Cloud Storage 281
Xu An Wang, Fatos Xhafa, Guang Ming Wu and Wei Wang

Elastipipe: On Providing Cloud Elasticity for Pipeline-structured Applications 293
Rodrigo da Rosa Righi, Mateus Aubin, Cristiano André da Costa, Antonio Marcos Alberti and Arismar Cerqueira Sodre

Semantic Summarization of News from Heterogeneous Sources 305
Flora Amato, Antonio d'Acierno, Francesco Colace, Vinenzo Moscato, Antonio Penta and Antonio Picariello

Self Planning in Critical Multi-Agent Systems 315
Flora Amato, Nicola Mazzocca and Francesco Moscato

Towards a coaching system for daily living activities: the use of kitchen objects and devices for cognitive impaired people 325
Alba Amato, Antonio Coronato and Giovanni Paragliola

Wellness & LifeStyle Server: a Platform for Anthropometric and LifeStyle Data Analysis 337
Giovanna Sannino, Alessio Graziani, Giuseppe De Pietro and Roberto Pratola

Semantic Information Retrieval from Patient Summaries 349
Mario Sicuranza, Angelo Esposito and Mario Ciampi

Adaptive Modular Mapping to Reduce Shared Memory Bank Conflicts on GPUs 361
Innocenzo Mungiello and Francesco De Rosa

A Review on Data Cleaning Technology for RFID Network 373
He Xu, Jie Ding, Peng Li and Wei Li

An Efficient RFID Reader Network Planning Strategy Based on P2P Computing 383
He Xu, Weiwei Shen, Peng Li and Cong Qian

Energy Optimization Algorithm of Wireless Sensor Networks based on LEACH-B 391
Peng Li, Wanyuan Jiang, He Xu and Wei Liu

On the Security of a Cloud-Based Revocable IBPRE Scheme for Data Sharing 405
Jindan Zhang and Baocang Wang

Semantic Integration and Correlation of Digital Evidences in Forensic Investigations 415
Flora Amato, Giovanni Cozzolino and Nicola Mazzocca

Efficient Batch and Online Kernel Ridge Regression for Green Clouds 425
Bo-Wei Chen, Seungmin Rho and Naveen Chilamkurti

Queuing-Oriented Job Optimizing Scheduling In Cloud Mapreduce 435
Ting-Qin He, Li-Jun Cai, Zi-Yun Deng, Tao Meng and Xu An Wang

A Novel Cloud Computing Architecture Oriented Internet of Vehicles 447
He Xu, Ye Ding, Peng Li and Ruchuan Wang

Design and implementation of Data Center Network Global Congestion Price Calculating method based on Software Defined Networking 459
Chen Xiao Long and Shan Chun

Business Process Mining based Insider Threat Detection System 467
Taiming Zhu, Yuanbo Guo, Jun Ma and Ankang Ju

Part II Workshop SMECS-2016: 9th International Workshop on Simulation and Modelling of Engineering & Computational Systems

Design of Intermediate Frequency digital processing module based on AD9269 481
Li Mei

Research on Adaptive Cooperative Spectrum Sensing 487
Ruilian Tan

Research on Directional AF Cooperative Communication System Based on Outage Probability 497
Zhe Li, Ruilian Tan, Kai Shi and Baoxiang Ren

Research on Efficient Fibre-Channel-based Token-Routing Switch-Network Communication Protocol 507
Wen-lang Luo, Jing-xiang Lv, Dong-sheng Liu and Xiao-hui Zeng

Path Optimization of Coal Mining Machine Based on Virtual Coal-Rock Interface 517
Dong Gang, Nie Zhen and Wang Peijun

Test Sequencing Optimization for a Kind of Radar Receiver 525
Liu Xiaopan, Cai Jinyan, Wu Shihao and Li Danyang

A Method for the Detection of Singularity in Pulse 533
Wei Qi, Hongwei Zhuang and Liqiong Zhang

Design of non-lethal kinetic weapon impact method based on hybrid-III dummy 543
Wei Qi, Hongwei Zhuang and Fadong Zhao

Review and Research Development of Optimization Algorithms Based on Wolf Intelligence 555
Sun Yixiao, Zhan Renjun, Wu Husheng, Han Zhexin and Ma Yanbin

SFCS-Broker: Supporting Feedback Credibility Brokering Service for Multiple Clouds Environments 567
Heba Kurdi, Sara Alfaraj and Asma Alatawi

Part III Workshop SMDMS-2016:The 7th International Workshop on Streaming Media Delivery and Management Systems

QoE Assessment of Multi-View Video and Audio Transmission with a Packet Discarding Method over Bandwidth Guaranteed IP Networks 579
Keita Furukawa and Toshiro Nunome

Evaluation of Division-based Broadcasting System over Wireless LAN 589
Yusuke Gotoh and Yuki Takagi

High Quality Multi-path Multi-view Video Transmission considering Path Priority Control 599
Tetta Ishida, Takahito Kito, Iori Otomo, Takuya Fujihashi, Yusuke Hirota and Takashi Watanabe

A Recursive Continuous Query Language for Integration of Streams and Graphs 609
Yousuke Watanabe

A Design of a Video Effect Process Allocation Scheme for Internet Live Broadcasting 619
Tomoya Kawakami, Yoshimasa Ishi, Satoru Matsumoto, Tomoki Yoshihisa and Yuuichi Teranishi

Part IV Workshop MWVRTA-2016: The 6th International Workshop on Multimedia, Web and Virtual Reality Technologies

Numerical Simulation of Impact Acceleration on the Key Parts of Packing Ammunition on Condition of Dropping 631
Xue Qing

Instruction Sequences Clustering and Analysis of Network Protocol's Dormant Behaviors 639
Yan-Jing Hu

Consideration of Educational Multimedia Contents Introducing Multimedia Switching Function 651
Kaoru Sugita

Mobile Spam Filtering base on BTM Topic Model 657
Jialin Ma, Yongjun Zhang and Lin Zhang

Construction of an Electronic Health Record System for supporting a Zoo Veterinarian 667
Tatsuya Oyanagi, Misaki Iyobe, Tomoyuki Ishida, Noriki Uchida, Kaoru Sugita and Yoshitaka Shibata

Part V Workshop CADSA-2016: 4th International Workshop on Cloud and Distributed System Applications

An Architecture for processing of Heterogeneous Sources 679
Flora Amato, Giovanni Cozzolino, Antonino Mazzeo and Sara Romano

Modeling Approach for Specialist Domain 689
Flora Amato, Giovanni Cozzolino, Francesco Moscato, Vincenzo Moscato and Antonio Picariello

Designing a Service Oriented System for social analytics 699
Angelo Chianese, Paolo Benedusi and Francesco Piccialli

GPU Profiling of Singular Value Decomposition in OLPCA Method for Image Denoising 707
Salvatore Cuomo, Pasquale De Michele, Francesco Maiorano and Livia Marcellino

A machine learning approach for predictive maintenance for mobile phones service providers 717
Anna Corazza, Francesco Isgrò, Luca Longobardo and R. Prevete

A Cloud-based Approach for Analyzing Viral Propagation of Linguistic Deviations by Social Networking: Current Challenges and Pitfalls for Text Analysis Tools 727
Fiammetta Marulli, Alessandro Nardaggio, Adriano Racioppi and Luca Vallifuoco

A Logic-based Clustering Approach for Cooperative Traffic Control Systems 737
Walter Balzano, Maria Rosaria Del Sorbo, Aniello Murano and Silvia Stranieri

Chain-of-Trust for Microcontrollers using SRAM PUFs: the Linux Case Study 747
Domenico Amelino, Mario Barbareschi and Antonino Mazzeo

Part VI Workshop DEM-2016: 3rd International Workshop on Distributed Embedded Systems

Hybrid Approach on Cache Aware Real-Time Scheduling for Multi-Core Systems 759
Thomas Huybrechts, Yorick De Bock, Haoxuan Li and Peter Hellinckx

Towards Recognising Activities of Daily Living through Indoor Localisation Techniques 769
Andriy Zubaliy, Anna Hristoskova, Nicolás González-Deleito and Elena Tsiporkova

Powerwindow: a Multi-component TACLeBench Benchmark for Timing Analysis 779
Haoxuan Li, Paul De Meulenaere and Peter Hellinckx

Predictive Algorithms for Scheduling Parameter Sweep Calculations in a Cloud Environment 789
Stig Bosmans, Glenn Maricaux and Filip Van der Schueren

Part VII Workshop BIDS-2016: International Workshop on Business Intelligence and Distributed Systems

A Study of the Internet Financial Interest Rate Risk Evaluation Index System in Context of Cloud Computing 801
Mu Sheng-dong, Tian Yi-xiang, Lili and Xu An Wang

A Dynamic System development Method for Startups Migrate to C loud 813
Asmaa Abdelrehim Selim Ibrahim and Mohameed Mamdouh Awny

Integrated Algorithms for Network Optimization of Dangerous Goods Transportation 825
Haixing Wang, Guiping Xiao and Tao Hai

Study on Dynamic Monitoring System for Railway Gaseous Tank Cars Transportation 835
Haixing Wang, Jujian Wang and Chao Li

Improvement of Study Logging System for Active Learning Using Smartphone 845
Noriyasu Yamamoto

Identifying Prime Customers Based on MobileUsage Patterns 853
Deepali Arora and Kin Fun Li

NetFlow: Network Monitoring and Intelligence Gathering 863
Vivek Ratan and Kin Fun Li

Company Web Site Analysis Based on Visual Information 869
Kosuke Takono and Yurika Moriwaki

A Study of the Standardization of Fire Forces Equipment Procurement . . 877
Zhang Jie

Part VIII Workshop-SiPML 2016: Workshop on Signal Processing and Machine Learning

Monitoring of Cardiac Arrhythmia Patterns by Adaptive Analysis 885
José Elías Cancino Herrera, Ricardo Rodríguez Jorge, Osslan Osiris Vergara Villegas, Vianey Guadalupe Cruz Sánchez, Jiri Bila, Manuel de Jesús Nandayapa Alfaro, Israel U. Ponce, Ángel Israel Soto Marrufo and Ángel Flores Abad

Pakistan Sign Language Recognition and Translation System using Leap Motion Device 895
Nosheela Raziq and Seemab Latif

Identifying stable objects for accelerating the classification phase of k-means 903
A. Mexicano, S. Cervantes, R. Rodríguez, J. Pérez, N. Almanza, M.A. Jiménez and A. Azuara

Expert System To *Engage* CHAEA Learning Styles, ACRA Learning Strategies and Learning Objects into an E-Learning Platform for Higher Education Students 913
José Angel Montes Olguín, Francisco Javier Carrillo García, Ma. de la Luz Carrillo González, Antonia Mireles Medina and Julio Zenón García Cortés

Part IX Workshop-A2LS-2016: International Workshop On Analytics & Awareness Learning Services

Security Analysis of OpenID Connect Protocol with Cryptoverif in the Computational Model 925
Jinli Zhang, Jintian Lu, Zhongyu Wan, Jing Li and Bo Meng

Towards a Particular Prediction System to Evaluate Student's Success 935
David Baneres

Evaluation of an eLearning Platform Featuring Learning Analytics and Gamification 947
David Gañán, Santi Caballé, Robert Clarisó and Jordi Conesa

MobilePeerDroid: A Platform for Sharing, Controlling and Coordination in Mobile Android Teams 961
Fatos Xhafa, Daniel Palou, Santi Caballé, Keita Matsuo and Leonard Barolli

Author Index 973

Part I
11th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC-2016)

A Configurable Shared Scratchpad Memory for GPU-like Processors

Alessandro Cilardo, Mirko Gagliardi, Ciro Donnarumma

Abstract During the last years Field Programmable Gate Arrays and Graphics Processing Units have become increasingly important for high-performance computing. In particular, a number of industrial solutions and academic projects are proposing design frameworks based on FPGA-implemented GPU-like compute units. Existing GPU-like core projects provide limited hardware support for shared scratchpad memory and particularly for the problem of bank conflicts, a major source of performance loss with many parallel kernels. In this paper, we present a configurable, GPU-like oriented scratchpad memory with built-in support for bank remapping. The core is fully synthetizable on FPGA with a contained hardware cost. We also validated the presented architecture with a cycle-accurate event-driven emulator written in C++ as well as an RTL simulator tool. Last, we demonstrated the impact of bank remapping and other parameters available with the proposed configurable shared scratchpad memory by evaluating the performance of two real-world parallelized kernels.

1 Introduction

Current trends in high-performance computing (HPC) are increasingly moving towards heterogeneous platforms [25], i.e. systems made of different computational units, like general-purpose CPUs, digital signal processors (DSPs), graphics pro-

Alessandro Cilardo
University of Naples Federico II and Centro Regionale ICT (CeRICT), Naples, Italy, e-mail: acilardo@unina.it

Mirko Gagliardi
University of Naples Federico II and Centro Regionale ICT (CeRICT), Naples, Italy, e-mail: mirko.gagliardi@unina.it

Ciro Donnarumma
University of Naples Federico II, Naples, Italy

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_1

cessing units (GPUs), co-processors, and custom acceleration logic, enabling significant benefits in terms of both power and performance.

While HPC covers today disparate applications [27, 7, 6], historically it has never extensively relied on FPGAs, mostly because of the reduced support for floating-point arithmetic. On the other hand, FPGAs and special-purpose hardware in general, e.g. used for arithmetic operations requiring specialized circuit solutions in various areas [11, 10, 15, 14], provide a huge potential for improved power efficiency compared to software-programmable platforms.

Furthermore, while numerous approaches exist for raising somewhat the level of abstraction for hardware design [18, 16], developing an FPGA-based hardware accelerator is still challenging as seen from a software programmer. Consequently, high-performance platforms mostly rely on general-purpose compute units such as CPUs and/or GPUs. However, pure general-purpose hardware is affected by inherently limited power-efficiency, i.e., low GFLOPS-per-Watt. Architectural customization can play here a key role, as it enables unprecedented levels of power-efficiency compared to CPUs/GPUs. This is the essential reason while very recent trends are putting more emphasis on the potential role of FPGAs.

In fact, recent FPGA families, such as the Xilinx Virtex-7 or the Altera Stratix 5, have innovative features, providing significantly reduced power, high speed, lower cost, and reconfigurability [24]. Due to these changes, in the very recent years many innovative companies, including Convey, Maxeler, SRC, Nimbix [25], have introduced FPGA-based heterogeneous platforms used in a large range of HPC applications, e.g. multimedia, bioinformatics, security-related processing, etc. [27, 25], with speedups in the range of 10x to 100x.

This paper explores the adoption of a deeply customizable scratchpad memory system for FPGA-oriented accelerator designs. At the heart of the proposed architecture is a multi-bank parallel access memory system for GPU-like processors. The proposed architecture enables a dynamic bank remapping hardware mechanism, allowing data to be redistributed across banks according to the specific access pattern of the kernel being executed, miminizing the number of conflicts and thereby improving the ultimate performance of the accelerated application. In particular, relying on an advanced configurable crossbar, on a hardware-supported remapping mechanism, and extensive parameterization, the proposed architecture can enable highly parallel accesses matching the potential of current HPC-oriented FPGA technologies. The paper describes the main insights behind by dynamic bank remapping as well as the key role that scratchpad memory might play for hardware-accelerated computing applications.

2 Related work

FPGAs have been used in a vast range of applications [5, 9], although the need for floating point operations has delayed their adoption in HPC. Recently Altera and Xilinx, the two prominent FPGA manufacturers, focused on overcoming FPGA

floating-point limitations. In particular, Altera, now part of Intel Corporation, has developed a new floating-point technology (called Fused Datapath) and toolkit (DSP Builder) intended to achieve maximum performance in floating-point design implementing on Altera FPGAs [3]. The other historical problem with FPGAs is programmability. Designing a complex architecture on FPGA, as mentioned above, requires a highly-skilled hardware designer. To overcome this limitation, Altera and Xilinx are bringing the GPU programming model to the FPGA domain. The Altera SDK for OpenCL [1] makes FPGAs accessible to non-expert users. This toolkit allows a user to abstract away the traditional hardware FPGA development flow, effectively creating custom hardware on FPGAs for each instruction being accelerated. Altera claims that this SDK provides much more power-efficient use of the hardware than a traditional CPU or GPU architecture. On other hand, similar to the Altera SDK for OpenCL, Xilinx SDAccel [30], enables traditional CPU and GPU developers to easily migrate their applications to FPGAs while maintaining and reusing their OpenCL, C, and C++ code. Driven by these innovations, FPGAs are becoming increasingly attractive for HPC applications, offering a fine grained parallelism and low power consumption compared to other accelerators.

In line with the above trends, academic and industrial research is focusing on GPU-like paradigms to introduce some form of programmability in FPGA design. In the last years, a few GPU-like projects have appeared. Kingyens and Steffan [23] propose a softcore architecture inspired by graphics processing units (GPUs) mostly oriented to FPGAs. The architecture supports multithreading, vector operations, and can handle up to 256 concurrent threads. Nyami/Nyuzi [12] is a GPU-like RISC architecture inspired by Intel Larrabee. The Nyami HDL code is fully parameterizable and provides a flexible framework for exploring architectural tradeoffs. The Nyami project provides an LLVM-based C/C++ compiler and can be synthesized on FPGA. Guppy [4] (GPU-like cUstomizable Parallel Processor prototYpe) is based on the LEON3 parameterizable soft core. Guppy main feature is to support CUDA-like threads in a lock-step manner to emulate the CUDA execution model. MIAOW [8] (Many-core Integrated Accelerator Of Wisconsin) provides an open-source RTL implementation of the AMD Southern Islands GPGPU ISA. MIAOW's main goal is to be flexible and to support OpenCL-based applications.

Data movement and memory access has traditionally been an important optimization problem and in many classes of systems it may significantly impact perfomance, along with the interconnection subsystem [21, 22]. This also applies for GPU-like processors. Cache hierarchy has been the traditional way to alleviate the memory bottleneck. However, cache coherence mechanisms are complex and not needed in some applications. Many modern parallel architectures utilize fast non-coherent user-managed on-chip memories, called *scratchpad memories* (SPM). Since NVIDIA Fermi family [2], GPUs are equipped with this kind of memories. In NVIDIA architectures this memory can be used to facilitate communication across threads, and it is hence referred to as shared memory. Typically, scratchpad memories are organized in multiple independently-accessible memory banks. Therefore if all memory accesses request data mapped to different banks, they can be handled in parallel. Bank conflicts occur whenever multiple requests are made for data

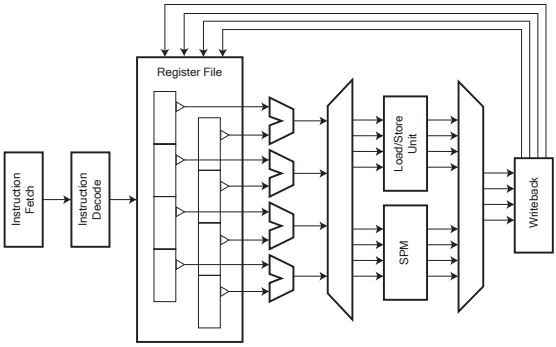


Fig. 1 High-level generic GPU-like core with scratchpad memory.

within the same bank [20]. If N parallel memory accesses request the same bank, the hardware serializes the memory accesses, causing an N-times slowdown [19]. In this context, a dynamic bank remapping mechanism, based on specific kernel access pattern, may help minimize bank conflicts.

Bank conflict reduction has been addressed by several scientific works during the last years. A generalized memory-partitioning (GPM) framework to provide high data throughput of on-chip memories using a polyhedral mode is proposed in [29]. GPM allows intra-bank offset generation in order to reduce bank conflicts. Memory partitioning adopted in these works are cyclic partitioning and block partitioning, as presented in [13]. In [17] the authors address the problem of automated memory partitioning providing the opportunity of customizing the memory architecture based on the application access patterns and the bank mapping problem with a lattice-based approach. While bank conflicts in shared memory is a significant problem, existing GPU-like accelerators [12, 8, 4, 23] lack bank remapping mechanisms to minimize such conflicts.

3 Architecture

SPM interface and operation. Figure 1 depicts a block diagram of the SPM in the context of a GPU-like core architecture. A GPU-like core has a SIMD structure with L multiple lanes. All lanes share the same control unit, hence in each clock cycle they execute the same instruction, although on different data. Every time a new instruction is issued, it is propagated to all execution lanes, each taking the operands from their corresponding portion of a vectorial register file addressed by the instruction. The typical memory instructions provided by a SIMD ISA offer gather and scatter operations. Such operations are respectively vectorial memory load and store memory accesses. If the SIMD core has a single-bank SPM with a single memory port, the previous instructions require al least L clock cycles. This is

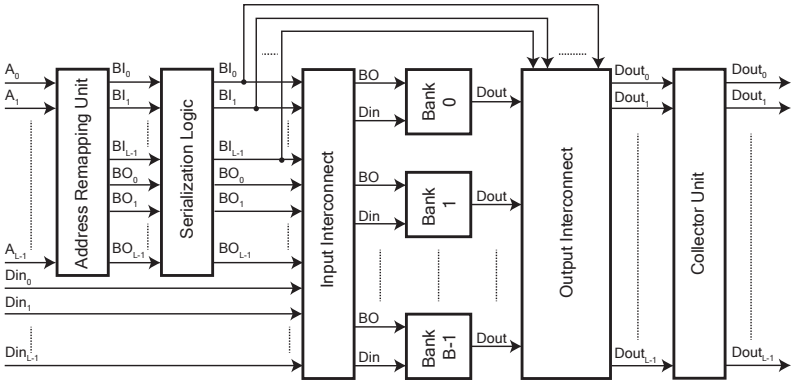


Fig. 2 Architecture overview.

because the L lanes cannot access a single memory port with different addresses in the same clock cycle.

Architecture. Figure 2 shows the internal architecture of the proposed SPM. The SPM takes as input L different addresses to provides support to the scattered memory access. It can be regarded as an FSM with two states: Ready and Busy. In the Ready state, the SPM is ready to accept a new memory request. In the Busy state, the SPM cannot accept any request as it is still processing the previous one, so in this state all input requests will be ignored. The Address Mapping Unit computes in parallel the bank index and the bank offset for each of the L memory addresses coming from the processor lanes. Bank index (BI_i in Figure 2) is the index of the bank to which the address is mapped. Bank offset (BO_i in Figure 2) is the address of the word into the bank. The Address Mapping Unit behaviour can be changed at run time in order to change the relationship between addresses and banks. This is a key feature in that it allows the adoption of the mapping strategy that best suits the executed workload. The Serialization Logic Unit performs the conflict detection and the serialization of the conflicting requests. Whenever an n-way conflict is detected, the Serialization Logic Unit puts the SPM in the busy state and splits the requests into n conflict-free requests issued serially in the next n clock cycles. When the last request is issued, the Serialization Logic Unit put the SPM in the ready state. Notice that for the Serialization Logic Unit, multiple accesses to the same address are not seen as a conflict, as in this occurrence a broadcast mechanism is activated. This broadcast mechanism provides an efficient way to satisfy multiple load requests for the same constant parameters. The Input Interconnect is an interconnection network that steers source data and/or control signals coming from a lane in the GPU-like processor to the destination bank. Because the Input Interconnect follows the Serialization Logic Unit, it only accepts one request per bank. Then, there are the B memory banks providing the required memory elements. Each memory bank receives the bank offset, the source data, and the control signal form the lane that addressed it. Each bank has a single read/write port with a byte-level write enable signal to support instructions with operand sizes smaller than word. Furthermore,

Cyclic mappig

Bank0	Bank1	Bank2	Bank3
0x00	0x04	0x08	0x0c
0x10	0x14	0x18	0x1c
0x20	0x24	0x28	0x2c
0x30	0x34	0x38	0x3c

Block mappig

Bank0	Bank1	Bank2	Bank3
0x00	0x10	0x20	0x30
0x04	0x14	0x24	0x34
0x08	0x18	0x28	0x38
0x0c	0x1c	0x2c	0x3c

Generalized Cyclic mappig

Bank0	Bank1	Bank2	Bank3
0x00	0x04	0x08	0x0c
0x1c	0x10	0x14	0x18
0x28	0x2c	0x20	0x24
0x34	0x38	0x3c	0x30

Fig. 3 The figure shows how addresses are mapped onto the banks. The memory is byte addressable and each word is four byte wide. In the case of generalized cyclic mapping the remapping factor is 1.

each lane controls a bit in an L-bit mask bus that is propagated through the Input Interconnect to the appropriate bank. This bit acts as a bank enable signal. In this way, we can disable some lanes and execute operations on a vector smaller than L elements. The Output Interconnect propagates the loaded data to the lane that requested it. Last, there is a Collector Unit which is a set of L registers that collect the results coming from the serialized requests outputting them as a single vector.

Remapping. As mentioned above, the mapping between addresses and banks can be changed at run time through the Address Mapping Unit. The technical literature presents several mapping strategies, including cyclic and block mapping [13, 29]. These strategies are summarized in Figure 3. Cyclic mapping assigns consecutive words to adjacent banks (Bank $B-1$ is adjacent to Bank 0). Block mapping maps consecutive words onto consecutive lines of the same banks. The block-cycle mapping is a hybrid strategy. With $B = 2^b$ banks, $W = 2^w$ bytes in a word, and $D = 2^d$ words in a single bank, a scratchpad memory address is made of $w+b+d$ bits. Figure 3 shows a cycling remapping, which can be easily obtained by repartitioning the memory address. The Address Mapping Unit described in this work implements a generalization of cyclic mapping, which we call *generalized-cyclic mapping*. By adopting this strategy, many kernels that generate conflicts with cyclic mapping change their pattern by accessing data on the memory diagonal, thereby reducing the number of conflicts.

Implementation details. The proposed configurable scratchpad memory was described in HDL and synthesized for a Xilinx FPGA device. In particular, we used Xilinx Vivado to synthesize the proposed architecture on a Xilinx Virtex7-2000t FPGA (part number xc7v2000tflg1925-2). We built our architecture with a variable number of banks and lanes. Figure 4 reports our SPM occupation in terms of LUTs and Flip-Flops (FFs) for a variable number of banks. On the other hand, Figure 5 reports our SPM occupation in terms of LUTs and Flip-Flops (FFs) for a variable number of lanes. Increasing the number of banks heavily affects LUTs occupation, while the number of lanes mostly affects the FF count.

The proposed SPM has been validated with the Verilator RTL simulator tool [28], which compiles synthesizable Verilog, SystemVerilog, and Synthesis assertions into a C++ behavioural model (called the *verilated* model), effectively converting RTL design validation into C++ program validation. In addition, a cycle accurate event-driven emulator written in C++ was purposely developed for verifying the proposed design. The SPM verilated model and the SPM emulator are integrated in a test-

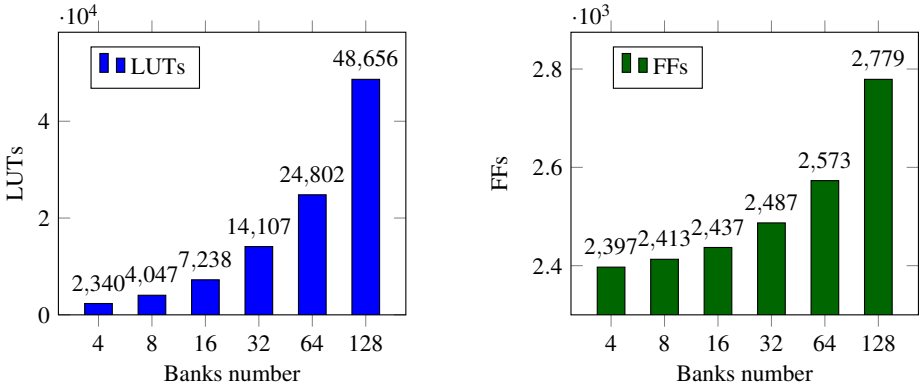


Fig. 4 LUTs and FFs occupation of the FPGA-based SPM design for a variable number of banks

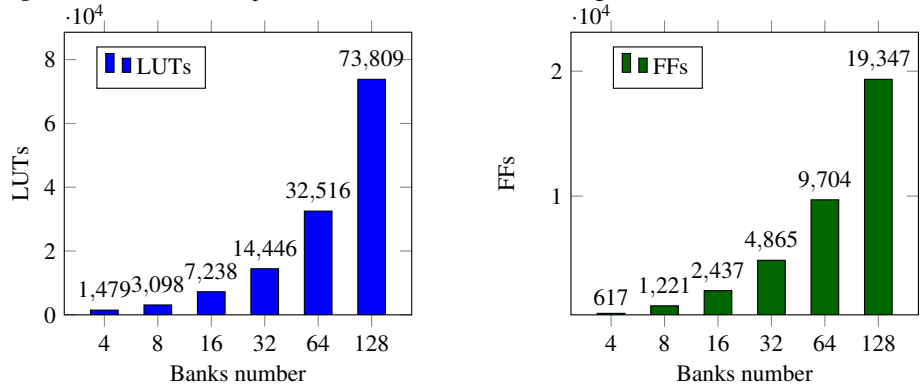


Fig. 5 LUTs and FFs occupation of the FPGA-based SPM design for a variable number of lanes

bench platform, that provides the same inputs to the emulator and the verilated model. The test platform compares the outputs from the two models at every simulation cycle, checking if the verilated model and the emulator generate the same responses. Notice that the Verilator tool supports the HDL code coverage analysis feature, which helped us create a test suite with full coverage of the SystemVerilog code.

Integration in a GPU-like core. The presented SPM, as previously explained, has a variable latency, which may be a potential complication for integration in GPU-like architectures. At issue time, before the operand fetch, the control unit is unaware of the actual latency of a scratchpad memory instruction and can not detect possible Writeback structural hazards. To avoid this problem, a core must support a dynamic on-the-fly structural hazard handler at Writeback stage.

4 Evaluation

The experimental evaluation was essentially meant to demonstrate to which extent the amount of bank conflicts can be reduced by changing the parameters in the proposed configurable scratchpad memory. In particular, to this end, our experiments assess how simultaneous memory accesses, as well as the bank remapping feature may affect the total bank conflict count.

Methodology. We first identified a few kernels that have potentially highly parallel memory accesses and that can benefit from the scratchpad memory support. Many such kernels exist in benchmark suites like PolyBench [26]. Next, we rewrote each of those kernels to increase the kernel memory access parallelism, as our aim was to study how conflicts vary with a variable number of parallel memory requests. We then extracted the access patterns for each kernel and we run it on our scratchpad emulator. The emulator is cycle-accurate, ensuring exact timings for the simulated accesses as the scratchpad memory and the emulator proceed in a lock-step fashion under the same inputs. Last, we collected the emulator response in terms of total bank conflicts for all the memory accesses issued by the kernel, through a counter that is incremented whenever a bank conflict occurs. We repeated this experiment for different remapping functions identified for the specific kernel as well as for a variable number of banks.

Matrix Multiplication. Square matrix-matrix multiplication is a classic bank conflict sensitive kernel. In this benchmark, we evaluated the square matrix access patterns and how the configurable parameters influence the scratchpad bank conflict count.

Listing 1 Matrix Multiplication parameterized on the number of lanes.

```
for (int i = 0; i < DIM; ++i)
  for (int j = 0; j < DIM; ++j)
    for (int k = 0; k < DIM/numLane; ++k)
          for (int lane = 0; lane < numLane; lane++){
                accessA[index][lane] = (i*DIM + k*numLane + lane)*4;
                accessB[index][lane] = ((k*numLane + lane)*DIM + j)*4;
          }
          index++;
```

We rewrote the code so as to maximize the exploitation of the available number of lanes in the target model of GPU-like processor. The inner cycle, shown in Listing 1, calculates which memory address will be accessed by each lane for both matrices. We have a fixed square matrix size $DIM = 128$. The number of hardware lanes is $numLane = [4,8,16,32]$ while the number of banks is $numBanks = [16,32,64,128,256,512,1024]$. The function bank remapping is $(Entry \cdot c + Bank)$ mod $(NUMBANK)$ with $c = [1,2,4,8,16]$.

The total scratchpad memory is kept constant and equal to $BANKnumber \times ENTRY perBank = 2 \times DIM^2$, so that the SPM can store both matrices completely.

Table 1 Matrix Multiplication results.

Lanes	Banks	Remapping factor				
		No Remap	1	2	4	8
4	16	262146	131072	262146	262146	262146
	32	262146	0	0	131072	262146
	64	262146	0	0	0	0
	128	262146	0	0	0	0
	256	131072	0	0	0	0
	512	0	0	0	0	0
	1024	0	0	0	0	0
8	16	183505	131073	183505	183505	183505
	32	183505	0	65536	131073	183505
	64	183505	0	0	0	65536
	128	183505	0	0	0	0
	256	131073	0	0	0	0
	512	65536	0	0	0	0
	1024	0	0	0	0	0
16	16	109230	91756	109230	109230	109230
	32	109230	32768	65538	91756	109230
	64	109230	0	0	32768	65538
	128	109230	0	0	0	0
	256	91756	0	0	0	0
	512	65538	0	0	0	0
	1024	32768	0	0	0	0
32	16	61696	58256	61696	61696	61696
	32	59768	32769	45878	54615	59768
	64	59768	0	16384	32769	45878
	128	59768	0	0	0	16384
	256	54615	0	0	0	0
	512	45878	0	0	0	0
	1024	32769	0	0	0	0

Results in Table 1 show that bank remapping has a greater impact than the other parameters. A remapping coefficient $c = 1$ drastically reduces bank conflicts, even with a limited number of banks, while adding little resource overhead compared to a solution relying on a large number of parallel banks.

Image Mean Filter 5×5. Mean filtering is a simple kernel to implement image smoothing. It is used to reduce noise in images. The filter replaces each pixel value in an image with the mean value of its neighbors, including itself. In our study a 5×5 square kernel is used.

Listing 2 shows our parallelized version of the mean filter. For this kernel we keep a fixed square matrix size $DIM = 128$ and a fixed number of lanes $numLane = 30$. The total scratchpad memory is kept constant and equal to $BANKnumber \times ENTRY\,perBank = DIM^2$. We evaluated the bank conflicts for a variable number of banks and for two bank remapping functions: no remap and $(Entry \cdot 5 + Bank)$ mod $(NUMBANK)$. The results are shown in Table 2. As in the case of the

Listing 2 Image Mean Filter 5x5.

```
#define OFFSET(x, y) (((x)*DIM + y)*4)

for (int i = 2; i < DIM - 3; ++i)
  for (int j = 2; j < DIM - 3; ++j) {
        for (int w1 = -W1; w1 <=W1; w1++ ){
          for (int w2 = -W2; w2 <= W2; w2++){
                a = baseA.getAddress() + OFFSET(i+w1, j+w2);
                l = (w1 + 2)*5 + (w2 + 2);
                accessA[index][l] = a;
          }
        }
        index++;
  }
```

matrix multiplication kernel, the remapping function has the largest impact on the bank conflict count.

Table 2 Image Mean Filter 5x5.

Banks	No Remap	Remap
16	7565	1722
32	7565	0
64	7565	0
128	7565	0
256	0	0
512	0	0
1024	0	0

5 Conclusion

In this work we presented a configurable GPU-like oriented scratchpad memory fully synthesizable on FPGAs. Various architectural aspects like the number of banks, the number of lanes, the bank remapping function, and the size of the total memory are parameterized. Reconfigurability helped explore architectural choices and assess their impact. We described the SPM design in HDL and extensively validated it. We also developed a software cycle accurate and event-driven emulator of our SPM component to support the experimental evaluation with real code. Through two case studies, a matrix multiplication and a 5×5 image mean filter, we showed the performance implications with different configurations and demonstrated the benefits of using a dedicated hardware bank remapping function over other architectural parameters. As a long-term goal of this research, we aim to integrate our

SPM architecture in an open source GPU-like core, enabling it to take full advantage of the underlying reconfigurable hardware technologies.

Acknowledgments. This work is supported by the European Commission in the framework of the H2020-FETHPC-2014 project n. 671668 - MANGO: exploring Manycore Architectures for Next-GeneratiOn HPC systems.

References

1. The Altera SDK for open computing language (OpenCL). https://www.altera.com/products/design-software/embedded-software-developers/opencl/overview.html
2. Nvidia's next generation cuda compute architecture. NVidia, Santa Clara, Calif, USA (2009)
3. An independent analysis of Altera's FPGA floating-point DSP design flow. Berkeley Design Technology, Inc (2011)
4. Al-Dujaili, A., Deragisch, F., Hagiescu, A., Wong, W.F.: Guppy: A GPU-like soft-core processor. In: Field-Programmable Technology (FPT), 2012 International Conference on, pp. 57–60 (2012)
5. Amato, F., Barbareschi, M., Casola, V., Mazzeo, A.: An FPGA-based smart classifier for decision support systems. Studies in Computational Intelligence **511**, 289–299 (2014)
6. Amato, F., Fasolino, A., Mazzeo, A., Moscato, V., Picariello, A., Romano, S., Tramontana, P.: Ensuring semantic interoperability for e-health applications. In: Proceedings of the International Conference on Complex, Intelligent and Software Intensive Systems, CISIS 2011, pp. 315–320 (2011)
7. Amato, F., Mazzeo, A., Penta, A., Picariello, A.: Building RDF ontologies from semi-structured legal documents. pp. 997–1002 (2008)
8. Balasubramanian, R., Gangadhar, V., Guo, Z., Ho, C.H., Joseph, C., Menon, J., Drumond, M.P., Paul, R., Prasad, S., Valathol, P., Sankaralingam, K.: Enabling GPGPU low-level hardware explorations with MIAOW: An open-source RTL implementation of a GPGPU. ACM Trans. Archit. Code Optim. **12**(2), 21:21:1–21:21:25 (2015)
9. Barbareschi, M., Del Prete, S., Gargiulo, F., Mazzeo, A., Sansone, C.: Decision tree-based multiple classifier systems: An FPGA perspective. In: International Workshop on Multiple Classifier Systems, pp. 194–205. Springer (2015)
10. Barbareschi, M., Iannucci, F., Mazzeo, A.: Automatic design space exploration of approximate algorithms for big data applications. In: 2016 30th International Conference on Advanced Information Networking and Applications Workshops (WAINA), pp. 40–45. IEEE (2016)
11. Barbareschi, M., Iannucci, F., Mazzeo, A.: An extendible design exploration tool for supporting approximate computing techniques. In: 2016 International Conference on Design and Technology of Integrated Systems in Nanoscale Era (DTIS), pp. 1–6. IEEE (2016)
12. Bush, J., Dexter, P., Miller, T.N.: Nyami: a synthesizable GPU architectural model for general-purpose and graphics-specific workloads. In: Performance Analysis of Systems and Software (ISPASS), 2015 IEEE International Symposium on, pp. 173–182 (2015)
13. Chatterjee, S., et al.: Generating local addresses and communication sets for data-parallel programs. SIGPLAN Not. **28**(7), 149–158 (1993)
14. Cilardo, A.: Exploring the potential of threshold logic for cryptography-related operations. IEEE Transactions on Computers **60**(4), 452–462 (2011)
15. Cilardo, A., De Caro, D., Petra, N., Caserta, F., Mazzocca, N., Napoli, E., Strollo, A.: High speed speculative multipliers based on speculative carry-save tree. IEEE Transactions on Circuits and Systems I: Regular Papers **61**(12), 3426–3435 (2014)
16. Cilardo, A., Durante, P., Lofiego, C., Mazzeo, A.: Early prediction of hardware complexity in HLL-to-HDL translation. pp. 483–488 (2010)

17. Cilardo, A., Gallo, L.: Improving multibank memory access parallelism with lattice-based partitioning. ACM Transactions on Architecture and Code Optimization (TACO) **11**(4), 45 (2015)
18. Cilardo, A., Gallo, L., Mazzeo, A., Mazzocca, N.: Efficient and scalable OpenMP-based system-level design. pp. 988–991 (2013)
19. Coon, B., et al.: Shared memory with parallel access and access conflict resolution mechanism. U.S. Patent No. 8,108,625 (2012)
20. Farber, R.: CUDA application design and development. Elsevier (2011)
21. Fusella, E., Cilardo, A.: H^2ONoC: A hybrid optical-electronic NoC based on hybrid topology. IEEE Transactions on Very Large Scale Integration (VLSI) Systems (2016)
22. Fusella, E., Cilardo, A.: Minimizing power loss in optical networks-on-chip through application-specific mapping. Microprocessors and Microsystems (2016)
23. Kingyens, J., Steffan, J.: The potential for a GPU-like overlay architecture for FPGAs. International Journal of Reconfigurable Computing (2011)
24. Kuon, I., Rose, J.: Measuring the gap between FPGAs and ASICs. In: Proceedings of the 2006 ACM/SIGDA 14th International Symposium on Field Programmable Gate Arrays, FPGA '06, pp. 21–30. ACM, New York, NY, USA (2006)
25. Paranjape, K., Hebert, S., Masson, B.: Heterogeneous computing in the cloud: Crunching big data and democratizing HPC access for the life sciences. Intel Corporation (2010)
26. Pouchet, L.N.: Polybench: The polyhedral benchmark suite. http://www. cs. ucla. edu/pouchet/software/polybench (2012)
27. Sarkar, S., et al.: Hardware accelerators for biocomputing: A survey. In: Proceedings of 2010 IEEE International Symposium on Circuits and Systems (2010)
28. Snyder, W., Wasson, P., Galbi, D.: Verilator (2007)
29. Wang, Y., Li, P., Cong, J.: Theory and algorithm for generalized memory partitioning in high-level synthesis. In: Proceedings of the 2014 ACM/SIGDA International Symposium on Field-programmable Gate Arrays, FPGA '14, pp. 199–208. ACM, New York, NY, USA (2014)
30. Wirbel, L.: Xilinx SDAccel: a unified development environment for tomorrow's data center. The Linley Group Inc (2014)

Research of Double Threshold Collaborative Spectrum Sensing based on OFDM

Ruilian TAN[1,2]
[1] Armed Police Engineering University Electronic Technology Laboratory of Information Engineering Department, Xi'an 710086, China
[2] College of Equipment Management and Safety Project, Air Force Engineering University, Xi'an 710051,China
madamtan@126.com

Abstract. In view of the shortage of the single threshold energy detection in spectrum sensing, the double threshold energy detection is studied. Combined with the characteristics of OFDM spectrum efficiency, the double threshold energy detection and OFDM system are combined. The method is on the basis of OFDM with the sub-channel as the user to be perceived. According to the available subcarrier number, the actual number of user to be perceived is determined, and the double threshold energy detection is conducted in sub-channels. The data fusion center makes decisions by means of "or" criterion. The simulation results show that the method can well identify the main users, thus controlling the data transmission of OFDM .

Key Words. Cognitive Radio; Collaborative Spectrum Sensing; Double Threshold Energy Detection; OFDM

1 Introduction

As the growth of the wireless communication business, spectrum resources increasingly tense, the FCC spectrum strategy task group pointed out that in the literature, the existing spectrum demand more nervous several hundred MHz to 3GHz wireless spectrum available resources are assigned to complete almost [1]. In view of the allocated spectrum resource utilization problem of low, Joseph Mitola put forward the Cognitive Radio (CR)) technology in 1999 [2], the concept of it to study the environmental awareness, and real-time adjust the transmission parameters to adapt to the change of external environment, which use has not been authorized band [3]. Spectrum sensing is the premise of cognitive radio, only spectrum hole, measured by the frequency spectrum can be achieved to dynamic spectrum access, it is in the normal communication does not interfere with the primary user of the key technology to improve spectrum efficiency under the premise of [4]. At present, the idle spectrum detection method mainly includes the matched filtering detection, cyclic stationary feature detection and energy detection. The filter detection needs to know the main users of a priori information, such as modulation method, and the prior information in

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_2

the actual process of communication is difficult to get [5]; complicated cycle stationary feature detection algorithm and the difficulties [6]. Literature points out that the energy detection is don't need any priori information of spectrum detection method[7], the starting point is the energy of the signal and noise is greater than the energy of the noise, belong to the blind detection algorithm, the algorithm is relatively simple, usually adopt single threshold energy, however, in actual communication performance is often in the multipath fading and shadow effect factors such as constraints, greatly decline the performance of the single threshold energy detection, aiming at this problem, can the method of using double threshold energy of collaborative detection [8].

In wireless communication, orthogonal frequency division multiplexing (OFDM) modulation technology has high spectrum efficiency, on the resistance to multipath fading, resist narrow-band interference obvious advantages, can improve the system's ability to non-line-of-sight propagation [9], and therefore OFDM transmission system is suitable for digital signal transmission, is widely used. This paper mainly studies the double threshold energy detection in cognitive radio spectrum perception, combined with the characteristics of OFDM spectrum utilization rate is higher, and puts forward the energies of the double threshold detection method combined with OFDM system. The method on the basis of OFDM system, the system as sub-channels to perception of users, according to the energy statistics to determine the available subcarrier number as the actual perception of users, in sub-channels, double threshold energy detection, data fusion center by means of "or" criterion. Simulation results show that the method can well identify the main users, in the main user when idle, OFDM system can be normal for data transmission, with the increase of the subcarrier number available, you can get a better recognition performance.

2 Double threshold detection energy model

2.1 Energy detection model

Energy spectrum detection is the most basic method, its judgment method is to preset a threshold method, through the energy detector and set threshold comparison, more than decision threshold method, the main users determine the spectrum exist, otherwise the frequencies in the idle state. Our energy detection can be defined as a binary hypothesis problem:

$$\begin{cases} H_0: y(k)=\omega(k) & H_0 \\ H_1: y(k)=s(k)+\omega(k) & H_1 \end{cases} \tag{1}$$

$y(k)$ is the signal received by the users, $s(k)$ is the detected signal, namely the main signal, $\omega(k)$ is additive white Gaussian noise, k = 1, 2..., n; n is the expressed received signals, H_0 is the idle stats, and the main user does not exist, H_1 is the occupation state, namely the primary user. When users are out of the frequency, $s(k)$ is zero, the detection is in H_0, the energy detector can be written as P [9]:

N is the vector dimension sampling sequence. Because of take up channel signal energy must be greater than the energy of the idle state, this time can determine whether the main user exists by P compared with preset threshold λ_E, λ_E is the

primary user occupancy state; otherwise it is idle, and it can be used access the idle spectrum.

Spectrum detection performance by two probability measure [2]: detection probability P_d and false alarm probability P_f. P_f is the occupied channel frequency when the channel is idle, the increase of P_f will make cognitive users lose the opportunity to access the idle spectrum. It can be expressed as:

$$P_d = probability(P > \lambda_E | H_1) \quad (3)$$

$$P_f = probability(P > \lambda_E | H_0) \quad (4)$$

P adheres to the chi square distribution with the degree of freedom 2υ:

$$P \sim \begin{cases} \chi^2_{2\upsilon} & H_0 \\ \chi^2_{2\upsilon}(2\eta) & H_1 \end{cases} \quad (5)$$

υ = TW, T is the observation time, W is the observation bandwidth, and η is the signal-noise ratio.

The probability distribution function of P is:

$$f_P(m) = \begin{cases} \frac{1}{2^{\upsilon}\Gamma(u)} m^{\upsilon-1} e^{-\frac{m}{2}} & H_0 \\ \frac{1}{2}\left(\frac{m}{2\eta}\right)^{\frac{\upsilon-1}{2}} e^{-\frac{2\eta+m}{2}} I_{\upsilon-1}\left(\sqrt{2\eta m}\right) & H_1 \end{cases} \quad (6)$$

$\Gamma(\bullet)$ is the gamma function, $I_{\upsilon}(\bullet)$ is the υ-order first kind of deformation Beisaier function.

Assumes that the cognitive radio network for the decline, the detection probability P_d and false alarm probability P_f can be obtained under additive white Gaussian noise:

$$P_d = probability(P > \lambda_E | H_1) = Q_{\upsilon}\left(\sqrt{2\eta}, \sqrt{\lambda_E}\right) \quad (7)$$

$$P_f = probability(P > \lambda_E | H_0) = \frac{\Gamma(\upsilon, \lambda_E / 2)}{\Gamma(\upsilon)} \quad (8)$$

Among them, $Q_{\upsilon}(a,x)$ is the general Marcum Q function.

In accordance with the above assumptions, the energy detector flow chart shown in figure 1, cognitive users the received signal y (t) first by band-pass filter to select the corresponding bandwidth and center frequency, and then through the square law to calculate the energy of the signal device, and through the integrator to implement energy accumulated over a period of time, the resulting energy detector statistics P, P into the judgment and decision of preset threshold λ_E for comparison; more than λ_E is the primary user occupancy state; otherwise, it is idle, used to access the idle spectrum.

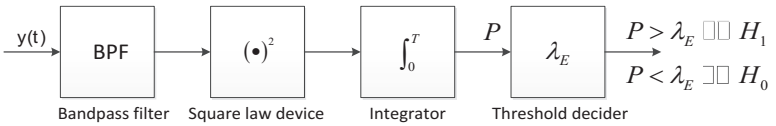


Fig 1 The flow chart of energy detector

2.2 Double threshold energy detection

Through the above analysis, we know, the traditional energy detection is based on the single threshold; the threshold is only one decision threshold. When the primary user signal power and noise is small, single threshold energy detection is a good way to work, namely under the condition of low SNR, there may not be able to identify primary user signal, single threshold energy detection may make a wrong decision, so as to make the false alarm rate increases, weaken in idle spectrum. In view of this, the literature [10] [11] based on the shortage of the single threshold energy detection, put forward the double threshold detection together, and through the computer simulation proves the feasibility of this method. The test set up two decision threshold λ_0 and λ_1, when the signal energy statistic value is greater than λ_1, the ruling Lord user exists, send data fusion center local decision 1; if signal energy statistic value is less than λ_0, it means that the main user does not exist, the data fusion center send local decision 0; if signal between λ_0 and λ_1, it needs to send signal energy statistics to the data fusion center again.

3 Double threshold collaborative spectrum sensing based on OFDM

In wireless communication, orthogonal frequency division multiplexing (OFDM) modulation technology has high spectrum efficiency, resistance to multipath fading, resist narrow-band interference on obvious advantages. OFDM basic idea is to decompose high-speed data stream into many low speed of data flow, the channel is divided into many sub orthogonal channel, the channel between the carrier of maintain orthogonality, spectrum overlap each other, in the form of parallel transmission in multiple child carrier, as a result of the orthogonal relation between subcarrier so as to eliminate the influence of data between the carrier, due to the spectral overlap each other so as to improve the spectrum utilization. The OFDM technology can be combined with a spectrum sensing technology used.The energy detection can use FFT module and OFDM modulation used when FFT module reuse, so as to reduce the system complexity. At the same time, the spectrum sensing consumed time period can also be guaranteed. In this paper, according to the principle of OFDM combined collaborative spectrum sensing, transmission model and collaborative spectrum sensing model OFDM system is obtained as shown in figure 2 and figure 3.

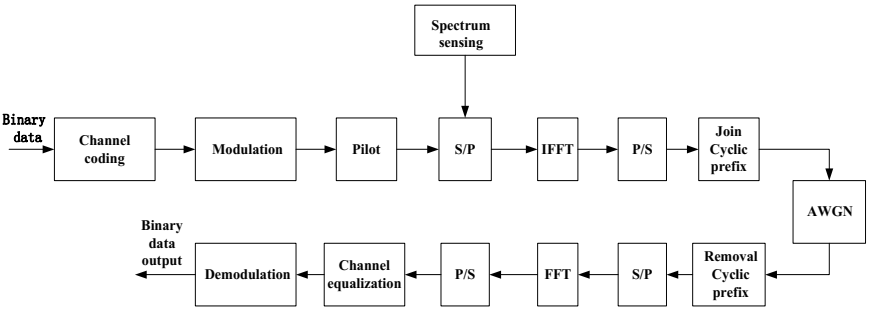


Fig 2 OFDM system transmission model

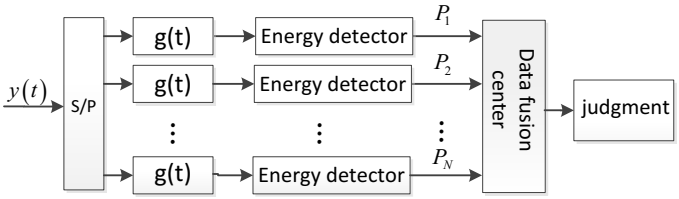


Fig 3 OFDM collaborative spectrum sensing model

The signal enters the OFDM transmission system, after a string of distribution and transform to N road rate lower, sub-channels is due to the rate of sub-channels to 1/N, symbol cycle expanded to N times. Subcarrier can by the channel bandwidth, the number of data throughput and useful symbol duration to decide. Assumption under the given conditions there are N-K carriers, K is carrier at this point, equivalent to N-K users perceived g (t), g (t) in the approach channel, get energy test statistics, various statistical values into the data fusion center, adopt corresponding "or" judgment standard, can be perceived. Feel the result control system is working correctly, to determine if the spectrum idle system normal to send data; to determine if the spectrum, is to stop sending data.

Due to the energy detector with double threshold judgment, the data fusion center at this time will receive the two kinds of data sent by N-K g(t), namely:

$$V_i = \begin{cases} A_i \ , \ \lambda_0 \le A_i \le \lambda_1 \\ B_i \ , \ A_i > \lambda_1 \ \ or \ \ A_i < \lambda_0 \end{cases} \tag{9}$$

According to the maximal-ratio combining principle:

$$D = \begin{cases} 0 \ , \ 0 \le \sum_{i=1}^{N-K} w_i \bullet A_i \le \lambda \\ 1 \ , \ \sum_{i=1}^{N-K} w_i \bullet A_i > \lambda \end{cases} \tag{10}$$

$w_i = \dfrac{\gamma_i}{\sqrt{\sum_{i=1}^{N-K} \gamma_j^{\,2}}}$ is the maximum ratio combining coefficient, and λ is the maximum merger decision threshold.

The decision after data fusion is as follows:

$$D_{final} = \begin{cases} H_0 \ : \ D=0 \\ H_1 \ : \ D=1 \end{cases} \tag{11}$$

In the data fusion decision criterion, it usually uses the "or" criterion, "or" criterion refers that as long as there is a cognitive users determine the main existence, you get to the final result, the principle was proved to be the one of the rule of data fusion has a better detection performance. Because each subcarrier in OFDM system is relatively independent, so entering the data fusion center of N-K energy statistics will also be independent identically distributed, the data fusion center by means of "or" criterion for the collaborative detection probability and false alarm probability are:

$$Q_{d.or} = 1 - \prod_{i=1}^{N-K} \left(1 - P_{d.i}\right) \quad (12)$$

$$Q_{f.or} = 1 - \prod_{i=1}^{N-K} \left(1 - P_{f.i}\right) \quad (13)$$

4 Simulation and Result Analysis

4.1 Detection performance analysis under different SNR

In AWGN channel under MATLAB simulation parameters are as follows: the primary user signal is $X = \sqrt{2\eta_i}\sin\left(2*\pi*T\right)$, adjust the signal amplitude according to the signal to noise ratio η_i; rate L = 300; sampling frequency Fs = 1000 Hz; carrier frequency f_0=fs/20; η=[-20,0]dB, step 0.5. N-K=14; P_{f_1} = 0.01, P_{f_2} =0.1, P_{f_3} =0.2, P_{f_4} =0.3.

As shown in figure 4, after BPSK modulation for output of the signal with noise, figure 5 energy values for after energy detector, it can be seen that the main user is noise and produced distortion, at the same time due to the superposition of the noise energy mix signal energy increases, so it is very important how to detect the primary users.

Figure 6 shows the different SNR performance of the energy detection. From formula (8), we can know when know that under the premise of false-alarm probability can be calculated each user perception detection threshold $\lambda_{E,i}$. The formula (7) shows that when the SNR η_i is determined, the detection probability $P_{d,i}$ can be obtained at this time. Formula (12) and (13) can be the collaborative detection probability and false alarm probability. From the simulation results, it can be seen that when the number of carrier and false alarm probability; when the signal-to-noise ratio at the same time, the greater the false-alarm probability, the greater the probability of detection. Too much means that cognitive users will lose the chance of more access to the spectrum hole, too low will make testing less than primary user, error-prone, so must balance between the detection probability and false alarm probability, selected appropriate threshold method.

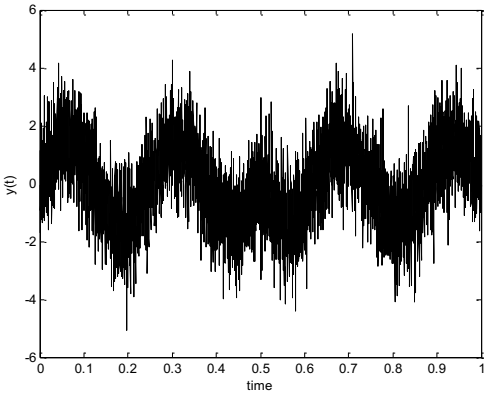


Fig 4 Primary user signal with noise

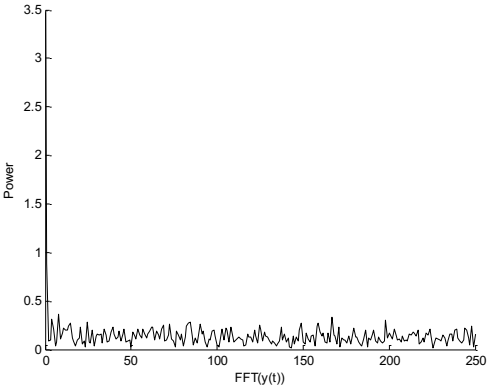


Fig 5 Energy detection

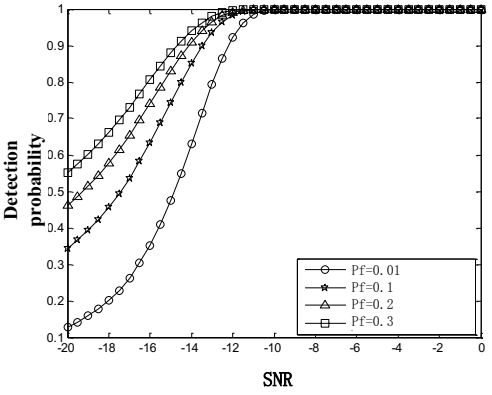


Fig 6 Signal-to-noise ratio and detection probability curve

4.2 Detection performance analysis under "or" fusion criterion

Set the MATLAB simulation parameters under AWGN channel as follows: primary user signal as randn (n, 1); W=5*10^{4}Hz; Fs = 2*W; SNR η = -10 dB; work subcarrier number N-K = 14; as shown in figure 7, conduct the energy detection performance under the "or" fusion criterion.

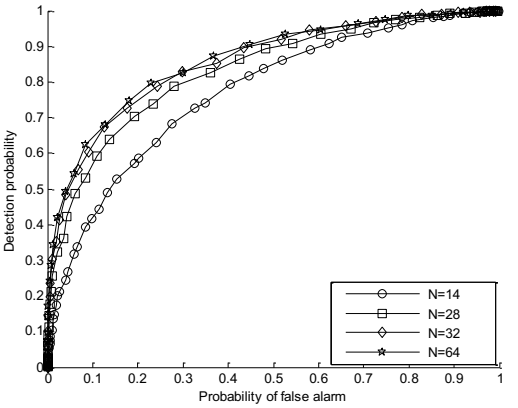


Fig 7 False-alarm probability and detection probability curve under "or" fusion criterion

From figure 7, it can be seen that under the "or" fusion criterion, when the number of carrier is defined, the detection probability will increase with the increase of false-alarm probability. When certain false-alarm probability is defined, with the increase of subcarrier number N - K, the detection probability will also increase. The formula (7) shows that in terms of a single user, the signal-to-noise ratio and false alarm probability are determined. The formula (12) shows that $Q_{d.or} = 1 - \prod_{i=1}^{N-K}(1 - P_{d.i})$. The joint detection probability $Q_{d.or}$ is mainly influenced by N-K, the more N-K number is, the greater the simulation results will be consistent with theoretical analysis. Detection probability performance, however, with the increase of the subcarrier number N - K number starts to degrade, when the carrier number is 64, its detection probability will be similar to the 32 subcarriers. With the increase of the false alarm probability, the performance becomes poor, and even in some places than the subcarrier number 32 detection performances, so the appropriate subcarrier number can achieve better detection probability. At the same time, because our main users of radon (n,1) function is randomly generated, the false alarm probability and detection probability curve are not smooth, but determined by the individual test results.

5 Conclusion

This paper studies the application of spectrum sensing algorithm in the cognitive radio, focuses on analyzing the deficiency of the single threshold energy detection and

advantage of double threshold energy detection, and proposes the combination of double threshold detection and OFDM system combined with the advantages of high spectrum efficiency of OFDM system. The simulation results prove the feasibility of the proposed method, and the experimental data and theoretical reasoning are consistent. The increasing subcarrier number can improve the performance of system identification, but when it increases to a certain number, the recognition performance will become poor. Therefore, the next step of research is to focus on the subcarrier number suitable for the system.

References

1. First Report and Order, Federal Communication Commission Std. FCC 02-48, Feb.2002.
2. P Qihang,Z Kun,W Jun,et al.A Distributed Spectrum Sensing Scheme Based on Credibility and Evidence Theory in Cognitive Radio Context. Proceedings of 17th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC) ,2006.
3. I F Akyildiz, W Y Lee, M C Vuran, S Mohanty. Next generation/dynamic spectrum access/cognitive radio wireless networks: a survey[J]. Computer Networks Journal(Elsevier),2006,50(13):2127-2159.
4. Cui Tao, Gao Feifei, Nallanathan A. Optimal ofcooperative spectrum sensing in cognitive radio. IEEETransactions on Vehicular Technology ,2011 .
5. Penna F, Garello R, Spirito M A. Cooperative spectrum sensing based on the limiting Eigen value ratio distribution in wishart matrices[J].IEEE Communications Letters. 2009,13(7):507-509.
6. Urkowitz H. Energy detection of unknown deterministic signals[J]. Proceedings of the IEEE.1967,55(4):523-531.
7. Kostylev V I. Energy detection of a signal with random amplitude[J]. ICC, New York, 2002:1606-1610.
8. PanLiang Deng,YuanAn Liu,Gang Xie,XiaoYu Wang,BiHua Tang,JianFeng Sun. Error exponents for two-hop Gaussian multiple source-destination relay channels[J]. Science China Information Sciences, 2012 (2) :388-391.
9. BOMFIN R C D V,GUIMARAES D A,SOUZA R A A D. On the probability of false alarm of the power spectral density split cancellation method. IEEE Wireless Communication Letters . 2016,8 (7):47-55.
10. Li Qunwei,Li Zan,Shen Jian,Gao Rui.A novel spectrum sensing method in cognitive radio based on suprathreshold stochastic resonance. 2012IEEE International Conference on Communications . 2012：2682-2686.
11. G Ganesan and Y Li. Cooperative spectrum sensing for cognitive radios under bandwidth constraints[C]. Proceedings of the Wireless Communications and Networking Conference.Jun.2007:1-5.

Research on particle swarm optimization of variable parameter

Zhe LI[1,a], Ruilian TAN[1,2,b] , Baoxiang REN[1,c]
[1] College of Equipment Management and Safety Project, Air Force Engineering University, Xi'an 710051,China
[2] Armed Police Engineering University Electronic Technology Laboratory of Information Engineering Department, Xi'an 710086, China
[a] kongyanshi@126.com,[b] madamtan@126.com, [c] ganbing1981@126.com

Abstract. Aimed at particle swarm optimization, since there are a fewer adjustable parameters, when solving the multi-dimensional function, it is easy to meet premature convergence problem, so an improved particle swarm optimization of variable parameters is proposed. According to particle movement characteristics, the formula of particle velocity updating is improved to make all integrated into the corresponding weight factor; through weight factor, the particle optimization performance is adjusted. Three standard test functions are used for test, with comparison with other algorithms, and the simulation results show that by setting different weight factors, the proposed algorithm has better optimization precision and ability to execute, and the better result can be achieved when solving the multi-dimensional function.

Key Words. Particle Swarm Optimization, Parameter Selection, Weight Factor, Convergence Analysis

1 Introduction

Particle swarm optimization (PSO) is an intelligent algorithm based on flock foraging, proposed by the American electrical engineer Eberhart and social psychologist Kennedy in 1995 [1, 2]. Because the algorithm is simple, with a fewer required parameters and easiness to realize, etc., the algorithm is widely used in science and engineering [3, 4]. Since the algorithm parameter is less, it also makes it easy to fall into local optimum and dimension disasters. To solve these problems, especially for algorithm parameter adjustment, related scholars put forward many improved algorithms [5-8]. Among them, as early as in the literature [5], Shi and Eberhart introduced inertia weight, and later standard particle swarm optimization (SPSO) that is generally believed, which also laid the foundation for subsequent scholars to conduct the further study on parameter adjustment; in literature [6], Clerc put forward a kind of particle swarm optimization with constriction factor (CPSO),

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_3

using the constriction factor to improve convergence of particle swarm optimization by balancing local development ability and the global search ability. However, it is a still research focus and difficulty for how to seek more appropriate parameters of the algorithm.

On the basis of previous studies, this paper proposes a new parameter adjustment scheme. The essences of SPSO algorithm and CPSO algorithm are integrated; according to the characteristics of the particle movement, the parameter adjustment is made, and the velocity updating formula is matched with the corresponding weight factor, thus improving the optimization performance of the algorithm. Three benchmark functions are tested, and compared with existing SPSO algorithm and CPSO algorithm, which proves that the proposed algorithm in this paper has better performance in terms of optimization precision and execution, and can effectively avoid dimension disaster in solving high dimensional function optimization.

2 Standard Particle Swarm Optimization

Assume X = (X1, X2,... Xm) is made up of m particles, then in a D-dimension search space, a possible solution for the case is the position vector of the particle i in D-dimension search space Xi = (xi1, xi2,..., xiD) T. If Pi = (Pi1, Pi2,..., PiD) T represents the best position that the particle i has experienced, Pg = (Pg1 and Pg2,..., PgD) T represents the global optimal position of the population, Vi = (Vi1, Vi2,..., ViD) T represents the flying speed of the particle i particles in D-dimension space, then in the process of each iteration, the particles will update its velocity and position according to the formula (1) (2):

$$V_{id}(k+1) = \omega V_{id}(k) + c_1 r_1 (P_{id} - X_{id}(k)) + c_2 r_2 (P_{gd} - X_{id}(k)) \quad (1)$$

$$X_{id}(k+1) = X_{id}(k) + V_{id}(k+1) \quad (2)$$

Wherein, ω is inertia weight; $i = 1, 2..., n$; $d = 1, 2..., D$; V_{id} is the speed of the particles; c_1 and c_2 are referred to as the acceleration factor, usually taking 2.0; r_1 and r_2 are random numbers uniformly distributed between [0, 1]. PSO algorithm with inertia weight is the standard particle swarm optimization (SPSO); due to its less parameters, simple operation, and better optimization effect, it widely used. According to the concept put forward by Shi and Eberhart first, inertia weight ω reduces in linear gradient according to the type (3):

$$\omega_1(k) = \omega_{start} - (\omega_{start} - \omega_{end}) \left(\frac{k}{K_{\max}} \right) \quad (3)$$

Wherein, ω_{start} is the initial inertia weight, ω_{end} is inertia weight when iteration reaches the largest number, k is the current iteration number, $K_{\max}$ is the number of maximum iterations. However, SPSO algorithm has a few adjustable parameters, making the mathematical basis of the algorithm relatively weak; in the face of optimization problem of nonlinear function or multi-dimension function, it often early falls into local optimum, and the optimization effect is not ideal.

3 Improved PSO

3.1 PSO Algorithm with Constraint Factor

In literature [6], Clerc put forward a kind of particle swarm optimization with constriction factor (CPSO), and the definition of a simple particle swarm optimization with constriction factor:

$$V_{id} = \chi[V_{id} + c_1 r_1 (P_{id} - X_{id}) + c_2 r_2 (P_{gd} - X_{id})] \quad (5)$$

$$\chi = \frac{2}{\left|2 - l - \sqrt{l^2 - 4l}\right|}, \quad l = c_1 + c_2, \quad l > 4 \quad (6)$$

Wherein, in the CPSO algorithm, Clerc set $l = 4.1$, $c_1 = c_2 = 2.05$, the contraction factor χ is 0.729, and the two coefficients are $0.729 * 2.05 r_i = 1.49445 r_i$, equivalent to all items in the new formula of speed updating multiplied by weighting factor, with considering the separate function of various items; although the algorithm is improved in terms of convergence, the algorithm must be constrained in advance according to certain functions; otherwise, under given iterations, it is very difficult to find global optimum, and application scope is limited.

3.2 Improved PSO

How to select appropriate parameters to make the algorithm with better convergence is the research hotspot and difficulty. In the early period of the iteration, we expect to get bigger inertia weight to make the algorithm preserve the stronger global search ability, whereas in the late period of iteration, we expect to get smaller inertia weight, which is advantageous for algorithm to conduct more accurate local search. Literature [6] shows that by setting the constraint factors, it can get better convergence, but in fact, we expect particle's individual optimal position can update faster under the condition of the global optimal position, so as to make the algorithm's global search ability and local search ability in equilibrium, and achieve better optimization effect. Integrating the advantages and disadvantages of SPSO and CPSO, this paper proposes a new parameter adjustment strategy named IPSO, and the velocity updating formula is matched with different weight factors. Its expression is as follows:

$$\begin{aligned} V_{id}(k+1) = \omega V_{id}(k) + \ & \omega\chi c_1 r_1 (P_{id} - X_{id}(k)) \\ & + \chi c_2 r_2 (P_{gd} - X_{id}(k)) \end{aligned} \quad (7)$$

Wherein, the first part of the formula (7) is the adjustment for the current speed, and the following two parts are the adjustment for the individual and the global position of the particles. In the first part of the formula, SPSO algorithm's inertia weight factor is kept, which can make particle search speed faster; in the third part, CPSO algorithm's constraint factor is kept, which can ensure that the particle has better global optimization performance. The second part integrates the essences of

SPSO algorithm and CPSO algorithm, weighting factor is that the inertia weight is multiplied by constraint factor, which can reach the purpose of the constraints, and ensure that the individual particles update with acceleration according to the previous speed, so as to achieve the balance of individual optimization and global optimization.

3.3 Convergence Analysis of Algorithm

Theorem 1 (convergence performance of PSO) when $0<\omega<1$, if $0<r_1+r_2<2(\omega+1)$, the standard PSO is in convergence.

Making $\gamma_1=\omega\chi c_1 r_1$, $\gamma_2=\chi c_2 r_2$, the formula （7）can be changed into

$$V_{id}(k+1)=\omega V_{id}(k)+\gamma_1(P_{id}-X_{id}(k))+\gamma_2(P_{gd}-X_{id}(k)) \quad (8)$$

Making $\gamma=\gamma_1+\gamma_2$, $\eta=\frac{\gamma_1 P_{id}+\gamma_2 P_{gd}}{\gamma}$, the formula （8）can be expressed as

$$V_{id}(k+1)=\omega V_{id}(k)+\gamma(\eta-X_{id}(k)) \quad (9)$$

Making $Y_{id}(k)=\eta-X_{id}(k)$, then

$$V_{id}(k+1)=\omega V_{id}(k)+\gamma Y_{id}(k) \quad (10)$$

$$Y_{id}(k+1)=-\omega V_{id}(k)+(1-\gamma)Y_{id}(k) \quad (11)$$

The form of constituting matrix

$$\begin{bmatrix} V_{id}(k+1) \\ Y_{id}(k+1) \end{bmatrix}=\begin{bmatrix} \omega & \gamma \\ -\omega & 1-\gamma \end{bmatrix}\begin{bmatrix} V_{id}(k) \\ Y_{id}(k) \end{bmatrix} \quad (12)$$

Setting $P_{id}(k)=\begin{bmatrix} V_{id}(k) \\ Y_{id}(k) \end{bmatrix}$, $A=\begin{bmatrix} \omega & \gamma \\ -\omega & 1-\gamma \end{bmatrix}$, the above formula can be changed into

$$P_{id}(k+1)=AP_{id}(k) \quad (13)$$

wherein the eigenvalue of matrix A is:

$$\lambda_{1,2}=\frac{\omega+1-\gamma\pm\sqrt{(\omega+1-\gamma)^2-4\omega}}{2} \quad (14)$$

Thus, it can get

$$\begin{bmatrix} V_{id}(k) \\ Y_{id}(k) \end{bmatrix}=Q\begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}Q^{-1}\begin{bmatrix} V_{id}(k) \\ Y_{id}(k) \end{bmatrix} \quad (15)$$

If and only if $|\lambda_{1,2}|<1$, the formula is in convergence, $\left|\frac{\omega+1-\gamma\pm\sqrt{(\omega+1-\gamma)^2-4\omega}}{2}\right|<1$ can be solved to get

$$0<\gamma<2\omega+2 \quad (16)$$

$$0<\omega<1 \quad (17)$$

Thus, the improved algorithm is in convergence.

3.4 Algorithm Realization

Step1: initialization parameter. Set the value range of inertia weight ω, random factors r1 and r2, the maximum number of iterations K_{max}, initial velocity V_{id}, the space dimension D, and randomly generate m particles in the defined search space to constitute the initial population.

Step2: calculate the fitness function value of each particle in population.

Step3: according to the formula (7), update flying speed of the particles; according to the updated speed and the formula (2), update each particle's position, resulting in a new population, implement individual extremum and group extremum update.

Step4: determine whether to meet the termination conditions, if satisfied, the algorithm ends, output the optimal solution; otherwise turn to step2, terminate until reaching the maximum number of iterations.

4 Simulation experiment and performance analysis

4.1 Test Function

To test the performance of the proposed algorithm in this paper, three benchmark functions that are usually used to test the optimization algorithm performance are selected for the test:

$$f_1(x) = \sum_{i=1}^{n} x_i^2 \tag{18}$$

$$f_2(x) = \sum_{i=1}^{n-1} [100(x_{i+1} - x_i^2)^2 + (x_i - 1)^2] \tag{19}$$

$$f_3(x) = \sum_{i=1}^{n} [x_i^2 - 10\cos(2\pi x_i) + 10] \tag{20}$$

Wherein, $f_1(x)$ is called *Sphere* function; it is nonlinear symmetric unimodal function, easy to implement; it is mainly used for testing algorithm's optimization precision. $f_2(x)$ is called *Rosenbrock* function; it is pathological quadratic function with difficulty in minimization, mainly used for performance evaluation of optimization algorithm. $f_3(x)$ is called *Rastrigrin* function; it is a typical complex multimodal function with a large number of local optimal points, easily plunged into local optimum, and is mainly used to test the local optimization ability of the algorithm. The search space and dimension of each test function is as shown in table 1.

Table 1. Search space and dimension number

Name of function	Search space	Dimension number	Optimal solution
$f_1(x)$	[-100,100]	10~30	0
$f_2(x)$	[-30,30]	10~30	0
$f_3(x)$	[-5.12,5.12]	10~30	0

4.2 Simulation Result and Analysis

The particle population size respectively takes 20, 40, 60; the maximum number of iterations in respectively 300,600, 900 times when the number of dimension is 10, 20, 30; inertia weight ω_{start} = 0.9, ω_{end} = 0.4; each benchmark test function is independent in SPSO algorithm, CPSO algorithm, and this algorithm runs 100 times, the optimal value of the average of the minimum fitness function is taken as evaluation standard. As shown in table 2 to table 4, they are the test results of each benchmark function in these three algorithms.

Table 2. Comparison of average optimal solution of *Sphere* function

size	dimension number	iterations	SPSO	CPSO	IPSO
20	10	300	1.5466e-004	2.1995e-004	1.3145e-004
	20	600	0.4180	0.7180	0.0029
	30	900	0.0166	1.1177	0.0091
40	10	300	3.7251e-005	9.2265e-005	1.3374e-005
	20	600	4.5622e-004	0.0159	0.1762
	30	900	0.0017	0.5055	0.0011
60	10	300	1.3011e-005	1.7066e-005	8.2671e-006
	20	600	1.0490e-005	0.0151	0.0370
	30	900	9.2004e-004	0.1038	7.0530e-004

Table 3. Comparison of average optimal solution of *Rosenbrock* function

size	dimension number	iterations	SPSO	CPSO	IPSO
20	10	300	7.4032	9.8691	7.9442
	20	600	17.0658	51.9569	18.6352
	30	900	25.1723	139.8640	23.7788
40	10	300	5.2178	6.6481	3.2203
	20	600	17.9740	26.8572	18.0530
	30	900	22.6215	78.6343	25.1159
60	10	300	2.5596	9.2300	3.8288
	20	600	18.5236	18.9232	16.2113
	30	900	23.5005	68.9229	22.3700

Table 4. Comparison of average optimal solution of *Rastrigrin* function

size	dimension number	iterations	SPSO	CPSO	IPSO
20	10	300	7.9982	10.0003	3.1124
	20	600	14.3013	22.7476	12.7407
	30	900	19.7945	86.0272	8.9303
40	10	300	5.0172	6.9648	4.9806
	20	600	14.9904	16.1784	13.9612

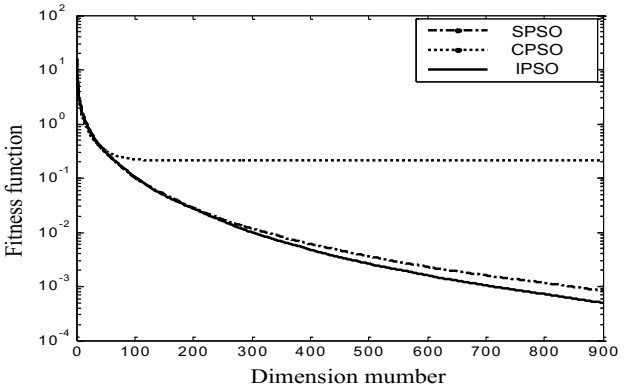
	30	900	14.1550	41.3476	20.1213
	10	300	4.9767	4.9748	1.9910
60	20	600	10.9693	13.1550	13.9743
	30	900	12.0231	35.4069	11.0401

By data contrast in table 2, in the optimization of unimodal function $f_1(x)$, when the population size is 20, space dimension is smaller and the average optimal solution of three algorithms is close to the optimal solution in theory; with the increase of space dimension and the number of iterations, optimization difficulty increases accordingly; this algorithm gets the value closer to the optimal value in theory under the same space dimension and the number of iterations; as the population size increases to 60, the optimal solution of the algorithm is one order of magnitudes lower than SPSO algorithm and CPSO algorithm, proving that the algorithm has better optimization precision in solving multi-dimensional function.

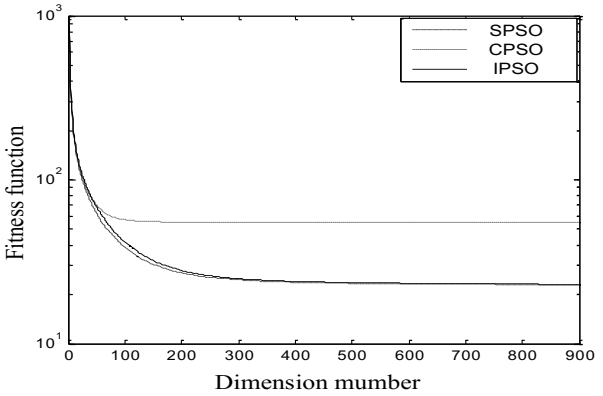
By data contrast in table 3, although $f_2(x)$ is pathological quadratic function with extreme difficulty in minimization, when the space dimension is smaller, the algorithm's optimum result is closer to the optimal value in theory than SPSO algorithm and CPSO algorithm; when the space dimension increases from 10 to 30, the algorithm also shows better optimization performance; with the increase of population size, optimization precision is better, which proves that the algorithm has better optimization performance in solving multi-dimensional function.

By data contrast in table 4, when spatial dimension is increased, the local optimization ability of the complex multimodal function with a large number of local optimal points and CPSO algorithm becomes weak gradually, and when the population size increases, the optimization result of SPSO algorithm hasn't changed a lot, which proves that it is most likely to fall into local optimum and can't jump out, and the proposed algorithm in this paper obtains better optimization numerical with the increase of population size, which proves the local optimization ability of the proposed algorithm in this paper is stronger in solving multi-dimensional function, with better overall performance.

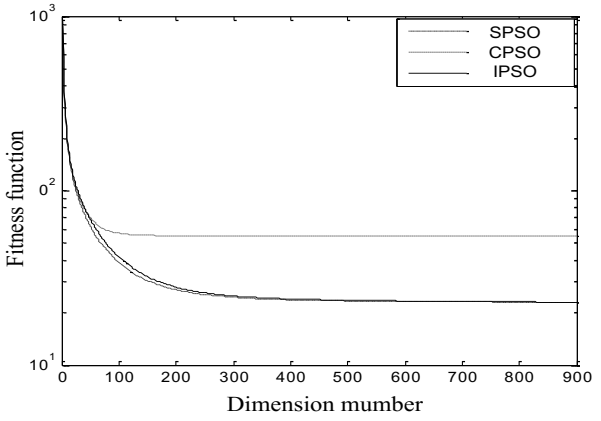
In order to more intuitively reflect the algorithm's optimization effect, SPSO, CPSO, and this algorithm are compared when the particle population size is 60 and the solution space is 30D; the convergence curve of three kinds of algorithm of test function is shown in figure 1. From figure 1, it shows that this algorithm also can obtain better performance of convergence in solving the optimization problem of multi-dimensional function, and effectively avoid the dimension disaster, which once again proves the effectiveness of the proposed algorithm in this paper.



(a) Fitness change of *Sphere* function



(b) Fitness change of *Rosenbrock* function



(c) Fitness change of *Rastrigrin* function

Figure 1. The evolution process of the optimal fitness of tested functions

5 Conclusion

This paper proposes an improved particle swarm optimization of variable parameters; the algorithm IPSO makes improvements on the basis of SPSO algorithm and CPSO algorithm, integrating the essence of the two parts, making the velocity update formula with different weight factor according to its movement characteristics; the algorithm operation is simple, and easy to implement. Through the simulation results of three benchmark test functions, the proposed algorithm in this paper has better optimization precision and ability to execute, and also can show superior convergence performance when solving optimization problem of multi-dimension function.

References

1. Perez R E, Behdinan K. Particle swarm approach for structural design optimization[J]. Computers&Structures,2007,85(19/20):1579-1588.
2. Coelho L S, Sierakowski C A. A software tool for teaching of particle swarm optimization fundamentals[J].Advances in Engineering Software,2008,39(11):877-887.
3. Fan S, Zahara E. A hybrid simplex search and particle swarm optimization for unconstrained optimization[J].European Journal of Operational Research,2007,181(2):527-548.
4. Li X D. Niching without niching paramenters:particle swarm optimization using a ring topology[J].IEEE Transactions on Evolutionary Computation,2010,14(1):150-169.Solitons&Fractals,2008,37(3):698-705.
5. Shi Y, Eberhart R C. A modified particle swarm optimizer. Proceedings of IEEE Enternational Conference on Evolutionary Computation, Anchorage,1998,69-73.
6. Clerc M. The swarm and the queen:Towards a deterministic and adaptive particle swarm optimization. Proceedings of the Congress of Evolutionary Computation, Washington,1999:1951-1957.
7. Jiao B, Lian Z G, Gu X S. A dynamic inertia weight particle swarm optimization algorithm[J].Chaos
8. G Ganesan and Y Li. Cooperative spectrum sensing for cognitive radios under bandwidth constraints[C]. Proceedings of the Wireless Communications and Networking Conference.Jun.2007:1-5.
9. Trelea I C. The particle swarm-explosion, stability, and convergence in a multidimensional complex space: optimization algorithm: Convergence analysis and parameter selection[J]; Information Process Letters,2003,85(6):317-325.

An Access Control Architecture for Protecting Health Information Systems

Angelo Esposito[1,2], Mario Sicuranza[1], Mario Ciampi[1]
[1] Institute for High Performance Computing and Networking
of the Italian National Research Council, Naples, Italy
[2] Department of Engineering, University of Naples "Parthenope", Naples, Italy
angelo.esposito@na.icar.cnr.it, mario.sicuranza@na.icar.cnr.it, mario.ciampi@na.icar.cnr.it

Abstract. The enormous benefits that Health Information Systems (HISs) can offer in terms of quality of care and reduction in costs have led many organizations to develop such systems in their domain. Many national and international organizations have developed their HISs in according to their needs, financial availability and organizational resources (such as technology infrastructure, number of involved structures, etc.), without taking into account the possibility of communicating with other systems satisfying common security policies for distributed authorization. For this reason, the solutions are not interoperable with each other. The main cause of the lack of interoperability is the development of "no open architectures" for communication with other systems and the adoption of different technologies. This paper illustrates a technological architecture based on a set of interoperability services to enable secure communication among heterogeneous HISs. In order to protect the interoperability services, having the aim of invoking services of local HISs, an appropriate access control model is part of the proposed architecture. This Access Control Architecture described in this paper allows different HISs to interoperate each other, ensuring the protection of interoperability services among different HIS systems through the integration of the XACML architecture with the HL7 PASS services. The main architectural components needed to perform the security checks established among heterogeneous HIS are shown in detail. Finally, the use of the architecture in the Italian context is shown.

1 Introduction

The use of Information and Communication Technologies (ICT) in healthcare has resulted in a considerable development of Health Information Systems (HISs) in order to both: i) enhance the quality of care; and ii) reduce the costs. The most important example of HIS is the Electronic Health Record (EHR), which can be developed at international, national, or regional level. The International Organization for Standardization (ISO) defines EHR as a "repository of patient data in digital form, stored and exchanged securely, and accessible by multiple authorized users". It contains retrospective, concurrent and prospective information and its primary purpose is to support continuing, efficient and quality integrated healthcare [1]. At national and international level, many healthcare organizations and institutions have developed their EHR in an autonomous way, with the consequence of a proliferation of hetero-

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_4

geneous systems, which are not able to communicate with other systems. The lack of interoperability is due to the use of proprietary representation models and of different technologies for the development of the systems.

In order to improve the provision of health services, it is very important that different systems are able to exchange information each other. So, with the purpose of enabling different HISs to homogeneously interpret exchanged health information, it is necessary to ensure full interoperability among the systems for the execution of shared business processes in which such systems are involved. Interoperability is the ability that two or more systems or components exchange information and use information that has been exchanged [20].

In the literature four levels of interoperability have been defined [2]:

- *Technical interoperability:* the systems share the communication protocols making possible, for example, the exchange of bytes among them.
- *Syntactic interoperability:* the systems are capable of communicating and exchanging data through the sharing of data formats.
- *Semantic interoperability:* the systems are able to exchange data and interpret the information exchanged in the same way; the systems can coordinate with each other.
- *Organizations/services interoperability:* This level is reached when business processes are shared between the different systems. The systems are able in this way to collaborate and cooperate in the best conditions.

The use of existing standards to meet interoperability is still a challenge in healthcare due to several reasons. The adoption of health interoperability standards is not trivial since it requires high effort, technical expertise as well as clinical domain knowledge [21]. To this aim, architectures based on the SOA pattern is often used, developed with the Web Services technology. These architectures permit indeed to design services able to easy making legacy systems to communicate each other. In order to protect such services, a dedicated service protection system has to be provided. A well-known, simple, and effective security approach is the Access Control (AC) model. Numerous AC techniques have been proposed. In [23], an access control scheme that addresses the lack of context-aware models for access control is proposed: it proposes an extended, trust-enhanced version of XML-based Role Based Access Control (X-RBAC) framework for Web-Services. Bhatia et al. [18] propose a framework that addresses client's privacy concerns in the context of Web Services environment, which provides automated reasoning techniques for matching the service provider's privacy policies for compliance with the client's privacy preferences. In [19], different extended role-based access control schemes for Cloud Computing environment are presented. The work in [22] shows the use of Model Driven techniques for e-Health systems with formal verification of privacy requirements. Another approach is shown in [17], where a semantic-based framework offers a valuable way to enable and support the definition of fine-grained access control policies over different EHRs.These techniques are designed to protect resources in specific contexts, resulting not adequate for being used in heterogeneous environments that require flexible and dynamic characteristics.

This paper illustrates the architecture intended to protect interoperability services that interact with clinical resources maintained by heterogeneous EHR systems. The architecture is based on i) appropriate services sharing the same interfaces conform to health informatics standards, and ii) an access control module able to suitably protect these services. The interoperability levels are covered in this way. In order to ensure technical interoperability, the services use common interfaces (implemented through the Web Services technology) able to exchange messages based on standard transactions conform to the IHE XDS profile. The syntactic and semantic interoperability is overcome by the use of the HL7 CDA Rel. 2 standard [11] and common coding systems, such as LOINC [14] and ICD-9 CM [15]. The access to the services provided by each EHR system has to meet a number of security requirements. The security requirements have to be well indicated in common specifications, even if the way to implement such constraints may be different.

The rest of the paper is organized in four sections. Section 2 illustrates the proposed access control architecture for the interoperability among different EHR systems. Section 3 presents the Italian context for the realization of EHR systems according to security regulations. Section 4 describes the process used by the proposed access control architecture, showing the details of the interaction of the different components of the architecture. Finally, Section 5 concludes the paper with some final remarks.

2 Access Control Architecture for EHR interoperability systems

This section describes the defined Access Control Architecture that allows different HIS systems to interoperate each other, by guarantying the protection of the interoperability services. The architecture consists of different modules that are illustrated and described below.

To achieve protection in a system of systems, the main issue is related to federated authentication and access control [4]. For this reason, an Attribute Based Access Control model [13] in order to allow only to the authorized users the access to the interoperability services and data has been defined. The attributes are associated both with the request made by the user and with the requested resource. The authorization access to documents and services is based on the evaluation of these attributes.

The necessary attributes for authorized access are classified in attributes related to the request, and those related to resources. The attributes related to the request are: i) role of the user who makes the request (a list of admissible roles in the system of systems is defined), ii) the purpose of use (a list of admissible purpose of use in the system of systems is defined), iii) the locality and the organization identification, etc. The attributes related to resources are for example the list of roles to which the access is permitted, the level of confidentiality, etc.

The scenario for the authorization phase in the proposed architecture provides that each user logs-in to own EHR system, which provides a set of attributes to the user. These attributes are asserted by EHR system and used for the interoperability service request.

The Access Control Architecture proposed is modular and designed to be easy to use in different system contexts. In the definition of the architecture has been considered standard requirements to protect resources in a distributed healthcare environment as described in "Privacy, Access and Security Services Security Labeling Service" in [10]. The architecture also has been defined following the standard OASIS eXtensible Access Control Markup Language (XACML) [5]. Figure 1 shows the proposed access control architecture. It consists of five modules, which we describe below. The architecture allows the protection of services (interoperability), intercepting the request messages, which are assessed and intercepting the response messages that are appropriately filtered. The service requests are XML messages forwarded to services. The messages are made by header and body, and in the header of the request message there are security assertions, containing relevant information about requestor, as indicated for the XACML architecture.

1. **Service Requestor** is an architecture external actor; it represents an interoperability service request from a user of the federation (system of systems).
2. **Access Control Interceptor,** this module allows intercepting all services requests and forwarding them to the other modules of the access control, in a transparent manner. The module manages the interoperability service requests and the responses, filtering the results according to the security label and the obligations received by the architecture modules via the component *Privacy Proxy Component*.
3. **The Privacy Proxy Component** is a sort of HL7 Privacy and Security Protective Service [10] module, that allows filtering of the resources in response, according to the obligations received by *Policy Enforcement Point (PEP)* responding to the service requestor.
4. **Policy Enforcement Point (PEP),** this component is inherited by XACML architecture, it makes a decision request to the *Policy Decision Point (PDP)* in order to obtain the access decision, and the obligations on the received decision. The *PEP* sends to *Privacy Proxy* Component the decision and the obligations received from the *PDP*.
5. **Policy Decision Point (PDP)**, this component is inherited by XACML architecture, it evaluates access request against authorization policies before issuing access decisions, implements the access control model defined and formalized in [9]. It is composed of several sub-component, through which verifies the message request correctness and it builds the obligations, which depending on the attributes associated with the request.

   The PDP sub-components are:

   - **Verification Signature Assertion Component** extracts the security assertions from the request message and verifies for each security assertions the validity of the digital signature. In the case of the signatures are not valid or not present, the output is DENY via *Obligation Builder Component*. In this case, the service is not called.

- **Verification Structure Assertion Component** receives the assertions (for which the signature has been verified) from the *Verification Signature Assertion Component*. For each assertion, the component checks that the structure is valid and that the assertion contains all needed information. This component calls *Header Component* or *Obligation Builder Component* (last option in case of verification is not passed).
- **Header Component** analyses all the information in the message header and consequently in the assertions. It verifies that all values are consistent, in particular that there is information consistency among assertions. In order to check the information consistency, the component accesses the Policy Stores via the *Policy Administration Point* and builds security labels depending on the analysed information. The defined security labels are stored in Attribute Stores via the *Policy Information Point*.
- **Body Component** analyses all information in the request body message and verifies the information consistency, for this operation (consistent check) the component accesses the Policy Store through the *Policy Administration Point* and generates security labels depending on the information. The defined security labels are stored in the Attributes Store via *Policy Information Point*.
- **Obligation Builder Component** makes the final check for the *PDP*, it compares the security labels defined by *Header Component* and those defined by the *Body Component*, then it builds obligations that the *PDP* will send to the *PEP*. The obligations will allow making filter of the resources providing in response. Only the authorized resources will be made available.

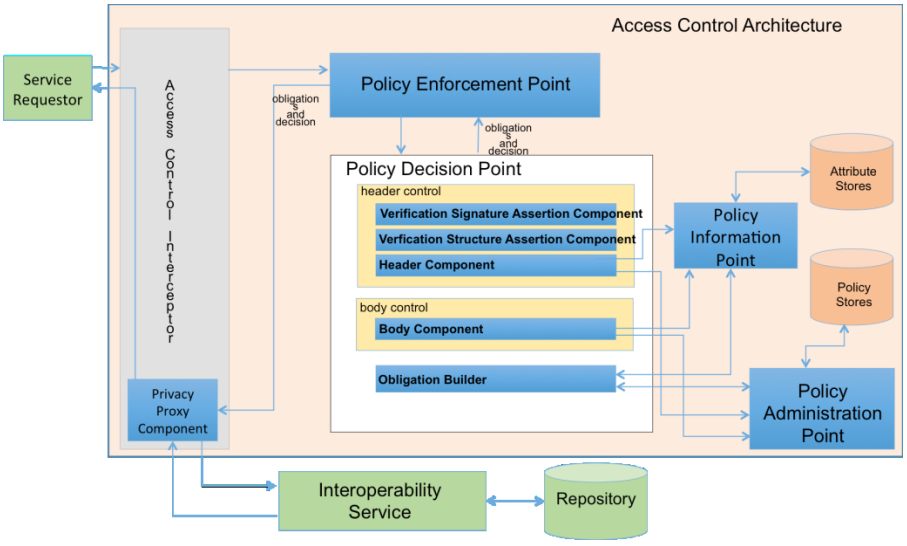


Figure 1 - Architecture for authorizing access to interoperability EHR services

3 Italian interoperability context and the security model

In this section, the solution adopted in Italy for making regional EHR systems able to interoperate each other is shown. In detail, the main interregional processes and the interoperability services to be provided by all regional EHR systems are outlined.

The Italian Constitution provides autonomy to the Regions in the management of the Health; consequently, each regional health organization has implemented or is developing its regional EHR system independently. The Italian government, in order to ensure interoperability among different regional solutions, has defined national specifications, for which each Region has to implement five services in order to satisfy the same business processes.

The regional EHR systems have to make available these services and be able to use the provided by other regions services, creating in this way a system of systems obtained through the federation of all the regional EHR systems. The architectural model undertaken in Italy to implement a nationwide interoperable EHR system is described in [3].

Each regional system acts as one of the following actors on the federation:

- **RDA** (*region of the patient's care*): is the region that manages (through metadata) clinical documents and security policies related to a patient which the region has in charge. The documents management is performed by memorizing specific metadata associated to documents, allowing thus the localization and management of the clinical resources.
- **RCD** (*region containing a document*): is the region in which a specific document has been created and maintained;
- **RDE** (*region of service delivery*): is the region that provides an health service to a patient;
- **RPDA** (*previous region of the patient's care*): is the region that previously has taken in charge a patient, managing his/her clinical documents.

The possible interoperability services invoked by the regional systems are:

1. ***Searching for documents***

This service allows a regional EHR system searching and locating the documents related to a specific patient that meet a set of search criteria. This request has to be forwarded to the RDA. After that RDE requires RDA to search documents, and in this way the RDA returns the list of documents for which the user in the RDE can access (according to security policies established directly by the patient). This service adheres to the standard transaction IHE Registry Stored Query [6].

2. ***Retrieving a document***

This service allows a regional EHR system to retrieve a specific clinical document.
The RDE requires RDA to retrieve a document. If RDA has the document requested, then it corresponds the document to the RCD, returning so the document if the user has access rights. Otherwise (if RDA is different from RCD), RDA forwards the request to RCD and operates as a proxy.
This service adheres to the transaction IHE Retrieve Document Set [7].

3. ***Creating or updating a document***
This service allows a regional EHR system to create or update a document.
RDE transmits to RDA the list of metadata relating to the document created/updated and RDA stores the metadata.
This service conforms to the transaction IHE Register Document Set-b [8].

4. Invalidating a document
This service allows a regional EHR system to request a logical delete of a previously created clinical document. In this case, the logical delete is the invalidation of the metadata associated with the document.
The RCD requires the logical delete of the document to RDA, which performs the elimination of metadata associated to the specified document. This service is conform to the standard transaction IHE Delete Document Set [12].

5. Transferring EHR index
This service allows a regional EHR system (of a new region of the patient's care) to request the transfer of the EHR index. The EHR index comprises all the metadata associated to all clinical documents and security policies of a specific patient. The new RDA requires to RPDA transferring the list of all metadata and privacy policies associated with a patient. The RPDA, after correctness checking, returns the information (metadata and policies) which will be registered in the new RDA. This service adheres to the standard transaction IHE Registry Stored Query Transaction [6].

The proposed Access Control (AC) Architecture is inserted in the described interoperability context so as shown in Figure 2. The AC Architecture allows to protect the interoperability services offered by regional EHR systems. In the federation, the requests for interoperability services are made through SOAP messages (a kind of XML message). All regional EHR systems make a circle of trust (CoT) [24] via Certification Authority (CA) [25]. The CA provides certificates to all regional EHR systems that provide and use the interoperability services on the federation. Each regional system has a digital certificate in X.509 format and a private key used for digital signature of the assertions containing all necessary information (attributes) for the request. The security assertions defined in the federation are:

i) *The identification assertion* that certifies the identity of the patient for which it is requested the resource;

ii) *The attribute assertion,* that certifies the information about the user that submits a request and the type of activities to perform; the assertion is issued by the regional system that intends to use an interoperability service;

iii) *The identification assertion of the RDA*, that certifies the identity of the RDA regional EHR system making the request. This assertion is used for retrieving a document available in RCD.

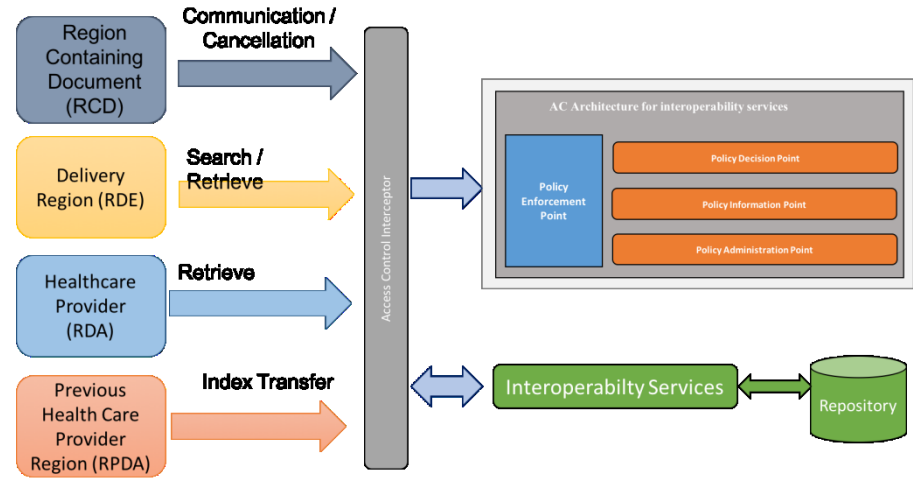


Figure 2 – Access Control Architecture in the business processes for the interoperability

4 Case scenario: the proposed architecture in the Italian context

This section presents how the proposed Access Control Architecture has been used in the Italian context. For each module of the architecture are shown the carried operations and interactions. Let us suppose that a regional EHR system (the RDE) requires an interoperability service to another regional EHR system (the RDA) in the federation. This service is offered and protected by the proposed Access Control Architecture. In this case, the RDE is the service requestor and the RDA provides the interoperability service.
Below, are shown and described the process steps:

1. In the first step, the RDE builds the attribute assertion and identification assertion in which are present a list of attributes related to the request, as:

   i) *User role*, is the role (shared in the federation) of the user that making the request for interoperability service. A user in RDE after log in on its regional system is associated with a specific role at the federation level. This information is in the attribute assertion.

   ii) *Patient identifier* represents, in the federation, the unique identifier of the patient to which requested resource refers. This information is in the *identification assertion*.

   iii) *Purpose of Use* specifies the reason for which the service has been requested, as for example emergency or clinical treatment.

   iv) *Interoperability Service* is the interoperability service that a service requestor requires, and the access control architecture protects.

   v) *OrganizationId and locality* specify the healthcare organization and, therefore the regional system from which the request was made (in this case the specific RDE).

2. The RDE builds the SOAP request message, in which the assertions are inserted, this message is sent to the interoperability service of a specific regional EHR system (the RDA);
3. The request is intercepted by the *Access Control Interceptor Component*, which forwards the SOAP message request to the *Policy Decision Point Component*;
4. The *PDP Component* makes an authorization decision and builds the obligations which are provided in response to the *Policy Enforcement Point Component*, through the use of several sub-components, that are:
    a. **Verification Signature Assertion Component:** it verifies digital signatures of the security assertions. In particular, the component verifies that the certificate has been realised by the trusted Certification Authority and validates the XML digital signature on SAML assertions.
    b. **Verification Structure Assertion Component**: checks the structure of SAML assertions.
    c. **Header Component**: collects the information in the SOAP message header, and checks all data related to the request. The component takes the "patient identifier" data from the identification assertion. This information is stored as security label in the Attribute Stores for the later consistency check. It picks up also the list of assistance regions for the patient and stores them in a security label.

        The component picks up on the attribute assertion the following information:

        i) *Patient identifier,* that is the id of the patient to which refers the requested resource. *Header component* performs the consistency checking between this value and the value obtained from the *identification assertion;*

        ii) *Organizationid* and *locality*, for which checks the coherence with the requested service and patient's RDA;

        iii) *User role*, this value must be present among "user roles known" in the federation and the indicated role has to be authorized for the interoperability service request. The component retrieves the security policy from the Policy Stores via *PAP Component* for the check;

        iv) *Purpose of use*, which must be compatible with the interoperability service requested and User role;

        v) *Document type*, which is stored as security label in the Attribute Stores;

        vi) *Clinic taking charge*, indicating the patient consent. If it is equal to false, the access is possible only in case of emergency;

vii) *Action*, which has to be consistent with the interoperability service requested, the component retrieves this information from the Policy stores through the *PAP Component*. Figure 3 shows the set of all the features that *Header Component* realizes.

d. **Body Component**: picks up all the information from the body SOAP message request, and provides the following functions: i) checks if the consents provided by the patient (consultation consent and supply consent) are in accordance with the service of interoperability, ii) stores all relevant information as security label in the Store Attributes, for example in the case of search for documents, the component stores the patient identification and document type.

e. **Obligation Builder Component**: picks up all security label stored in Attribute stores for the access control. In addition, the component recovers all security policies by querying the policy stores via *PAP*; Performs a series of checks for consistency between the security label stored by *Header Component* and those stored by the *Body Component*. For example, it checks that the patient ID indicated in the *identification assertion* (and *attribute assertion*) has the same value of that patient identifier in the SOAP message body request. Another check is on the patient assistance region (the information is taken from the security label), in fact on value of patient assistance region depends if the interoperability service can be served. Other control regards the type of resource required. In particular, the document type requested has to be the same of the type stored in the security label by **Header Component**. According to these checks, the component builds obligations and the final decision of the *PDP*.

5. The PEP forwards obligations and the decision to *Privacy Proxy Component*.
6. The Privacy Proxy Component (PPC) sends the SOAP request message to the interoperability service.
7. After receiving the response from interoperability service, the PPC builds a message response, filtering the message received in according to the obligations.
8. The PPC returns message built to the service requestor via Access Control Interceptor.

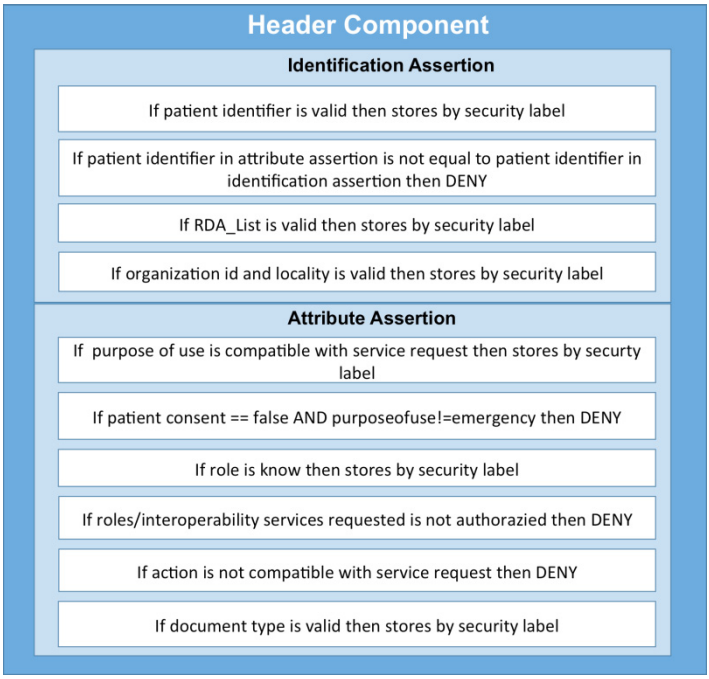


Figure 3 – Header Component

5 CONCLUSION AND FUTURE WORK

In this paper the process and the Access Control Architecture defined for the interoperability of EHR systems has been described. The proposed architecture can be used to protect Health Information Systems that have to interoperate with each other by exchanging health information.

The proposed architecture is composed by: i) a set of interoperability services, having the aim to simplify the interaction from one HIS to another in obtaining health information; ii) an access control module based on the integration of XACML architecture and HL7 PASS services, for authorizing the access to services and health information. In order to validate the proposed architecture, a prototypal system for realizing an interoperable EHR system conform to the Italian guidelines and technical implementation has been developed. The prototype is able to control the access to the services, filtering clinical information through the definition and the use of obligations, according to specific user that makes request.

As future work, we intend to generalize the proposed architecture in order to realize a Security as a Service to be used in Service Oriented Architectures.

ACKNOWLEDGEMENTS

The work presented in this paper has been partially supported by the joint project between the Agency for Digital Italy and the National Research Council of Italy titled "Realization of services of the national interoperability infrastructure for the Electronic Health Record", det. AgID 61/2015.

References

[1] ISO/TR 20514:2005, Health informatics -- Electronic health record -- Definition, scope and context

[2] D. Kalra and B.G. Blobel, "Semantic interoperability of EHR systems", Stud Health Technol Inform. 2007;127:231-45.

[3] M.T. Chiaravalloti, M. Ciampi, E. Pasceri, M. Sicuranza, G. De Pietro, and R. Guarasci, "A model for realizing interoperable EHR systems in Italy", 15th International HL7 Interoperability Conference (IHIC 2015), Prague, Czech Republic

[4] M. Deng, R. Scandariato, D. de Cock; B. Preneel and W. Joosen, "Identity in federated electronic healthcare," in Wireless Days, 2008. WD '08. 1st IFIP , vol., no., pp.1-5, 24-27 Nov. 2008,doi: 10.1109/WD.2008.4812919

[5] OASIS eXtensible Access Control Markup Language (XACML), online at https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=xacml (Access date: 30 January 2016)

[6] IHE IT Infrastructure Technical Framework - Registry Stored Query Transaction for XDS Profile [ITI 18]

[7] IHE IT Infrastructure Technical - Retrieve Document Set for XDS Profile [ITI-43].

[8] IHE IT Infrastructure Technical - Register Document Set-b for XDS Profile [ITI-42]

[9] M. Sicuranza, A. Esposito and M. Ciampi "A View-Based Access Control Model for EHR Systems", Intelligent Distributed Computing VIII p. 443-452,2015 Springer International Publishing

[10] HL7 Version 3 Standard: Privacy, Access and Security Services Security Labeling Service (SLS)

[11] HL7 Version 3 Clinical Document Architecture (CDA) Release 2, https://www.hl7.org/implement/standards/product brief.cfm?product id=7 (Access date: 30 August 2016)

[12] IHE IT Infrastructure Technical - Delete Document Set [ITI-62]

[13] V. C. Hu, D. R. Kuhn and D. F. Ferraiolo, "Attribute-Based Access Control", in Computer, vol. 48, no. 2, pp. 85-88, Feb. 2015.

[14] Logical Observation Identifiers Names and Codes - https://loinc.org (Access date: 30 August 2016)

[15] The International Classification of Diseases, 9th Revision, Clinical Modification - http://www.salute.gov.it/portale/temi/p2_6.jsp?id=1278&area=ricoveriOspedalieri&menu=classificazione (Access date: 30 August 2016)

[16] Integrating the Healthcare Enterprise (IHE) - https://www.ihe.net (Access date: 30 August 2016)

[17] Flora Amato, Giuseppe De Pietro, Massimo Esposito, Nicola Mazzocca, An integrated framework for securing semi-structured health records, Knowledge-Based Systems, Volume 79, May 2015, Pages 99-117, ISSN 0950-7051.

[18] R. Bhatia and M. Singh, "An Implementation Model for Privacy Aware Access Control in Web Services Environment", Proceedings of International Conference on ICT for Sustainable Development: ICT4SD 2015 Volume 1, pp. 475-484, 2016

[19] Hongjiao Li, Shan Wang, Xiuxia Tian, Weimin Wei and Chaochao Sun "A Survey of Extended Role-Based Access Control in Cloud Computing", Proceedings of the 4th International Conference on Computer Engineering and Networks, pp 821-831, 2015

[20] "IEEE Standard Glossary of Software Engineering Terminology", IEEE Std 610.12-1990.

[21] I. Macía, "Towards a semantic interoperability environment," e-Health Networking, Applications and Services (Healthcom), 2014 IEEE 16th International Conference on, Natal, 2014, pp. 543-548.

[22] Flora Amato and Francesco Moscato. 2015. A model driven approach to data privacy verification in E-Health systems. Trans. Data Privacy 8, 3 (December 2015), 273-296.

[23] R. Bhatti, E. Bertino, A. Ghafoor, "A Trust-Based Context-Aware Access Control Model for Web-Services", Distributed and Parallel Databases, pp 83-105, 2005.

[24] L. Boursas and V. A. Danciu, "Dynamic inter-organizational cooperation setup in Circle-of-Trust environments," NOMS 2008 - 2008 IEEE Network Operations and Management Symposium, Salvador, Bahia, 2008, pp. 113-120.

[25] J. Classen, J. Braun, F. Volk, M. Hollick, J. Buchmann and M. Mühlhäuser, "A Distributed Reputation System for Certification Authority Trust Management," Trustcom/BigDataSE/ISPA, 2015 IEEE, Helsinki, 2015, pp. 1349-1356.

Intelligent management system for small gardens Based on wireless sensor network

Xiao-hui ZENG[1,2], Man-sheng LONG[1,2], Qing LIU[1,2], Xu-an WANG[3], Wen-lang LUO[1,2]*

[1] School of Electronic and Information Engineering, Jinggangshan University, China
[2] Key laboratory of watershed ecology and geographical environment monitoring, NASG, China
[3] Engineering University of CAPF, China
*Corresponding author Email: 8102011wen@163.com

Abstract. An intelligent cultivation management system is proposed. In the system, by using ZigBee wireless sensor network monitoring the temperature, humidity, light, the concentration of CO2, and other environmental factors based on solar power supply. Thus small gardens crop growth conditions are obtained. Through wireless sensor network, the irrigation and fertilization for small gardens crop growth are controlled by management end software, which aims at remote wireless elaborating intelligent management, so the economic benefit is improve.

1 Introduction

Our country is an economic crops country, but in the development of agriculture our country is still faced with many problems and challenges. Most areas in China is mainly small gardens, and mostly suitable agriculture intelligence need further research and development [1,2,3]. For a long time, due to high investment for high productivity, agricultural resources were wasted, and farmland ecological environment deteriorated. If real-time monitoring on the soil and environment can be achieved, those information about plant growth status, plant diseases, and insect pests can be obtained in the process of agricultural production management. Through scientific regulation of growth environment factors, the rational use of agricultural resources can be realized, reducing the human cost, improving the environment, increasing crop yield and quality. Crop growth and production are affected by air temperature and humidity, soil temperature and humidity, wind speed, wind direction, rainfall, and light. Accurate and stable environment parameters monitoring is the important foundation to realize precision agriculture and modern management.

Environmental control is by changing the environment factors (such as temperature, humidity, light, CO2 concentration, etc.) to get the best conditions for crop growth. At present, the domestic monitoring and control system mainly adopts cable way to realize information transmission through upper machine and lower machine. This kind of transmission mode makes the signal lines and power lines inside the greenhouse with lower reliability, inconvenient installation and maintenance, which is

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_5

not conducive to mobile devices such as agricultural robot's work. In this way, the expansibility of monitoring system function and flexibility of sensor placement is poor [4,5,6,7].

Facility agriculture has the features of high land utilization, short production cycle, high technical content, and low labor costs. However, there are some problems in the current facility agriculture:

1) Equipment condition bad, lack of the advanced production process control equipment and modern information technology management method;

2) Cucumber in nutrient solution irrigation technology is not popular, and water is wasted too much. Not only low utilization rate of resources is serious, but also the production quality of crop is directly affected. Thus, productivity per unit area is low. At present, the enclosed cultivation way in Europe and Japan has been widely used. The European Union countries have irrigated that greenhouse production must use enclosed cultivation system, and the Netherlands is a representative of the European facility agriculture.

According to the characteristics of the small gardens environment and accurate monitoring requirements, we developed a crop precise monitoring system based on solar power and ZigBee wireless sensor network (WSN) technology. The system can overcome the shortcomings of traditional monitoring system, achieving the fundamental purpose of convenient management, increased crop yield, improved crop quality, and added economic benefit.

2 System design and implementation

2.1 System architecture

In recent years, researchers have done a lot of work for the wireless sensor network [8,9,10,11,12]. The whole monitoring system makes up of solar energy power supply system, wireless sensor network, and monitoring and control center of three parts. Based on ZigBee wireless sensor network protocol, our wireless sensor network, the monitoring node, the coordinator node, and routing nodes. Sensor nodes are responsible for collecting crops measurement parameter information which will be transmitted via the wireless router and gateway to the server, and remote monitoring sites connect servers through the Internet. The system adjusts threshold according to circumstances information, and then control information is sent to the control node in the gardens to open or close the electromagnetic valve. Monitoring and control center need to monitor the working status of sensor nodes, and work tasks of node are adjusted in real-time according to the change data from the sensors. In addition, when the crops water requirement or other requirement is in the critical state, the server will also send alarm messages to the user's mobile phone via the GSM network, which will be convenient for the user to manage the crops.

2.2 Hardware design

The solar energy power supply system consists of solar modules, solar controller, battery. The figure 1 shows the solar energy power supply system structure. Solar energy is a green, clean energy, and it is also helpful for energy conservation.

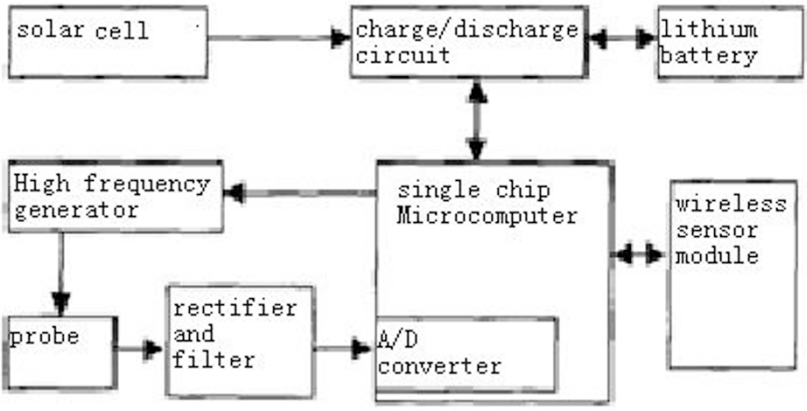


Fig. 1. structure of solar energy power supply system

In order to guarantee the stability of the system to work in the wet weather, battery capacity should be able to meet the normal equipment load for 7 days. In addition, from the economic consideration, as the overall system power consumption is low, which is similar to the smart phone mode, the lithium battery power supply scheme is designed. The outputted Solar panels electricity during the day is stored in the lithium battery, and at the same time it supply equipment load. When the battery power reaches its limit, the system will stop charging, preventing the battery to be overcharged.

In the evening, the lithium battery power supply equipment load, and at the same time the system prevents the reverse charge circuit from charging the solar panels. When the battery power is insufficient, the system cuts off the power supply circuit in time, preventing excessive battery discharge to protect the lithium battery [13,14,15,16]. The I2C bus digital temperature and humidity sensor is adopted as air temperature and humidity sensor, and their features are small size, low energy consumption, two line digital interface, temperature range for -40℃ to 85℃, relative humidity range from 0 ~ 100%. TSL2550D is chosen as light intensity sensor, and its

power consumption can meet the requirements of low power wireless sensor system design; moreover, its bus is also easy to combine with the TI company products.

This system chooses the TI company wireless products, which is a system on a chip based on ZigBee wireless sensor network/IEEE802.15.4 solution. The TI company wireless product is the kind of low power consumption communication chip complied with the standard of ZigBee, which is fast arousal and searches external devices, makes sensor nodes to be dormant at more time and save power consumption. Its parameters of the channel frequency and power consumption can be flexibly set [17,18].

The sensor nodes of wireless sensor network are distributed in the small garden, the sensor node main parts are the sensor, RF wireless single chip microcomputer, crop control device and its control circuit, etc. Crop control switch can be controlled by the remote server software or client software.The system chooses STH-01 type soil moisture sensor, and its measurement accuracy is 3% or so. Its parameters are 12VDC working voltage, 4~20mA output signal, stable time for 2 s, and around 35mA working current. Its sealing material is completely waterproof, and it can real-time monitor soil moisture to meet the system requirements. The structure of the wireless sensor node is shown in the figure 2.

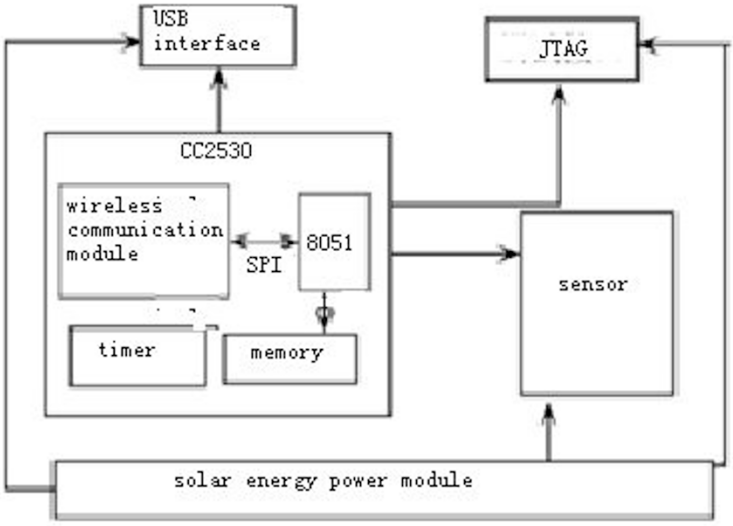


Fig. 2. structure of wireless sensor node

2.3 Software design

Our software includes the parameter setting module, data acquisition module, data analysis module, control output module, and data management module. The software system control hardware to complete data acquisition, operation, compare, and communication with other equipments, and make evaluation conclusion by understanding the status of the crops; what's more, The software system will determine whether to process. The system will save the data, and decision will be given on the basis of the collected data[19, 20, 21, 22].

This system adopts the distributed multi-sensor architecture, and three points data are the basis of information fusion.

When the coordinator receives information, the sensor network address or data from the sensor is judged according to the data of the first identification character. If it is the sensor network address, then the network address is stored in the address list, and the network address will be sent via a serial port to the upper machine to do further processing. If it is data from the sensor, further judgment must be made through the identifier. If the user's request is monitoring the local data, the data is directly displayed to the user, or the user needs the integration data of the whole monitoring area. At this time, the data are stored in the temporary array, and the next sensor data information is collected according to the address table. The sensors' data acquisition for the whole monitoring area of is completed, and finally the control decisions will be made according to the integrated data in the temporary array.

When the system is running, the coordinator and the sensor node power on to initialize, then the network starts and automatically form a self-organized network. The sensor nodes judge whether the required data can be collected in the system after receiving the acquisition data signal. Node dormant in your leisure time, minimize power consumption. Nodes will be in the dormant state at the idle time, which will minimize power consumption.

In this system, the monitoring interface of software can display the network topological structure, signal strength, and gateways according to operation collection. It can also show the curve of soil moisture, temperature, and light sensor data value, and distinguish by different colors. After data are collected 10 times, the calculated average is displayed in the interface. All data can be saved in text format and imported to the interface to display in the curve mode, and the data can be conveniently searched in the future. The software can set the acquisition step length, store interval and execution method.

In order to avoid unnecessary operation and too often wake behavior, the sensor nodes can adjust their acquisition interval time according to its environmental information. The sensor nodes can selectively update the data to the management end, thus saving the nodes energy, and meet the real-time requirement for the environmental information. In this system, we design a sensor adjustment method based on the difference of the sensor data. For example, the sensor node will compare the measurement value of current temperature with the measured value of the historical temperature, if the difference is in the error range of the system, the node will discard the current measured value, giving up updating the current temperature data to the management side, and enter a dormant state directly.

3. Conclusion

After the whole system is installed, the environment information can be smoothly gathered and transmitted from our wireless sensor network, but the correctness of the collected data must be tested. We compare the weather stations with our sensor network, and the results show that their data are consistent, which indicates our sensor network can run normally.

Acknowledgment

This work is supported in part by the Key laboratory of watershed ecology and geographical environment monitoring (Grant No. WE2015013, WE2015002), the Opening Project of Key Laboratory of Embedded System and Service Computing of Ministry of Education (Grant No. 2010-09), the key technology R&D program of Jiangxi province (Grant No. 20151BBE50056, 2010BGA0090), the Science & Technology Transformation Program for Universities in Jiangxi Province (Grant No.KJLD13068, KJLD13066, KJLD14063).

References

[1]. Xiao X, He Q, Fu Z, et al. Applying CS and WSN methods for improving efficiency of frozen and chilled aquatic products monitoring system in cold chain logistics[J]. Food Control, 2016, 60: 656-666.
[2]. Zhang X, Wen Q, Tian D, et al. PVIDSS: Developing a WSN-based Irrigation Decision Support System (IDSS) for Viticulture in Protected Area, Northern China[J]. Applied Mathematics & Information Sciences, 2015, 9(2): 669.
[3]. Sun H, Li M. Precision Agriculture in China[J]. Precision Agriculture Technology for Crop Farming, 2015: 231.
[4]. Yang C, Yuling S, Zhongyi W, et al. Connectivity of wireless sensor networks in greenhouse for plant growth[J]. International Journal of Agricultural and Biological Engineering, 2016, 9(1): 89-98.
[5]. Jiang S, Wang W, Hu Y, et al. Design of Wireless Monitoring System for Environment Monitoring in Greenhouse Cultivation[C]//Proceedings of the 6th International Asia Conference on Industrial Engineering and Management Innovation. Atlantis Press, 2016: 219-228.
[6]. Liu Q, Jin D, Shen J, et al. A WSN-based prediction model of microclimate in a greenhouse using extreme learning approaches[C]//2016 18th International Conference on Advanced Communication Technology (ICACT). IEEE, 2016: 730-735.
[7]. Liu Y, Han W, Zhang Y, et al. An Internet-of-Things solution for food safety and quality control: A pilot project in China[J]. Journal of Industrial Information Integration, 2016, 3: 1-7.
[8]. Chew C C, Funabiki N, Maruyama W, et al. An extended active access-point selection algorithm for link speed changes in wireless mesh networks[J]. International Journal of Space-Based and Situated Computing, 2014, 4(3-4): 184-193.
[9]. Morreale P, Goncalves A, Silva C. Mobile ad hoc network communication for disaster recovery[J]. International Journal of Space-Based and Situated Computing, 2015, 5(3): 178-186.
[10]. Yerra R V P, Rajalakshmi P. Effect of relay nodes and transmit power on end-to-end delay in multi-hop wireless ad hoc networks[J]. International Journal of Space-Based and Situated Computing 9, 2014, 4(1): 26-38.
[11]. Bahrepour M, Meratnia N, Poel M, et al. Use of wireless sensor networks for distributed event detection in disaster management applications[J]. International Journal of Space-Based and Situated Computing, 2012, 2(1): 58-69.

[12]. Xia J, Yun R, Yu K, et al. A coordinated mechanism for multimode user equipment accessing wireless sensor network[J]. International Journal of Grid and Utility Computing, 2014, 5(1): 1-10.
[13]. Ongaro F, Saggini S, Mattavelli P. Li-ion battery-supercapacitor hybrid storage system for a long lifetime, photovoltaic-based wireless sensor network[J]. IEEE Transactions on Power Electronics, 2012, 27(9): 3944-3952.
[14]. Gutiérrez J, Villa-Medina J F, Nieto-Garibay A, et al. Automated irrigation system using a wireless sensor network and GPRS module[J]. IEEE transactions on instrumentation and measurement, 2014, 63(1): 166-176.
[15]. Aziz A A, Sekercioglu Y A, Fitzpatrick P, et al. A survey on distributed topology control techniques for extending the lifetime of battery powered wireless sensor networks[J]. IEEE communications surveys & tutorials, 2013, 15(1): 121-144.
[16]. Jelicic V, Magno M, Brunelli D, et al. Context-adaptive multimodal wireless sensor network for energy-efficient gas monitoring[J]. IEEE Sensors Journal, 2013, 13(1): 328-338.
[17]. Sran S S, Kaur L, Kaur G, et al. Energy Aware Chain based data aggregation scheme for wireless sensor network[C]//2015 International Conference on Energy Systems and Applications. IEEE, 2015: 113-117.
[18]. Li M, Li Z, Vasilakos A V. A survey on topology control in wireless sensor networks: Taxonomy, comparative study, and open issues[J]. Proceedings of the IEEE, 2013, 101(12): 2538-2557.
[19]. Zeng, X.,M. Li,W. Luo. Research on a remote network monitoring model for large-scale materials manufacturing[C]. in 2011 International Conference on Advanced Materials and Computer Science. 2011. Chengdu, China: Trans Tech Publications Ltd.
[20]. Gonzalez M, Schandy J, Wainstein N, et al. Wireless image-sensor network application for population monitoring of lepidopterous insects pest (moths) in fruit crops[C]//2014 IEEE International Instrumentation and Measurement Technology Conference (I2MTC) Proceedings. IEEE, 2014: 1394-1398.
[21]. Srbinovska M, Gavrovski C, Dimcev V, et al. Environmental parameters monitoring in precision agriculture using wireless sensor networks[J]. Journal of Cleaner Production, 2015, 88: 297-307.
[22]. Abbasi A Z, Islam N, Shaikh Z A. A review of wireless sensors and networks' applications in agriculture[J]. Computer Standards & Interfaces, 2014, 36(2): 263-270.

An AHP Based Study Of Coal-Mine Zero Harm Safety Culture Evaluation

Hongxia Li[1,2,3,4],Hongxi Di[1,2,3,4],Xu An Wang[5]
*[1]School of Management, Xi'an University of Science and Technology, Xi'an 710054,China;
[2]School of Energy Engineering, Xi'an University of Science and Technology, Xi'an 710054, China;
[3]Key Laboratory of Western Mine Exploitation and Hazard Prevention, Ministry of Education, Xi'an University of Science and Technology, Xi'an 710054,China;
[4]Shanxi Provincial Audit Office, Xi'an 710054,China;
[5]Engineering University of CAPF, Xi'an 710054,China
Email:416139865@qq.com

ABSTRACT. At present, the Coal-mine industry calls for a reliable method for evaluating "zero harm" safety cultural construction performance. On the basis of an analysis of various factors affecting "zero harm" safety cultural construction performance, a comprehensive index system for evaluating safety cultural construction performance is built. The analytical hierarchy process (AHP) and the theory of Fuzzy Comprehensive Evaluation (FCE) are employed to build an AHP-FCE Model for coal-mine zero harm safety culture, thus providing a scientific and practical quantitative method for systematic analyses and comprehensive evaluations of coal-mine zero harm safety culture.

This model is used to analyze the "zero harm" safety cultural construction performance of BLA. Analytical results show that the AHP-based "zero harm" safety culture evaluation index system has a great practical applicability. It can be applied to provide a solid foundation for enterprises to improve their strategic goals of "zero harm" safety culture construction", so it should be popularized and widely applied.

Keywords: safety culture, zero harm, AHP, FCE

1 Introduction

Coal-mine safety culture [1] is a new concept of safety management which deepens cognition of safety problems in coal-mine safety production by extending from the natural science to the human science. Coal-mine safety culture puts people first. It is a culture about management and survival, reflecting enterprise workers' pursuit of personal safety and health [2] - [6].

"Zero harm" safety culture with distinct features is proposed based on long-term production practice of coal-mine enterprises in China." 100-1 = 0" is the core concept

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_6

in its safety value construction. The “zero harm” safety culture advocates safety cognitions that “production safety should be placed first”, “one industrial incident will deny all achievements”, and “absolute safety means zero accident rate” [7] - [10].

Zero harm safety culture has rich connotations. The goal of zero harm safety culture is “zero harm”. Safety culture works for safety production and guarantees safety. However, in the daily safety production, the weakness safety consciousness and poor safety quality are greatest hidden troubles for achieving coal-mine safety production. Therefore, in safety cultural construction, it is an important link to cultivate safety awareness of employees, improve their safety quality and implement the “zero harm” safety culture idea concept.

In this paper, a zero harm safety culture model is proposed. On this basis, the AHP is chosen to make a scientific and complete evaluation of zero harm safety culture construction effects, in order to reflect strengths and weaknesses of the coal-mine enterprise in safety culture construction, thereby improving safety culture construction effects and maintaining sustainable development of safety culture of the enterprise.

2. Establishment Of Safety Cultury Evaluation Model For Enterprise

2.1 Establishment Of Evaluation Factors

On the basis of investigating existing research results, considering the actual national condition that China is still in the exploratory stage of zero harm safety culture construction and following the principle of establishing the index system, the author presents four 1st level indictors for the zero harm evaluation index system, including zero harm safety concept culture, safety institution culture, safety behavior culture and safety material culture. Four 1st level indicators get three 2nd level indicators, respectively and a total of 12 2nd level indicators. Zero harm safety concept culture, as one 1st level indicator, covers three 2nd level indicators including zero harm safety values, enterprise zero harm safety concept and enterprise zero harm safety thinking mode. Zero harm safety institution culture includes enterprise safety leadership system, enterprise safety institutional system and enterprise safety organizational structure. Zero harm safety behavior culture gets three 2nd level indicators including enterprise safety production style, safety production decision-making and field operation. Zero harm safety material culture includes enterprise safety material products, enterprise safety material technology and enterprise safety material environment. Each 2nd level indicator gets several evaluation factors and forms some 3rd level indicators. The AHP is employed to build an enterprise zero harm safety culture evaluation model, as shown in Figure 1.

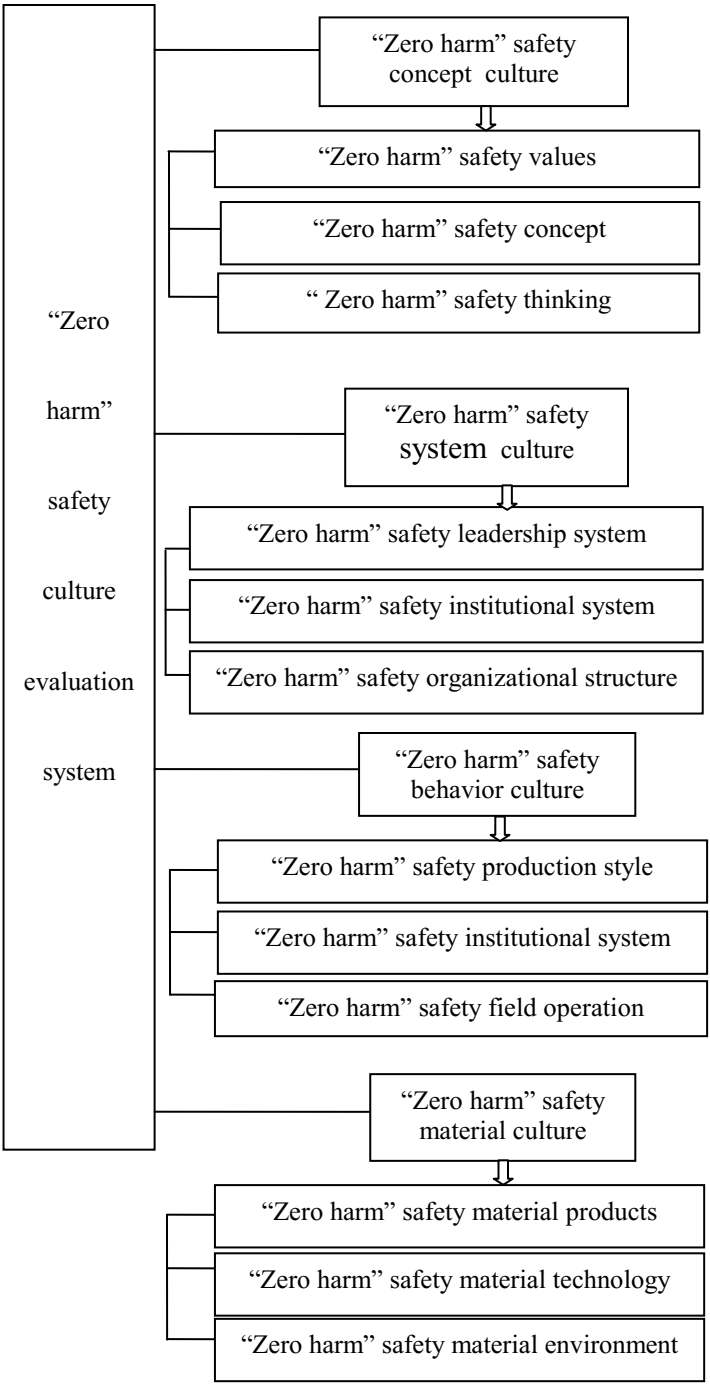


Figure. 1. "Zero harm" safety cultural evaluation index.

Each evaluation factor gets following meanings.

① “Zero harm” safety concept culture

“Zero harm” safety values, “zero harm” safety concept and “zero harm” safety thinking mode constitute “zero harm” safety concept culture.

② “Zero harm” safety institution culture

“Zero harm” safety institution culture mainly discusses the “zero harm” safety leadership system, “zero harm” safety institutional system and “zero harm” safety organizational structure.

③ “Zero harm” safety behavior culture

“Zero harm” safety behavior culture mainly explores the “zero harm” safety production style, “zero harm” safety production decision-making and “zero harm” safety field operation.

④ “Zero harm” safety material culture

“Zero harm” safety material culture mainly investigates the “zero harm” safety material products, “zero harm” safety material technology and “zero harm” safety material environment.

2.2 Determining Weights Of Evaluation Factors

2) Calculation steps of the AHP

① Build a hierarchical model .

② Construct judgment matrixes.

Judgment matrixes are constructed based on comparisons between a factor in the higher layer and factors in its upper layer. The relative importance of each pair of factors in the same layer is compared to determine the corresponding weight. Results of relative importance comparisons are shown through the 1-9 scale method. Each scale gets its corresponding meaning, as shown in Table 1.

Table 1. Meaning of Scale 1 to 9.

Scale	Meaning
1	Indicating that the two factors are equally important
3	Indicating that one factor is slightly more important than the other
5	Indicating that one factor is obviously more important than the other
7	Indicating that one factor is greatly more important than the other
9	Indicating that one factor is extremely more important than the other
2，4，6，8	Between values of two neighboring judgments
Reciprocal	The relative importance scale of the latter to the former when two factors are compared.

The relative importance of each pair of factors is compared to get following results, as shown in Table 2.

Table 2 . Comparative results of relative importance.

	A_1	A_2	…	A_n
A_1	a_{11}	a_{12}	…	a_{1n}
A_2	a_{21}	a_{22}	…	a_{2n}
…	…	…	…	…
A_n	a_{n1}	a_{n2}	…	a_{nn}

Results obtained through comparisons of the relative importance can be used to get a comparison matrix A:A$\{a_{ij}\}$.

Corresponding values in judgment matrixes should meet conditions: $a_{ij}>0$, $a_{ij}=\frac{1}{a_{ji}}$, $a_{ii}=1$.

③ Ranking of factors in the same layer and consistency test .

Judgment matrixes are used to calculate weight vectors of various factors in one layer to factors in the upper layer. In addition, consistency tests are made.

The summation process is chosen to calculate weight vectors, by following steps.

Calculate the weight vector using "Mediation Method" procedure is as follows:

Step 1: Normalize the vectors in each row of A to get $\overline{\omega}_{ij}=\frac{a_{ij}}{\sum_{i=1}^{n}a_{ij}}$ $(j=1,2,...n)$;

Step 2: add weight vectors $\overline{\omega}_{ij}$ of all rows of A to get a summation $\overline{\omega}_i=\sum_{j=1}^{n}\overline{\omega}_{ij}$ $(i=1,2,...n)$;

Step 3: normalize $\overline{\omega}_{ij}$ to get $\omega_i=\frac{\overline{\omega}_i}{\sum_{i=1}^{n}\overline{\omega}_i}$ and $w=(\omega_1,\omega_1,...,\omega_n)^T$ is approximate eigenfunction;

Step 4: calculate the maximum eigenvalue $\lambda_{\max}=\frac{1}{n}\sum_{i=1}^{n}\frac{(Aw)_i}{\omega_i}$.

When CR <0.1, the pairwise comparison matrix A is considered to show a good consistency. The normalized eigenvector of $\lambda_{\max}$ as the maximum eigenvalue of A is taken as a weight vector for the comparison matrix. When CR≥0.1, the wise comparison matrix must be adjusted, until a good consistency is obtained.

3)Weight calculations of safety culture evaluation indictors

Experienced leaders, safety management experts and on-site safety supervision personnel are invited to grade weights of indicators in different layers, combined with actual on-site situation.

(1)Weight calculation of Level 1 “zero harm” safe culture indicators and consistency test

1st level indicator set, $U_n = \{U_1, U_2, U_3, U_4\}$ = {“Zero harm” safety concept culture, “Zero harm” safety institution culture, “Zero harm” safety behavior culture, “Zero harm” safety material culture}. Calculations are shown in Table 3.

Table 3. Weight calculation of Level 1 “zero harm” safe culture indicators and consistency test.

U	U_1	U_2	U_3	U_4	Weight W_i	W_i^0	λ_{mi}
U_1	1	4	3	2	2.213	0.476	4.178
U_2	1/4	1	2	1	0.841	0.181	4.128
U_3	1/3	1/2	1	1/3	0.485	0.105	4.149
U_4	1/2	1	3	1	1.107	0.238	4.074

$$\lambda_{max} = \frac{1}{4}(4.178 + 4.128 + 4.149 + 4.074) = 4.132$$

$$C.I. = \frac{\lambda_{max} - n}{n-1} = \frac{4.132 - 4}{4-1} = 0.044 < 0.1$$

$$C.R. = \frac{C.I.}{R.I.} = \frac{0.044}{0.89} = 0.049 < 0.1$$

Because CR=0.049<0.1, the judgment matrix has a good consistency. Therefore, calculated values of weights can be used.

(2) Weight calculation of Level 2 “zero harm” safe concept culture indicators and consistency test

Level 2 indicator set: “Zero harm” safety concept culture $U_1 = \{U_{11}, U_{12}, U_{13}\}$ = {Enterprise “zero harm” safety values, “zero harm” safety concept, “zero harm” thinking modes}. Calculation results are listed in the following Table 4.

Table 4. Weight calculation of Level 2 “zero harm” safe concept culture indicators and consistency test.

U_2	U_{21}	U_{22}	U_{23}	Weight W_i	W_i^0	λ_{mi}
U_{21}	1	4	1/2	1.260	0.359	3.108
U_{22}	1/4	1	1/3	0.437	0.124	2.953
U_{23}	2	3	1	1.817	0.517	3.108

$$\lambda_{max} = \frac{1}{3}(3.108 + 2.953 + 3.108) = 3.056$$

$$C.I.=\frac{\lambda_{max}-n}{n-1}=\frac{3.056-3}{3-1}=0.028<0.1$$

$$C.R.=\frac{C.I.}{R.I.}=\frac{0.028}{0.52}=0.054<0.1$$

CR=0.0087<0.1 indicates that the judgment matrix passes the consistency test, so calculations of weights can be used.

(3) Weight calculation of Level 2 "zero harm" safe institution culture indicators and consistency test

Level 2 indicator set: "Zero harm" safety institution culture $U_2=\{U_{21},U_{22},U_{23}\}=$ {enterprise "zero harm" safety leadership system, "zero harm" safety institutional system, "zero harm" safety organizational structure}. Calculations are shown in Table 5.

Table 5. Weight calculation of Level 2 "zero harm" safe institution culture indicators and consistency test.

U_1	U_{11}	U_{12}	U_{13}	Weight W_i	W_i^0	λ_{mi}
U_{11}	1	3	2	1.817	0.545	3.018
U_{12}	1/3	1	1	0.693	0.210	3.020
U_{13}	1/2	1	1	0.794	0.240	3.017

$$\lambda_{max}=\frac{1}{3}(3.018+3.020+3.017)=3.018$$

$$C.I.=\frac{\lambda_{max}-n}{n-1}=\frac{3.018-3}{3-1}=0.009<0.1$$

$$C.R.=\frac{C.I.}{R.I.}=\frac{0.009}{0.52}=0.017<0.1$$

CR=0.0087<0.1 indicates that the judgment matrix passes the consistency test, so calculations of weights can be used.

(4) Weight calculation of Level 2 "zero harm" safe behavior culture indicators and consistency test

Level 2 indicator set: "Zero harm" safety institution culture $U_3=\{U_{31},U_{32},U_{33}\}=$ { Enterprise "zero harm" safety production style, "Zero harm" safety production decision-making, field operation}. Calculation results are listed in the following Table 6.

Table 6. Weight calculation of Level 2 "zero harm" safe behavior culture indicators and consistency test.

U_3	U_{31}	U_{32}	U_{33}	Weight	W_i^0	λ_{mi}

				W_i		
U_{31}	1	1/2	1/3	0.550	0.163	3.010
U_{32}	2	1	1/2	1	0.297	3.009
U_{33}	3	2	1	1.817	0.540	3.009

$$\lambda_{max}=\frac{1}{3}(3.010+3.009+3.009)=3.009$$

$$C.I.=\frac{\lambda_{max}-n}{n-1}=\frac{3.009-3}{3-1}=0.0045<0.1$$

$$C.R.=\frac{C.I.}{R.I.}=\frac{0.0045}{0.52}=0.0087<0.1$$

CR=0.0087<0.1 indicates that the judgment matrix passes the consistency test, so calculations of weights can be used.

(5) Weight calculation of Level 2 "zero harm" safe material culture indicators and consistency test

Level 2 indicator set: "Zero harm" safety material culture $U_4=\{U_{41},U_{42},U_{43}\}=\{$ Enterprise "zero harm" safety material products, Enterprise "zero harm" safety material technology, Enterprise "zero harm" safety material environment$\}$. Calculation results are shown in the following Table 6.

Table 7. Weight calculation of Level 2 "zero harm" safe material culture indicators and consistency test.

U_4	U_{41}	U_{42}	U_{43}	Weight W_i	W_i^0	λ_{mi}
U_{41}	1	2	4	2	0.558	3.019
U_{42}	1/2	1	3	1.145	0.320	3.018
U_{43}	1/4	1/3	1	0.437	0.122	3.250

$$\lambda_{max}=\frac{1}{3}(3.019+3.018+3.250)=3.096$$

$$C.I.=\frac{\lambda_{max}-n}{n-1}=\frac{3.096-3}{3-1}=0.048<0.1$$

$$C.R.=\frac{C.I.}{R.I.}=\frac{0.048}{0.52}=0.092<0.1$$

As CR=0.0087<0.1, it indicates that the judgment matrix passes the consistency test, so calculations of weights can be used.

(2) Calculation of evaluation results

1) Calculation methods

On the basis of using the AHP to determine weights of various factors, the FCE is employed to grade "zero harm" safety culture of BLA as a coal-mine enterprise. The FCE is a method using the fuzzy set theory to evaluate systems or programs. It is hard to use traditional mathematical methods to solve problems with various evaluation factors and fuzzy evaluation standards or natural state. However, the FCE can well solve them. Before score assignment, work is done to set a total of six evaluation ranks. The evaluation set is V= $(v_1, v_2, v_3, v_4, v_5, V_6)$ =(quite important, important, general, somewhat important, less important, and quite unimportant). A corresponding score is assigned to each evaluation rank, as shown in Table 8.

Table 8. Score assignments of different evaluation ranks.

Evaluation ranks	Quite high	High	General	A bit low	Low	Quite low
Interval value in Hundredmark system	90 100	80 90	70 80	60 70	60	Below 60
Class midvalue	95	85	75	65	60	30

The AHP is used to determine that weights of Level 2 indicators are W1 = (48, 18, 10, 24); W2 = (55, 21, 24); W3 = (16, 30, 54); W4 = (56, 32, 12). Meanwhile, a total of 10 professors, assistant professors and lecturers specialized in safety engineering from universities and relevant doctoral students and graduate students were gather o form an expert team to mark the zero harm safety culture effects. Concrete grading results are listed in Table 9-1、9-2.

Table 9-1. Marking table for experts.

Evaluation factors		"Zero harm" safety culture U				"Zero harm" institution culture I		
		U1	U2	U3	U4	U21	U22	U23
Evaluation scale	90	4	1	5	3	2	6	2
	80	3	5	2	4	3	2	5
	70	2	4	1	2	5	2	1
	60	1	0	2	1	0	0	1
	60	0	0	0	0	0	0	1
	30	0	0	0	0	0	0	0

Table 9-2. Marking table for experts.

"Zero harm" concept culture **C**			"Zero harm" material culture **M**			"Zero harm" behavior culture **B**		
U11	U12	U13	U41	U42	U43	U31	U32	U33
4	0	3	4	6	5	1	3	6
2	5	4	6	3	2	5	2	2
2	3	1	0	1	2	3	2	1
2	2	1	0	0	1	1	3	0
0	0	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1

For U1, 4 experts consider it to be quite important; 3 experts choose "important"; 2 experts choose "general"; and 1 expert choose "somewhat important". Following grading results can be obtained. U11 = 1/10 = 0.1; U12 = 4/10 = 0.4; U13 = 4/10 = 0.4; U14 = 1/10 = 0.1; U15 = 0, U16 = 0. These values are membership degrees of corresponding evaluation scales. In the same way, membership degrees of other factors can be calculated. Membership degree matrixes of other factors are as follows.

$$I=\begin{bmatrix} 0.2 & 0.3 & 0.5 & 0 & 0 & 0 \\ 0.6 & 0.2 & 0.2 & 0 & 0 & 0 \\ 0.2 & 0.5 & 0.1 & 0.1 & 0.1 & 0 \end{bmatrix} \quad C=\begin{bmatrix} 0.4 & 0.2 & 0.2 & 0.2 & 0 & 0 \\ 0 & 0.5 & 0.3 & 0.2 & 0 & 0 \\ 0.3 & 0.4 & 0.1 & 0.1 & 0.1 & 0 \end{bmatrix}$$

$$M=\begin{bmatrix} 0.4 & 0.6 & 0 & 0 & 0 & 0 \\ 0.6 & 0.3 & 0.1 & 0 & 0 & 0 \\ 0.5 & 0.2 & 0.2 & 0.1 & 0 & 0 \end{bmatrix} \quad B=\begin{bmatrix} 0.1 & 0.5 & 0.3 & 0.1 & 0 & 0 \\ 0.3 & 0.2 & 0.2 & 0.3 & 0 & 0 \\ 0.6 & 0.2 & 0.1 & 0 & 0 & 0.1 \end{bmatrix}$$

The same method can be employed to construct the membership degree matrix of the factor in the target layer

$$U=\begin{bmatrix} 0.4 & 0.3 & 0.2 & 0.1 & 0 & 0 \\ 0.1 & 0.5 & 0.4 & 0 & 0 & 0 \\ 0.5 & 0.2 & 0.1 & 0.2 & 0 & 0 \\ 0.3 & 0.4 & 0.2 & 0.1 & 0 & 0 \end{bmatrix}$$

In accordance with the above-mentioned evaluation steps, the comprehensive evaluation vector of the factor U in the target layer is:

$$T=W*R=(0.48\ \ 0.18\ \ 0.1\ \ 0.24\ \ 0\ \ 0)\begin{bmatrix}0.4 & 0.3 & 0.2 & 0.1 & 0 & 0\\ 0.1 & 0.5 & 0.4 & 0 & 0 & 0\\ 0.5 & 0.2 & 0.1 & 0.2 & 0 & 0\\ 0.3 & 0.4 & 0.2 & 0.1 & 0 & 0\end{bmatrix}=(0.4\ 0.35\ 0.23\ 0.092\ 0\ 0)$$

Normalize T to get the final evaluation result (0.37 0.33 0.21 0.09 00).

Quantify evaluation ranks to calculate the overall score of the "zero harm" safety culture evaluation for BLT.

$$U=(0.37\ 0.33\ \ 0.21\ 0.09\ 0\ 0)\begin{bmatrix}90\\ 80\\ 70\\ 60\\ 60\\ 30\end{bmatrix}=79.8\ (\text{points})$$

On the basis of a calculation of the "zero harm" safety culture level of BLT as a coal-mine enterprise, the calculation result (79.8 points) can help to determine the development stage of "zero harm" safety culture of BLT, in order to provide useful references to BLT to make plans for developing its "zero harm" safety culture.

Table 10 shows the division of "zero harm" safety culture levels of a coal-mine enterprise

Table 10. Level division of "zero harm" safety culture a coal-mine enterprise.

Valuation	"Zero harm" safety culture levels	Development stage	Suggestions
[95,100]	Level 5	Most developed	"Zero harm" safety culture should be preserved;
[85,95]	Level 4	More developed	"Zero harm" safety culture should be perfected;
[75,85]	Level 3	Medium-developed	"Zero harm" safety culture should be further developed;
[60,75]	Level 2	Less developed	"Zero harm" safety culture should be constructed
[0,65]	Level 1	Least developed	"Zero harm" safety culture requires improvement;

4. Conclusion

The AHP is used to determine weights of "zero harm" safety culture of BLT, and the FCE is chosen to mark the safety culture development of BLT. The total points for "zero harm" safety culture of BLT are 79.8.

This score indicates that BLT is at the self-management stage, as an intermediate development stage of “zero harm” safety culture.

BLT does not complete get rid of the passive restrained state. Therefore, BLT should timely build a mechanism to make employees participate in discussion and decision-making of safety issues, so that employees can realize the great importance and value of safety for them, and individual employees and production groups can voluntarily make commitment to and compliance with safety culture. In this way, BLT can fully realize self-management, proceed in an orderly way, and finally move towards the advanced stage of “zero harm” safety culture.

Acknowledgment

The work was supported by National Natural Science Foundation of China (71271169, 71273208).

References

1. Kastenberg W E. Ethics, Risk and Safety Culture. Reflections on the Fukushima Daiichi Nuclear Accident, pp.165-187，2015.
2. MA Yue，FU Gui，ZANG Ya-li． Evaluation index system of enterprise safety culture construction level.China SafetyScience Journal，vol. 24(4) ，pp.124 － 129,2014.
3. Guldenmund F W. The nature of safety culture: a review of theory and research.Safety Science,vol. 34(1),pp:215–257,2000.
4. Liu C, Liu J, Wang J X. Fuzzy Comprehensive Evaluation of Safety Culture in Coal Mining Enterprises. Applied Mechanics & Materials, vol. 724, pp.373-377,2015.
5. QIAN Li-jun LI，Shu-quan． Study on assessment model for aviation safety culture based on rough sets and artificial neuralnetworks.China Safety Science Journal，19(10),pp. 132 － 138,2009.
6. LIU Fang． Study on safety culture evaluation of construction enterprise,Ph.D. thesis, Harbin: Harbin Institute ofTechnology，2010．
7. QIN Bo-tao，LI Zeng-hua． Application of improved AHP method in safety evaluation of mine． Xi′ an University of Science & Technology Journal，22(2),pp. 126 － 129． 2002.
8. Piyatumrong, et al. "A multi-objective approach for high quality tree-based backbones in mobile ad hoc networks." International Journal of Space-Based and Situated Computing 2.2(2012):83-103.
9. MLABao, Sarenna, and T. Fujii. "Learning-based p-persistent CSMA for secondary users of cognitive radio networks." International Journal of Space-Based and Situated Computing 3.2(2013):102-112.
10. Wen, Yean Fu, and C. L. Chang. "Load balancing consideration of both transmission and process responding time for multi-task assignment."International Journal of Space-Based and Situated Computing4.2(2014):100-113.

Analysis of Interval-Valued Reliability of Multi-State System in Consideration of Epistemic Uncertainty

Gang Pan, Chao-xuan Shang, Yu-ying Liang, Jin-yan Cai, Dan-yang Li
Department of Electronic and Optic Engineering
Mechanical Engineering College
050003, Shijiazhuang
Email: pg605067394@163.com

Abstract. Since it is hard to obtain adequate performance data of high-reliability component, resulting in epistemic uncertainty on component' degradation law, system reliability cannot be accurately estimated. For the purpose of accurate estimation of system reliability, assuming the component' performance distribution parameter is the interval parameter, a component' performance distribution model based on interval parameter variable is built, the definition of interval continuous sequences of component' state performance and a computational method of the interval-valued state probability are provided, the traditional universal generating function method is improved, the interval-valued universal generating function and its algorithm are defined, an assessment method of interval-valued reliability of multi-state system in consideration of epistemic uncertainty is proposed, and verification and illustration are conducted with simulation examples. This method overcomes the shortcoming that an inaccurate reliability analysis model of the component is built on account of epistemic uncertainty, which features great universality and engineering application value.

1 Introduction

Systems are only in "normal working" and "complete failure" in the traditional reliability analysis but, for some systems, traditional Binary State System (BSS) assumption is unable to accurately describe some probable states in system operation. These systems have multiple working (or failure) states except "normal working" and "complete failure" or can operate under multiple performance levels, which can be called Multi-State System (MSS) [1]. MSS model can precisely define component' multiple state performance and more flexibly and exactly represent the influence of component' performance changes on system performance and reliability compared with "BSS" model [2].

Research on MSS reliability has been widely concerned after it was raised in 1970s [3, 4]. From the perspective of theoretical methods, [1, 2, 5, 6] references have a detailed description of basic concept, assessment method, and optimal design, etc. of MSS reliability. Ref. [7] has an in-depth research on change and maintenance decisions of incompletely maintained MSS. With regard to engineering application, related theories of MSS reliability have been applied to electric power [8, 9], network [10, 11], and machinery [2, 12, 13], etc.

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_7

Components' state performance and state probability are usually assumed as accurate values and are given in traditional MSS theories. However, material and components update speeds are accelerated along with technological development and the improvement of industrial level, which enables components to present "integrated, intellectualized, and complicated characteristics" and has an increasingly shorter production cycle, and components reliability is tremendously improved in the meantime. It is hard to get accurate and effective components or system failure data for systems constituted by high-reliability components in normal conditions. Therefore, there are many difficulties in estimating system's accurate probability and state performance by gaining accurate failure data. Some scholars promote traditional MSS theories against the above-mentioned problems. Ding et al. [14, 15] have given a general definition and a reliability analysis method of fuzzy MSS. Yan Minqiang et al. [16] have proposed a computational method of fuzzy MSS reliability in consideration of incomplete fault coverage against the problem that MSS performance and probability distribution cannot be accurately gained and incompletely covered in engineering application. Li et al. [17] have analyzed interval-valued reliability of MSS by the use of interval analysis theory and universal generating function. Sebastien et al. [18] have combined random set theory and universal generating function method to analyze MSS reliability of epistemic uncertainty. Liu et al. [19] have analyzed fuzzy MSS reliability by combining fuzzy Markov models and universal generating function. In ref. [20], probability convolution and fuzzy expansion are combined to propose an analytical method on MSS reliability based on mixed universal generating function method against MSS reliability analysis under aleatory uncertainty and epistemic uncertainty. In references [21-23], fuzzy mathematical theory and Bayesian networks are combined to analyze MSS fuzzy reliability from different perspectives.

There are usually two problems in the analysis of reliability of MSS constituted by high-reliability components: (1) epistemic uncertainty on components performance distribution appears because accurate performance degradation data of components cannot be gained, which means parameters are inaccurate; (2) incomplete understanding of performance degradation mechanism of systems or components leads to inaccurate reliability analysis models and even great deviation. In addition, the state performance and state probability of MSS are usually given in the research on MSS reliability with epistemic uncertainty through the analysis of researches of the afore-mentioned scholars, which does not conform to engineering application.

In view of the afore-said insufficiencies, an analytical method of interval-valued reliability of MSS in consideration of epistemic uncertainty is proposed. First, a components' performance distribution model based on interval parameters is built; second, the components' state is divided in the form of interval continuous sequence and components' state interval probability is obtained according to the sequence in order to more accurately describe components' state information and define its performance interval continuous sequence; finally, the traditional universal generating function is improved, definition and algorithm of interval-valued universal generating function are provided, and an analytical model of interval-valued reliability of MSS in consideration of epistemic uncertainty is built.

2 Paper Preparation

2.1 Performance analysis of performance degraded components

In engineering application, since accurate and effective data of high-reliability performance-degraded components cannot be obtained within a short time, based on which, the built performance degradation distribution model is usually inaccurate, there may be great deviation in the analysis result. For that reason, the components' performance distribution parameter can be regarded as an interval variable, then, the performance distribution with the parameter as an interval variable is analyzed before the following assumption is made:

(1) A continuous sequence of interval number is defined, $[x_i]=[\underline{x}_i,\overline{x}_i]\in \mathbf{I}(\mathbf{R})$ is assumed as the interval number, if the sequence is constituted by $[x_1],[x_2],\cdots,[x_n]$ and meets $[x_1]\le[x_2]\le\cdots\le[x_n]$, it can be called a continuous sequence of interval number (interval continuous sequence in short), noted as: $\left[x^{\mathrm{I}}\right]=\left[[x_1],[x_n]\right]$, thereinto, $i=1,2,\cdots,n$.

(2) The components has only one performance parameter x, which corresponds to one performance degradation process, and the degradation process is irreversible;

(3) At any time t, assuming components performance as $x(t)$, which obeys normal distribution with the mean value of $\mu_x(t)$ and variance of $\sigma_x^2(t)$, thereinto, $\mu_x(t)$ and $\sigma_x^2(t)$ are random variables which respectively comply with uniform distribution in $[\mu(t)]$ and $[\sigma^2(t)]$, $x(t)$ is independent identical distribution.

(4) At any given time t, the distribution parameter of components performance $x(t)$ is a random variable which obeys uniform distribution, so the performance distribution function of the components is shown as follows:

$$F(Y)=\int_{-\infty}^{y}\int_{\underline{\mu}(t)}^{\overline{\mu}(t)}\int_{\underline{\sigma}^2(t)}^{\overline{\sigma}^2(t)} f(x|u_x(t),\sigma_x^2(t))h(u_x(t))m(\sigma_x^2(t))du_x d\sigma_x^2 dx \qquad (1)$$

Thereinto,

$$f(x|u_x(t),\sigma_x^2(t))=\frac{1}{\sqrt{2\pi\sigma_x^2(t)}}\exp\left(\frac{-(x-\mu_x(t))^2}{2\sigma_x^2(t)}\right)$$

$$h(u_x(t))=\begin{cases}\dfrac{1}{\overline{\mu}_x(t)-\underline{\mu}_x(t)} & \underline{\mu}_x(t)<u_x<\overline{\mu}_x(t)\\ 0 & else\end{cases}$$

$$m(\sigma_x^2(t))=\begin{cases}\dfrac{1}{\overline{\sigma}^2(t)-\underline{\sigma}^2(t)} & \overline{\sigma}^2(t)<\sigma_x^2(t)<\underline{\sigma}^2(t)\\ 0 & else\end{cases}$$

To ensure that the reproduction of your illustrations is of a reasonable quality, we advise against the use of shading. The contrast should be as pronounced as possible.

If screenshots are necessary, please make sure that you are happy with the print quality before you send the files.

2.2 State probability analysis of performance-degraded components

When a components' state performance is defined, the computational accuracy of the universal generating function is increasingly improved with the increase of components state number but the calculation amount will also be sharply increased, so as to cause "curse of dimensionality" [1]. The state performance is divided in the form of interval continuous sequences in order to remain the state number unchanged, reduce error influence caused by epistemic uncertainty, and improve computational accuracy as much as possible. Assuming the interval continuous sequence of the state performance at t as $\left[g_{i,k_i}^{\mathrm{I}}\right]=\left[[\underline{g}_{i,k_i}],[\overline{g}_{i,k_i}]\right]$ and meets $[\underline{g}_{i,k_i}]=[\underline{x}_{i,ki},\overline{x}_{i,ki}]$, $[\overline{g}_{i,k_i}]=[\underline{y}_{i,ki},\overline{y}_{i,ki}]$, and $\overline{x}_{i,ki}<\underline{y}_{i,ki}$.

According to the analysis of assumption 4 in Section 2.1, the upper and lower boundary s, $\underline{p}(t)$ and $\overline{p}(t)$, of the interval probability of components' state performance at a given t are respectively:

$$\begin{cases}\underline{p}(t)=\min\limits_{y_i(t)\in[\underline{y}_{i,ki},\overline{y}_{i,ki}]}\left(F(y_i(t),u,\sigma)\right)-\max\limits_{x_i(t)\in[\underline{x}_{i,ki},\overline{x}_{i,ki}]}\left(F(x_i(t),u,\sigma)\right)\\ \overline{p}(t)=\max\limits_{y_i(t)\in[\underline{y}_{i,ki},\overline{y}_{i,ki}]}\left(F(y_i(t),u,\sigma)\right)-\min\limits_{x_i(t)\in[\underline{x}_{i,ki},\overline{x}_{i,ki}]}\left(F(x_i(t),u,\sigma)\right)\end{cases} \tag{2}$$

Then, the interval continuous sequence of the components at t is $\left[g_{i,k_i}^{\mathrm{I}}\right]=\left[[\underline{g}_{i,k_i}],[\overline{g}_{i,k_i}]\right]$ and the interval probability of state performance is $[p(t)]=[\underline{p}(t),\overline{p}(t)]$.

3 Algorithm

The family of interval continuous sequences of state performance is divided into upper boundary of interval continuous sequences, mid-value of interval continuous sequences, and lower boundary of interval continuous sequences for discussion and analysis in order to reduce calculation difficulty. Take the upper boundary of interval continuous sequences as an example for analysis.

3.1 Definition of interval-valued universal generating function

The family of interval continuous sequences of state performance of components i, is defined as follows:

$\{\boldsymbol{g}_i^{\mathrm{I}}\}=\left\{\left[g_{i,1}^{\mathrm{I}}\right],\cdots,\left[g_{i,k_i}^{\mathbf{I}}\right],\cdots,\left[\boldsymbol{g}_{i,M_i}^{\mathbf{I}}\right]\right\}$, in which $\left[g_{i,k_i}^{\mathrm{I}}\right]=\left[[\underline{g}_{i,k_i}],[\overline{g}_{i,k_i}]\right]$ represents the k_i interval continuous sequence of state performance of components i, $[\underline{g}_{i,k_i}]=[\underline{g}_{i,k_i}^{L},\underline{g}_{i,k_i}^{U}]$, $[\overline{g}_{i,k_i}]=[\overline{g}_{i,k_i}^{L},\overline{g}_{i,k_i}^{U}]$.

The family of interval continuous sequences of state performance of component i is:

$\{\boldsymbol{p}_i\}=\left\{\left[p_1\right],\cdots,\left[p_{k_i}\right],\cdots,\left[p_{M_i}\right]\right\}$, in which $\left[p_{k_i}\right]=\left[\underline{p}_{k_i},\overline{p}_{k_i}\right]$ represents the interval probability of the k_i state of component i.

The interval-valued universal generating function of component i is defined as follows:

$$u_i(z)=\sum_{k_i=1}^{M_i}\left[p_{i,k_i}\right]\cdot z^{[\overline{g}_{i,k_i}]} \tag{3}$$

Thereinto, $i=1,2,\cdots,n$, $k_i=1,2,\cdots,M_i$.

3.2 Algorithm of interval-valued universal generating function

Before the algorithm of interval-valued universal generating function is defined, interval extension theory is first briefly described as follows:

Definition 1 [24] Assuming that ψ is a real-valued function of n real variables ($x_1,x_2\cdots,x_n$), $f(x_1,x_2\cdots,x_n)$, if the interval-valued function, $\Psi(\mathbf{x}^{\mathrm{I}})=\Psi(x_1^{\mathrm{I}},x_2^{\mathrm{I}}\cdots,x_n^{\mathrm{I}})$, of n interval variables, $\mathbf{x}^{\mathrm{I}}=(x_1^{\mathrm{I}},x_2^{\mathrm{I}}\cdots,x_n^{\mathrm{I}})$, meets

$$\Psi(x_1,x_2\cdots,x_n)=\psi(x_1,x_2\cdots,x_n)$$

Ψ is called the interval extension of ψ.

Definition 2 [24] (natural interval extension) assuming $\psi(\boldsymbol{x})=\psi(x_1,x_2\cdots,x_n)$ is the real-valued function of n real variables, $x_1,x_2\cdots,x_n$, if the interval variable $\boldsymbol{x}^{\mathrm{I}}$ replaces the real variable $\boldsymbol{x}$ and corresponding interval arithmetic replaces real arithmetic in $\psi(x_1,x_2\cdots,x_n)$, the obtained rational interval function $\Psi(\mathbf{x}^{\mathrm{I}})=\Psi(x_1^{\mathrm{I}},x_2^{\mathrm{I}}\cdots,x_n^{\mathrm{I}})$ obviously features inclusion relation monotonicity and Ψ is called natural interval extension of ψ.

Assuming the interval continuous sequences of state performances of component i and i' are respectively $\left[g_{i,k_i}^{\mathrm{I}}\right]$ and $\left[g_{i',k_{i'}}^{\mathrm{I}}\right]$, the interval-valued state probabilities are respectively $\left[p_{i,k_i}\right]$ and $\left[p_{i',k_{i'}}\right]$, in which $\left[g_{i,k_i}^{\mathrm{I}}\right]=\left[[\underline{g}_{i,k_i}],[\overline{g}_{i,k_i}]\right]$, $\left[g_{i',k_{i'}}^{\mathrm{I}}\right]=\left[[\underline{g}_{i',k_{i'}}],[\overline{g}_{i',k_{i'}}]\right]$, $i=1,\cdots,n$, $i'=1,\cdots,n$. $\left[g_{i,k_i}^{\mathrm{I}}\right]=\left[[\underline{g}_{i,k_i}],[\overline{g}_{i,k_i}]\right]$ and $\left[g_{i',k_{i'}}^{\mathrm{I}}\right]=\left[[\underline{g}_{i',k_{i'}}],[\overline{g}_{i',k_{i'}}]\right]$ are applied to the two pieces of component, take the upper boundary of the family of interval continuous sequences of component state performance as an example for illustration, assuming ψ meets $g_{k_s}=\psi(g_{i,k_i},g_{i',k_{i'}})$, it is known from the given function interval extension definition that the interval

extension of $g_{k_s}=\psi(g_{i,k_i},g_{i',k_{i'}})$ is $\left[g_{k_s}\right]=\psi\left([\overline{g}_{i,k_i}],[\overline{g}_{i',k_{i'}}]\right)$. Therefore, the universal generating function of interval vector of $\left[g_{k_s}\right]$ can be acquired through the following operation:

$$\begin{aligned} U(z,t) &= \Omega(u_i(z),u_{i'}(z)) \\ &= \sum_{i=1}^{M_i}\sum_{i'=1}^{M_{i'}}\left[p_{i,k_i}\right]\cdot\left[p_{i',k_{i'}}\right]\cdot z^{\psi\left([\overline{g}_{i,k_i}],[\overline{g}_{i',k_{i'}}]\right)} \end{aligned} \tag{4}$$

The following operator is defined according to system structure characteristics:

(1) Operator δ_1 is defined when $\left[g_{k_s}\right]$ is the sum of $[\overline{g}_{i,k_i}]$ and $[\overline{g}_{i',k_{i'}}]$:

$$\delta_1(U_i(z,t),U_{i'}(z,t))=\sum_{i=1}^{M_i}\sum_{i'=1}^{M_{i'}}\left[p_{i,k_i}\right]\left[p_{i',k_{i'}}\right]\cdot z^{[\overline{g}_{i,k_i}]+[\overline{g}_{i',k_{i'}}]}$$

(2) Operator δ_2 is defined when $\left[g_{k_s}\right]$ is the maximum of $[\overline{g}_{i,k_i}]$ and $[\overline{g}_{i',k_{i'}}]$:

$$\delta_2(U_i(z,t),U_{i'}(z,t))=\sum_{i=1}^{M_i}\sum_{i'=1}^{M_{i'}}\left[p_{i,k_i}\right]\left[p_{i',k_{i'}}\right]\cdot z^{\max\{[\overline{g}_{i,k_i}],[\overline{g}_{i',k_{i'}}]\}}$$

(3) Operator δ_3 is defined when $\left[g_{k_s}\right]$ is the minimum of $[\overline{g}_{i,k_i}]$ and $[\overline{g}_{i',k_{i'}}]$:

$$\delta_3(U_i(z,t),U_{i'}(z,t))=\sum_{i=1}^{M_i}\sum_{i'=1}^{M_{i'}}\left[p_{i,k_i}\right]\left[p_{i',k_{i'}}\right]\cdot z^{\min\{[\overline{g}_{i,k_i}],[\overline{g}_{i',k_{i'}}]\}}.$$

4 Analysis of MSS reliability based on interval continuous sequences

According to the analysis of the afore-said algorithm, assuming the gained interval-valued universal generating function of MSS as:

$$U(z,t)=\sum_{k_s=1}^{M_s}\left[p_{k_s}\right]\cdot z^{[g_{k_s}]} \tag{5}$$

Thereinto, $[g_{k_s}]$ is a state performance interval.

The minimum performance requirement interval of MSS is defined as $[w]$ and, then, the interval-valued reliability of MSS is:

$$[R(t)]=P\{[\boldsymbol{G}^{\mathbf{I}}(t)]\geq[w]\}=\sum_{k_s=1}^{M_s}[p_{k_s}]\cdot p([\mathrm{g}_{k_s}(t)]-[w]\geq 0) \tag{6}$$

Thereinto, $P\{[\boldsymbol{G}^{\mathrm{I}}(t)]\geq[w]\}$ represents the probability of $[\boldsymbol{G}^{\mathrm{I}}(t)]-[w]\geq 0$, $[\boldsymbol{G}^{\mathrm{I}}(t)]=\{[\mathrm{g}_1(t)],\cdots,[\mathrm{g}_{M_s}(t)]\}$.

The key to the afore-said problem is the relation between the family of intervals, $[\mathrm{g}_{k_s}(t)]-[w]=[\underline{\mathrm{g}}_{k_s}(t)-\overline{w},\overline{\mathrm{g}}_{k_s}(t)-\underline{w}]$, and 0 if $[\mathrm{g}_{k_s}(t)]$ and $[w]$ are interval variables.

(1) if $\underline{\mathrm{g}}_{k_s}(t)-\overline{w}>0$ and $\underline{\mathrm{g}}_{k_s}(t)-\overline{w}>0$, $p([\mathrm{g}_{k_s}(t)]-[w]\geq 0)=1$;

(2) if $\underline{\mathrm{g}}_{k_s}(t)-\overline{w}<0$ and $\overline{\mathrm{g}}_{k_s}(t)-\underline{w}<0$, $p([\mathrm{g}_{k_s}(t)]-[w]\geq 0)=0$;

(3) if $\underline{\mathrm{g}}_{k_s}(t)-\overline{w}<0$ and $\overline{\mathrm{g}}_{k_s}(t)-\underline{w}>0$, the probability interval of $[\mathrm{g}_{k_s}(t)]-[w]$ can be defined as follows:

$$p([\mathrm{g}_{k_s}(t)]-[w]\geq 0)=\frac{\overline{\mathrm{g}}_{k_s}(t)-\underline{w}}{\overline{\mathrm{g}}_{k_s}(t)-\underline{\mathrm{g}}_{k_s}(t)+\overline{w}-\underline{w}}$$

To sum up, the probability interval of $[\mathrm{g}_{k_s}(t)]-[w]$ can be defined as follows:

$$p([\mathrm{g}_{k_s}(t)]-[w]\geq 0)=\frac{\max\{\overline{\mathrm{g}}_{k_s}(t)-\underline{w},0\}}{\max\{\overline{\mathrm{g}}_{k_s}(t)-\underline{\mathrm{g}}_{k_s}(t)+\overline{w}-\underline{w},\overline{\mathrm{g}}_{k_s}(t)-\underline{w}\}}$$

Therefore, the interval probability of MSS is:

$$[R(\mathrm{t})]=P\{[\boldsymbol{G}(t)]\geq[w]\}=\sum_{k_s=1}^{M_s}[p_{k_s}]\cdot\frac{\max\{\overline{\mathrm{g}}_{k_s}(t)-\underline{w},0\}}{\max\{\overline{\mathrm{g}}_{k_s}(t)-\underline{\mathrm{g}}_{k_s}(t)+\overline{w}-\underline{w},\overline{\mathrm{g}}_{k_s}(t)-\underline{w}\}} \tag{7}$$

5 Analysis of examples

The multi-state serial-parallel system as shown in Diagram 1 is constituted by three subsystems and includes six pieces of performance-degraded component. In the diagram, Component 1 and Component 2 belong to the same type and Component 3, Component 4, and Component 5 belong to the same type. Assuming the component' degraded performances are shown in Table 1, the minimum performance requirement interval of the system is $[w]=[90,110]$, system reliability when t=10000h shall be solved.

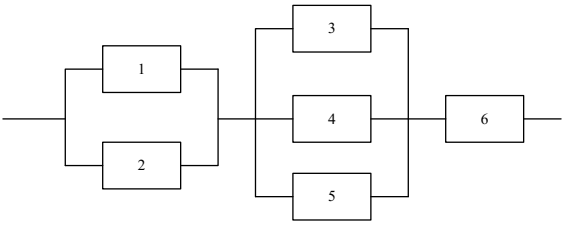


Fig. 1 Multi-State Serial-Parallel System

Table 1 Component' Degraded Performances, Subsystems' Structure Functions, and System's Structure Functions

Component	Component' degraded performance distribution	Subsystems' structure function	System's structure function
1	G_1 obeys normal distribution within [0, [70]], thereinto, $[\mu_1(t)]=[59.75,60.25]-[1.25,1.75]\times10^{-4}\cdot t$ $[\sigma_1(t)]=[1.75,2.25]+[2.75,3.25]\times10^{-5}\cdot t$	Subsystem 1, $X_1(t)= G_1(t)+ G_2(t)$	When $G_6(t) \geqq 70$, $Y(t)=\min\{X_1(t), X_2(t)\}$ $G_6(t)<70$ $Y(t)=0$
3	G_3 obeys normal distribution within [0, [50]], thereinto, $[\mu_3(t)]=[39.75,40.25]-[0.75,1.25]\times10^{-4}\cdot t$ $[\sigma_3(t)]=[1.75,2.25]+[1.75,2.25]\times10^{-5}\cdot t$	Subsystem 2, $X_2(t)= G_3(t)+ G_4(t)+ G_5(t)$	
6	G_6 obeys normal distribution within [0, [100]],, thereinto, $[\mu_6(t)]=[79.75,80.25]-[2.75,3.25]\times10^{-4}\cdot t$ $[\sigma_6(t)]=[3.75,4.25]+[2.75,3.25]\times10^{-5}\cdot t$	Subsystem 3, $X_3(t)= G_6(t)$	

5.1 Interval-valued universal generating functions of component and subsystems

When $t=10000h$, the family of interval continuous sequences of components state performance and probability intervals can be obtained according to the Section 2. Interval-valued universal generating functions of the afore-said six pieces of component can be obtained according to the definition of the family of interval continuous sequences of state performances and the solving method of state probabilities of performance-degraded component, which are shown as follows:

$$u_1(z,t)=[0.0005,0.0011]\cdot z^{[50.75,51.25]}+[0.0518,0.0803]\cdot z^{[54.75,55.25]}+[0.8815,0.9186]\cdot z^{[69.75,70.25]}$$

$$u_2(z,t)=u_1(z,t)$$

$$u_3(z,t)=[0.0289,0.0467]\cdot z^{[34.75,35.25]}+[0.9274,0.9533]\cdot z^{[49.75,50.25]},$$

$$u_3(z,t)=u_4(z,t)=u_5(z,t).$$

It is known from Diagram 1 that Subsystem 1 is constituted by Component 1 and Component 2, and the performance of Subsystem 1 is the sum of the performances of Component 1 and Component 2. According to Operator δ_1 defined in the algorithm of Section 3, the interval-valued universal generating function of Subsystem 1 can be obtained as followings:

$$\begin{aligned}U_{sub1}(z,t)&=\delta_1(u_1(z,t),u_2(z,t))\\&=[0.0001,0.0002]\cdot z^{[105.5,106.5]}+[0.0027,\ 0.0065]\cdot z^{[109.5,110.5]}\\&\quad+[0.0010,\ 0.0020]\cdot z^{[120.5,121.5]}+[0.0913,\ 0.1476]\cdot z^{[124.5,125.5]}\\&\quad+[0.7770,\ 0.8438]\cdot z^{[139.5,140.5]}\end{aligned}$$

Subsystem 2 is constituted by Component 3, Component 4, and Component 5, and its performance is the sum of the performances of the three pieces of component. The interval-valued universal generating function of Subsystem 2 can be obtained in the same manner as follows:

$$\begin{aligned} U_{sub2}(z,t) &= \delta_1(u_3(z,t),u_4(z,t),u_5(z,t)) \\ &= [0,0.0001]\cdot z^{[104.25,105.75]} + [0.0023,\ 0.0062]\cdot z^{[119.25,120.75]} \\ &+ [0.0745,\ 0.1273]\cdot z^{[134.25,135.75]} + [0.7976,\ 0.8664]\cdot z^{[149.25,150.75]} \end{aligned}$$

5.2 Analysis of MSS interval-valued reliability

Operator δ_3 defined in the algorithm of Section 3 is used on Subsystem 1 and Subsystem 2 to solve the interval-valued universal generating function of an integrated subsystem of Subsystem 1 and Subsystem 2, which is shown as follows:

$$\begin{aligned} U_{1-2} &= [0,0.0001]\cdot z^{[104.25,105.75]} + [0,0.0001]\cdot z^{[105.5,106.5]} + [0.0023,0.0063]\cdot z^{[109.5,110.5]} \\ &+ [0.0020,0.0062]\cdot z^{[119.5,120.75]} \quad + [0.0008,0.0020]\cdot z^{[120.5,121.5]} \\ &+ [0.0796,0.1466]\cdot z^{[124.5,125.5]} + [0.0578,0.1074]\cdot z^{[134.25,135.75]} \\ &+ [0.6197,0.7310]\cdot z^{[139.5,140.5]} \end{aligned}$$

Therefore, the interval-valued reliability of Subsystem 1 and Subsystem 2 can be solved according to formula (7):

$$R_{1-2} = [0.7624,\ 0.9998]$$

The reliability of Subsystem 3 is:

$$R_3 = 0.9475$$

Finally, the interval-valued reliability of the system when t=10000h can be solved as follows:

$$R(t) = R_{1-2}(t)\cdot R_3(t) = [0.7224,\ 0.9473]$$

Monte Carlo (MC) simulation method is used for analysis when t=10000h, during which the maximum and minimum simulation results are taken as the maximum and minimum of interval-valued reliability when L=1000 in order to sufficiently state possibilities of calculation results gained from Monte Carlo (MC) in all simulated calculations, then the interval-valued reliability of MC can be solved to be [0.7838, 0.9551].

If a traditional method is used to analyze system reliability, assuming the component has only two states, component' failure threshold is determined according to the total output requirement of the system. Subsystem 1 is normal when the performances of Component 1 and Component 2 are not less than [54.75, 55.25], Subsystem 2 is normal when performances of Component 3, Component 4, and Component 5 are not less than [35.75,36.25], and Subsystem 3 is normal when the performance of Component 6 is not less than 70. Therefore, interval-valued reliability of the system can be obtained to be [0.5516, 0.6570] with a traditional method.

Comparison of interval-valued reliability of systems respectively obtained with the method proposed in this Paper, a traditional reliability method, and Monte Carlo simulation method is shown in Table 2.

Table 2 Contrastive Analysis of System's Interval-Valued reliability Respectively Obtained with the Three Methods

t/*h*	7000	8000	9000	10000	11000
The lower boundary of the family of interval continuous sequences	[0.6046,0.7062]	[0.5800,0.6911]	[0.5543,0.6745]	[0.5278,0.6367]	[0.5006,0.6367]
The mid-value of the family of interval continuous sequences	[0.7816,0.9348]	[0.7546,0.9231]	[0.7264,0.9100]	[0.6973,0.8953]	[0.6674,0.8789]
The upper boundary of the family of interval continuous sequences	[0.7986,0.9691]	[0.7741,0.9628]	[0.7486,0.9555]	[0.7224,0.9473]	[0.6957,0.9381]
MC simulation method	[0.8515,0.9778]	[0.7890,0.9737]	[0.7915,0.9659]	[0.7838,0.9551]	[0.7664,0.9434]
Traditional method	[0.6670,0.7584]	[0.6297,0.7265]	[0.5911,0.6926]	[0.5516,0.6570]	[0.5117,0.6198]

The following conclusions can be drawn from Table 2 after analysis:

(1) When the lower boundary , mid-value, and upper boundary of the family of interval continuous sequences are respectively taken with the method proposed in this Paper, the calculation results generally present an increase trend. Both of the upper boundary and lower boundary of the family of interval continuous sequences do not completely include state information of the family of interval continuous sequences; since the mid-value is the median of the upper boundary and lower boundary , the result is relatively accurate and approaches to the objective value; when the lower boundary of the family of interval continuous sequences is taken, the result has greater deviation from the rest two as time goes by, which will not be analyzed any more in subsequent research.

(2) After the comparison of results respectively obtained with the traditional method, the method of taking the mid-value of the family of interval continuous sequences, and Monte Carlo simulation method, it is known that the traditional method has a greater error than the rest two methods and, when it is used for system reliability estimation, component' failure threshold is usually determined according to its statistical law and cannot be specifically analyzed according to the specific system, which may have some deviations at times.

(3) Although the result obtained with the method of taking the mid-value of the family of interval continuous sequences has some errors compared with the result obtained with Monte Carlo simulation method, this method represents system reliability level to some extent. Therefore, within the range of allowable error, the method of taking the mid-value of the family of interval continuous sequences proposed in this Paper not only resolves the problem that great error occurs in the traditional method, but also overcomes the shortcoming of great scale and long time of Monte Carlo simulation method, which is easily achieved.

6 Conclusion

In consideration of epistemic uncertainty, an analytical method of MSS interval-valued reliability based on epistemic uncertainty is proposed, an analytical flow of multi-state serial-parallel system reliability is provided and verification is conducted with simulated examples. The possible value of system reliability is given in the form of interval value, users can make a reasonable preventive maintenance plan according to the actual demand, to reduce the operation risk of the system.

7 Acknowledgement

This project is supported by the Science Fund of China (No. 61271153). This is partially supported by the Science Fund of China (No. 61372039).

References

1. Lisnianski A. And Levitin G.. Multi-State System Reliability: Assessment, Optimization and Applications. 2003, Singapore: World Scientific.
2. Chun-yang Li. Research on Reliability Analysis and Optimization Based on the Multi-State System Theory, 2010, National University of Defense Technology: Changsha. (in Chinese)
3. Barton, R.M. and Damon W.. Reliability in a multi-state system[C], in Proceedings of the Sixth Annual Southeastern Symposium on Systems Theory. 1974: Louisiana.
4. Barton, R.E. and A.S. Wu . Coherent systems with multi-state components. Mathematical of Operations Research, 1978(3): pp.275-281.
5. Natvig, B., Multi-state Systems Reliability Theory with Applications. 2010, Hoboken, NJ, USA: Wiley.
6. Levitin, G. . The Universal Generating Function in Reliability Analysis and Optimization. 2005, London: Springer.
7. Yu Liu. Multi-State Complex System Reliability Modeling and Maintenance Decision, 2011, UESTC: Chengdu. (in Chinese)
8. Massim, Y., Zeblah A. and Meziane R..Optimal design and reliability evaluation of multi-state series-parallel power systems. Nonlinear Dynamics, 2005(4): pp.309-321.
9. Taboada, H., J. Espiritu and D.W. Coit.Design allocation of multi-state series-parallel systems for power systems planning: a multiple objective evolutionary approach. Journal of Risk and Reliability, 2008. 222(3): pp.381-391.
10. Yeh, W.C.. A simple universal generating function method for estimating the reliability of general multi-state node networks[J]. IIE Transactions, 2009(41): pp. 3-11.
11. Jane, C.C. and Y.W. Laih. A practical algorithm for computing multi-state two-terminal reliability. IEEE Transactions On Reliability, 2008. 57(2): pp. 295-302.
12. Krzysztof, K. and Joanna S.. On multi-state safety analysis in shipping. International Journal of Reliability, Quality and Safety Engineering, 2007. 14(6): pp. 547-567.
13 Kolowrocki, K. and Kwiatuszewska-sarnecka B. . Reliability and risk analysis of large systems with ageing components. Reliability Engineering and System Safety, 2008(93): pp.1821-1829.
14 Ding Yi, M.J.Z.. Fuzzy Multi-State Systems: General Definitions, and Performance Assessment. IEEE Transactions on Reliability, 2008. 57(4):pp. 589-594.

15 Ding, Yi. and Lisnianski A., Fuzzy universal generating functions for multi-state system reliability assessment. Fuzzy Sets and Systems, 2008(159): pp.307-324.
16 Min-qiang Yan, Bo Yang and Zhan Wang. Reliability Assessment for Multi-state System Subject to Imperfect Fault Coverage. Journal of Xi'an Jiao tong University, 2011(10): pp.109-114. (in Chinese)
17 Chun-yang Li, X.C.X.Y., Interval-Valued Reliability Analysis of Multi-State Systems. IEEE Transactions on Reliability, 2011. 60(1): pp.323-330.
18 Destercke S., and Sallak M., An extension of Universal Generating Function in Multi-State Systems Considering Epistemic Uncertainties. Reliability Engineering and System Safety, 2013. 62(2): pp.504-514.
19 Liu, Y. and H. Huang. Reliability assessment for fuzzy multi-state system. International Journal of Systems Science, 2010. 41(4): pp. 365-379.
20 Li, Y., Ding Yi and E. Zio. Random Fuzzy Extension of the Universal Generating Function Approach for the Reliability Assessment of Multi-State Systems Under Aleatory and Epistemic Uncertainties. IEEE Transactions on Reliability, 2014. 63(1): pp.13-23.
21 Dongning Chen and Chengyu Yao. Reliability Analysis of Multi-state System Based on Fuzzy Bayesian Networks and Application in Hydraulic System. Chinese Journal of Mechanical Engineering, 2012(16): pp.175-183. (in Chinese)
22 Dezhong Ma, ZHOU Zhen etc. Reliability Analysis of Multi-state Bayesian Networks Based on Fuzzy Probability, Systems Engineering and Electronics, 2012, 34(12):pp.2607-2611. (in Chinese)
23 Dongning Chen, Chengyu Yao and Zhen Dang Dang. Reliability Analysis of Multi-state Hydraulic System Based on T-S Fuzzy Fault Tree and Bayesian Network. China Mechanical Engineering, 2013(07): pp.899-905. (in Chinese)
24 Zhiping Qiu. Convex Method Based on Non-Probabilistic Set-Theory and Its Application. Beijing: National Defense Industry Press, 2005. (in Chinese)

Toward Construction of Efficient Privacy Preserving Reusable Garbled Circuits

Xu An Wang

Abstract In this paper, we propose an efficient way to construct privacy preserving reusable garbled circuits (RGC) with input privacy (IP) and circuit privacy (CP)(which we denote as RGC − IP − CP) based on two-to-one recoding (TOR) scheme. Currently the only way to achieve reusable garbled circuits (RGC) with input privacy (IP) and circuit privacy (CP) heavily rely on FHE, which is Goldwasser et al.'s work. Compared with GKPVZ13, our work can achieve reusable garbled circuits (RGC) with input privacy (IP) and circuit privacy (CP) with high efficiency.

1 Introduction

In early 1980s, Yao first proposed the technique of garbled circuit [13, 12] to implement the mechanisms of secure two-party computation. This technique runs as the following: First, the sender (or the generator) simultaneously garbled the inputs and the circuit corresponding to the function which they agree to compute, and then send the garbled inputs (keys) and garbled circuits (ciphertext table) to the receiver (or the evaluator); Second, the receiver (or the evaluator) obtains the garbled keys corresponding to the receiver's inputs by running the oblivious transfer protocol with the sender; Third, the receiver (or the evaluator) evaluates the circuit on the received garbled inputs and get the output garbled results; Fourth, the sender recover the final function's output from the garbled results by using the trapdoor used in the beginning garbling process.

The correctness and security requirement for garbled circuits are the following: In this process, the receiver (or the evaluator) know nothing about the sender's inputs, but it can correctly compute the function on both parties' inputs. Although not knowing the receiver's inputs, the sender (or the generator) can ensure the receiver

Xu An Wang
Key Laboratory of Cryptology and Information Security, Engineering University of CAPF, Xi'an 710086, China, wangxazjd@163.com

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_8

has computed the desired function correctly. Furthermore, the receiver (or the evaluator) can not know the concrete function he computed for he just knows the garbled ciphertext table, thus achieve privacy preserving for the sender.

But Yao's garbled circuits can not be reusable, that is, the sender needs to garble again the circuit and the inputs once the inputs have been changed. In 2013, Goldwasser et al. [5] make a first step toward this direction, the first reusable garbled circuits based on functional encryption with fully homomorphic encryption has been constructed, but it is not efficient for the inefficiency of FHE. Recently, a new cryptographic primitive named fully key homomorphic encryption was proposed by Boneh et al.[2] and they constructed such a scheme based on lattice, they also show this primitive can be used to construct reusable garbled circuits with better size, but their result is also not efficient due to the complicated mechanism of fully key homomorphic encryption. Gorbunov et al.[6] constructed reusable garbled circuits based on the elegant primitive named two-to-one recoding (TOR) scheme, however their construction can not achieve input privacy and circuit privacy.

In this paper, we try to give a new way to construct privacy preserving reusable garbled circuits (RGC), we put forward a new method to construct reusable garbled circuits with input privacy (IP) and circuit privacy (CP)(which we denote as RGC − IP − CP) based on two-to-one recoding (TOR) scheme, which is a simple and easily understandable primitive. We organize the papers as the following. In section 2, we review the preliminaries which need to understand our work. In section 3, we review of Gorbunov et al.[6]'s RGC construction based on TOR and show why their scheme can not achieve input privacy and circuit privacy. In section 4, we present our construction of RGC with input privacy and circuit privacy based on their RGC, with some novel techniques we have been developed. We also roughly analysis our construction security. Finally we conclude our work with some interesting open problems.

2 Preliminaries

RGC: Definition and Security Model. We use the term reusable garbled circuit to refer to the most interesting variant of garbled circuits: the ones that can run on an arbitrary number of encoded inputs without compromising the privacy of the circuit or of the input. We recall the definition of a reusable garbled circuit presented in [5].

Definition 1. A reusable garbling scheme for a family of circuits $\mathscr{C} = \{C_n\}$ with $\mathscr{C}_n$ a set of boolean circuits taking as input n bits, is a tuple of p.p.t algorithms RGb=(RGb.Garble, RGb.Enc, RGb.Eval) such that

- RGb.Garble$(1^\lambda, C)$ takes as input the security parameter λ and a circuit $C \in \mathscr{C}_n$ for some n, and outputs the garbled circuit Γ and a secret key sk.
- RGb.Enc(sk, x) takes as input $x \in \{0,1\}^*$ and outputs an encoding c.
- RGb:Eval(Γ, c) takes as input a garbled circuit Γ, an encoding c and outputs a value y which should be $C(x)$.

Correctness. For any polynomial $n(\cdot)$, for all sufficiently large security parameters λ, for $n = n(\lambda)$, for all circuits $C \in C_n$ and all $x \in \{0,1\}^n$,

$$Pr[(\Gamma, sk) \leftarrow RGb.Garble(1^\lambda, C); c \leftarrow RGb.Enc(sk, x); y \leftarrow RGb.Eval(\Gamma, c) : C(x) = y] = 1 - negl(\lambda)$$

Efficiency. There exists a universal polynomial $p = p(\lambda, n)$ (p is the same for all classes of circuits C) such that for all input sizes n, security parameters λ, for all boolean circuits C of with n bits of input, for all $x \in \{0,1\}^n$,

$$Pr[(\Gamma, sk) \leftarrow RGb.Garble(1^\lambda, C); |sk| \leq p(\lambda, n); and \ \ runtime(RGb.Enc(sk, x)) \leq p(\lambda, n)] = 1$$

Security of Reusable Garbled Circuits. Here, we present the security definition of the reusable garbled circuits, as presented in [5]

Definition 2. (Input and circuit privacy with reusability)
Let RGb be a garbling scheme for a family of circuits $C = \{C_n\}_{n \in N}$. For a pair of p.p.t. algorithms $A = (A_1, A_2)$ and a p.p.t. simulator $S = (S_1, S_2)$, consider the following two experiments:

$exp^{\mathsf{real}}_{\mathsf{RGb,A}}(1^\lambda)$:	$exp^{\mathsf{ideal}}_{\mathsf{RGb,A,S}}(1^\lambda)$:
1. $(C, \mathsf{state_A}) \leftarrow A_1(1^\lambda)$	1. $(C, \mathsf{state_A}) \leftarrow A_1(1^\lambda)$
2. $(gsk, \Gamma) \leftarrow \mathsf{RGb.Garble}(1^\lambda, \mathsf{C})$	2. $(\tilde{\Gamma}, \mathsf{state_S}) \leftarrow \mathsf{RGb.Garble}(1^\lambda, \mathsf{C})$
3. $\alpha \leftarrow A_2^{\mathsf{RGb.Enc(gsk,\cdot)}}(C, \Gamma, \mathsf{state_A})$	3. $\alpha \leftarrow A_2^{\mathsf{O}(\cdot,\mathsf{C})[[\mathsf{state_S}]]}(C, \Gamma, \mathsf{state_A})$
4. output α	4. output α

In the above, $\mathsf{O}(\cdot, \mathsf{C})[[\mathsf{state_S}]]$ is an oracle that on input x from A_2 runs S_2 with inputs $C(x)$, $1^{|x|}$, and the latest state of S; it returns the output of S_2 (storing the new simulator state for the next invocation).

We say that the garbling scheme RGb is input- and circuit-private with reusability if there exists a p.p.t. simulator S such that for all pairs of p.p.t. adversaries $A = (A_1, A_2)$, the following two distributions are computationally indistinguishable:

$$\left\{ exp^{\mathsf{real}}_{\mathsf{RGb,A}}(1^\lambda) \right\}_{\lambda \in N} \overset{c}{\approx} \left\{ exp^{\mathsf{ideal}}_{\mathsf{RGb,A,S}}(1^\lambda) \right\}_{\lambda \in N}$$

3 Review of GVW's RGC Construction Based on TOR

TOR.

Definition 3. [6] Formally, a TOR scheme over the input space $S = \{S_\lambda\}$ consists of six polynomial-time algorithms (Params, Keygen, Encode, ReKeyGen, Sim-

ReKeyGen, Recode) and a symmetric-key encryption scheme (E, D) with the following properties:

- Params$(1^\lambda, d_{max})$ is a probabilistic algorithm that takes as input the security parameter λ and an upper bound d_{max} on the number of nested recoding operations (written in binary), outputs "global" public parameters pp.
- Keygen(pp) is a probabilistic algorithm that outputs a public/secret key pair (pk, sk).
- Encode(pk, s) is a probabilistic algorithm that takes pk and an input $s \in S$, and outputs an encoding ψ.

In addition, there is a recoding mechanism together with two ways to generate recoding keys: given one of the two secret keys, or by programming the output public key.

- ReKeyGen$(pk_0, pk_1, sk_0, pk_{tgt})$ is a probabilistic algorithm that takes a key pair (pk_0, sk_0) another public key pk_1, a target public key pk_{tgt}, and outputs a recoding key rk.
- SimReKeyGen(pk_0, pk_1) is a probabilistic algorithm that takes two public keys pk_0, pk_1 and outputs a recoding key rk together with a "target" public key pk_{tgt}.
- Recode(rk, ψ_0, ψ_1) is is a deterministic algorithm that takes the recoding key rk, two encodings ψ_0, ψ_1, and outputs an encoding $\psi_t gt$.

Until now, there are two TOR schemes [6, 11], here we omit the concrete construction.

ABE **for Circuits Based on** TOR**.** Here we review an ABE for circuits (RGC) based on TOR [4].

(Circuit Representation.) Let C_λ be a collection of circuits each having $l = l(\lambda)$ input wires and one output wire. Define a collection $\mathscr{C} = \{\mathscr{C}_\lambda\}_{\lambda \in \mathbb{N}}$. For each $\mathscr{C} \in \mathscr{C}_\lambda$ we index the wires of $\mathscr{C}$ in the following way. The input wires are indexed 1 to l, the internal wires have indices $l+1, l+2, \cdots, |C|-1$ and the output wire has index $|C|$, which also denotes the size of the circuit. We assume that the circuit is composed of arbitrary two-to-one gates. Each gate g is indexed as a tuple (u, v, w) where u and v are the incoming wire indices, and $w > max\{u, v\}$ is the outgoing wire index. The gate computes the function $g_w : \{0,1\} \times \{0,1\} \to \{0,1\}$. The "fan-out wires" in the circuit are given a single number. That is, if the outgoing wire of a gate feeds into the input of multiple gates, then all these wires are indexed the same.

The ABE scheme ABE=(Setup, Enc, KeyGen, Dec) is defined as follows.

Setup$(1^\lambda, 1^l, d_{max})$: For each of the l input wires, generate two public/secret key pairs. Also, generate an additional public/secret key pair:

$$(pk_{i,b}, sk_{i,b}) \leftarrow \mathsf{KeyGen}(\mathsf{pp}) \qquad for\ i \in [l], b \in \{0,1\}$$
$$(pk_{out}, sk_{out}) \leftarrow \mathsf{KeyGen}(\mathsf{pp})$$

output

$$mpk := \begin{pmatrix} pk_{1,0} & pk_{2,0} & \cdots & pk_{l,0} & \\ pk_{1,1} & pk_{2,1} & \cdots & pk_{l,1} & pk_{out} \end{pmatrix} \qquad msk := \begin{pmatrix} sk_{1,0} & sk_{2,0} & \cdots & sk_{l,0} \\ sk_{1,1} & sk_{2,1} & \cdots & sk_{l,1} \end{pmatrix}$$

$\mathsf{Enc}(mpk, ind, m)$: For $\mathsf{ind} \in \{0,1\}^l$, choose a uniformly random $s \leftarrow \$S$ and encode it under the public keys specified by the index bits:

$$\psi_i \leftarrow \mathsf{Encode}(pk_{i,ind_i}, s)\ for\ all\ i \in [l]$$

Encrypt the message m:

$$\tau \leftarrow \mathsf{E}(\mathsf{Encode}(pk_{out}, s), m)$$

output the ciphertext

$$\mathsf{ct}_{\mathsf{ind}} := (\psi_1, \psi_2, \cdots, \psi_l, \tau)$$

$\mathsf{KeyGen}(msk, C)$:

1. For every non-input wire $w = l+1, \cdots, |C|$ of the circuit C, and every $b \in \{0,1\}$, generate public/secret key pairs:

$$(pk_{w,b}, sk_{w,b}) \leftarrow \mathsf{Keygen}(\mathsf{pp})\ if\ w < |C|\ or\ b = 0$$

   and set $pk_{|C|,1)} := pk_{out}$
2. For the gate $g = (u, v, w)$ with outgoing wire w, compute the four recoding keys $rk^w_{b,c}$ (for $b, c \in \{0,1\}$):

$$rk^w_{b,c} \leftarrow \mathsf{ReKeyGen}(pk_{u,b}, pk_{v,c}, sk_{u,b}, pk_{w,g_w(b,c)})$$

   Output the secret key which is a collection of $4(|C| - l)$ recoding keys

$$sk_C := (rk^w_{b,c} : w \in [l+1, |C|], b, c \in \{0,1\})$$

$\mathsf{Dec}(sk_C, ct_{ind})$: We tacitly assume that ct_{ind} contains the index ind, For $w = l+1, \cdots, |C|$, let $g = (u, v, w)$ denote the gate with outgoing wire w, Suppose wires u and v carry the values b^* and c^*, so that wire w carries the value $d^* := g_w(b^*, c^*)$. Compute

$$\psi_{w,d^*} \leftarrow \mathsf{Recode}(rk^w_{b^*,c^*}, \psi_{u,b^*}, \psi_{v,c^*})$$

If $C(ind) = 1$, then we would have computed $\psi_{|C|,1}$. Output the message

$$m \leftarrow \mathsf{D}(\psi_{|C|,1}, \tau)$$

If $C(ind) = 0$, outputs $\perp$.

4 RGC Based on ABE for Circuits

The definition of RGC:

1. RGb.Garble($1^{\lambda}, C$): takes as input the security parameter λ and a circuit $C \in \mathscr{C}_n$ for some n, and outputs the garbled circuit Γ and a secret key sk.
2. RGb.Enc(sk, x) takes as input $x \in \{0,1\}^*$ and outputs an encoding c.
3. RGb:Eval(Γ, c) takes as input a garbled circuit Γ, an encoding c and outputs a value y which should be $C(x)$.

RGC without input privacy and circuit privacy was first constructed by Gorbunov et. al. [6], their construction is actually an attribute based encryption scheme for circuits, Figure 1 expresses this concrete construction process. They show that ABE for circuits schemes can have the authenticity guarantee of GC, and they are reusable garbled circuits without IP and CP properties. Until now there are a few ABE for circuits schemes [2, 3, 4], but all of them can not achieve IP and CP.

5 Our Construction of RGC − IP − CP

Our First Construction of RGC − CP**.** In this subsection, we show how to achieve RGC with CP. The core idea is the following:

1. We first permutate the table of re-encryption keys (**Table**(rk)).
2. We then use SIGNAL to indicate which one row in **Table**(rk) corresponding to the two gate inputs.
3. We add to each wire and **Table**(rk) the indexes to smoothly running this SIGNAL technique.
4. Figure 2 shows such an example.

The idea of using SIGNAL technique has been used in Yao's GC to reduce the computation overhead [7, 8, 9], which is rooted in [10]. We here extended the technique of SIGNAL by just itemizing **Table**(rk)with (1,2,3,4) instead of $((0,0),(0,1),(1,0),(1,1))$. From **Table**($rk$), the adversary can not test whether the gate is a OR, or AND Gate, thus this proposal achieves circuit privacy except leaking the information on the topology and the XOR gates (We will show in the next subsection to show how to hide also the information on the XOR gates), which is acceptable in many interesting application [1].

Our Second Construction of RGC − IP − CP**.** In this subsection we show how to achieve input privacy based on RGC − CP. We need solve the following issues:

1. **Layered Circuits.**

   We need organize the circuits in its layered form, otherwise we can not easily apply our INPUT PRIVACY TECHNIQUE below smoothly. Fortunately every circuits can be organized in its layered form for the following statements in [11, 3]

   > "Without loss of generality, we consider the class of circuits $\mathscr{C} = \{C_\lambda\}$ where each circuit $C \in C_\lambda$ is a layered circuit consisting of input wires, gates, internal wires, and a single output wire. Recall that in a layered circuits gates are arranged in layers where

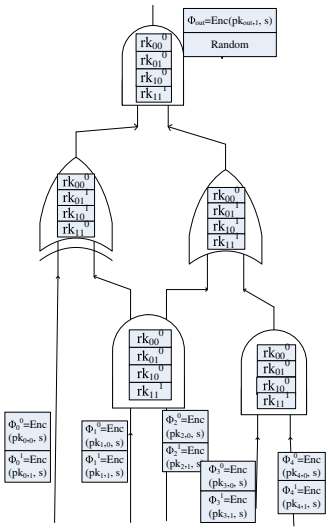


Fig. 1 RGC without IP and CP

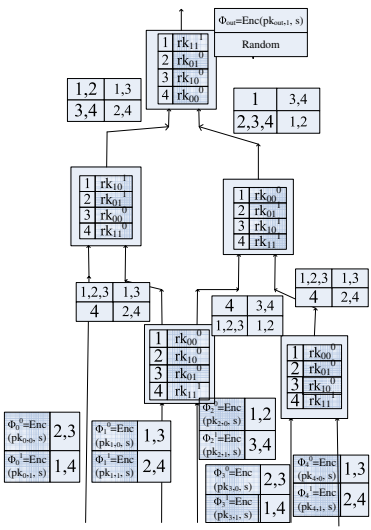


Fig. 2 RGC – CP

every gate at a given layer has a pre-specified depth. The lowest row has depth 1 and depth increases by one as we go up. A gate at depth i receives both of its inputs from wires at depth $i-1$. The circuit has $l = l(\lambda)$ input wires, numbered from 1 to l. The size of the circuit is denoted by $|C|$, and all internal wires are indexed from $l+1, \cdots, |C|-1$; the output wire has index $|C|$. Every gate is a boolean-gate with exactly two input wires and one output wire."

2. How to hide the the information on the XOR gates?

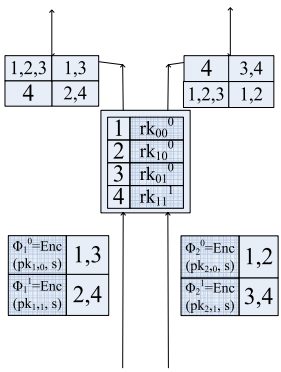


Fig. 3 An input gate

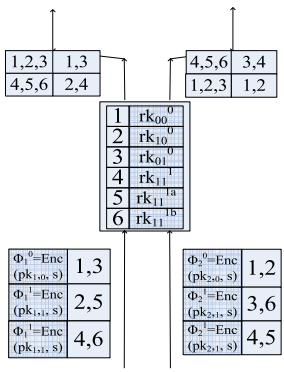


Fig. 4 The transformed input gate

We add one more type of gate to the circuits: the NXOR gate. Why the adversary can easily decide which gate is the XOR gate? For the table associate with the

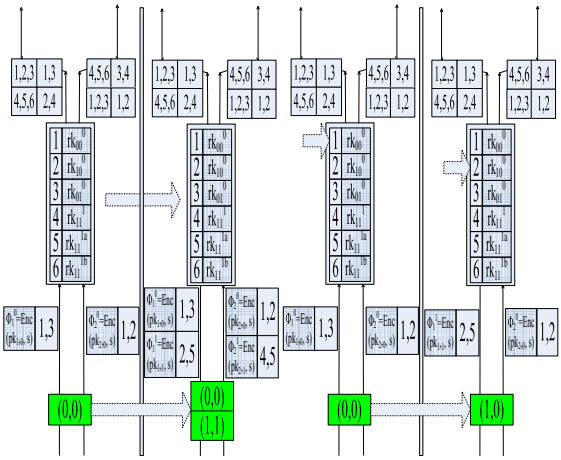


Fig. 5 Twinning the input message

Fig. 6 Substituting the input message

XOR gate has the form of $\begin{array}{|c|c|}\hline 1,2 & 1,3 \\ \hline 3,4 & 2,4 \\ \hline \end{array}$ while the OR and the AND gate has the form of $\begin{array}{|c|c|}\hline 1 & 3,4 \\ \hline 2,3,4 & 1,2 \\ \hline \end{array}$. By adding NXOR gates, there are two types of gate has the form of $\begin{array}{|c|c|}\hline 1 & 3,4 \\ \hline 2,3,4 & 1,2 \\ \hline \end{array}$, just like there are AND gate and OR gate of the form of $\begin{array}{|c|c|}\hline 1,2 & 1,3 \\ \hline 3,4 & 2,4 \\ \hline \end{array}$. In this situation, the adversary can not distinguish XOR gate and NXOR gate, OR gate and AND gate, thus the circuit privacy can be achieved in this way.

3. **How to make the SIGNAL's distribution be uniform?**

   a. We need to on the probability distribution of the SIGNAL in the associate tables with the lowest layer gate. For example, we consider the gate in Figure 3 which is the first input gate in Figure 2. We can observe the tables associated with the output gate has the form of $\begin{array}{|c|c|}\hline 1,2,3 & 1,3 \\ \hline 4 & 2,4 \\ \hline \end{array}$ and $\begin{array}{|c|c|}\hline 4 & 3,4 \\ \hline 1,2,3 & 1,2 \\ \hline \end{array}$, thus the adversary (including the evaluator Bob) can easily derive that the 4th rk in **Table**(rk) should be have the form of rk^b_{00} or rk^b_{11}, for the output of OR gate or AND gate have the 3:1 probability distribution on $\{1,0\}$ or $\{0,1\}$. If the output of the OR gate is 0, then the two input must be 0 and 0; and if the output of the AND gate is 1, then the two input must be 1 and 1. In both cases, the two inputs corresponding to the output fall in the little probability (1/4) space must be the same. Thus if Bob is guided to choose 4th rk in **Table**(rk) as the re-encryption key, then he can derive that the two inputs of this gate are same, which breaking the input privacy of Alice.

   b. We solve this issue by adding **Table**(rk) with two rows. For example, considering the AND gate, we add two rows are of the form (i, rk^{1a}_{11}) and (j, rk^{1b}_{11}) which actually are two fresh rk^1_{11}. We can do this because the re-encryption keys in TOR can be generated probabilistic [6, 11]. i and j are the permutated position in **Table**(rk), which can be arbitrary in $\{1, \cdot, 6\}$. The modification

to the OR gate can be processed similarly except the added two rows are of the form (i, rk_{00}^{0a}) and (j, rk_{00}^{0b}). See Figure 4 for this modification. In Figure 4, this time the tables associated with the output gate will be have the form of $\boxed{1,2,3}\boxed{1,3}$ $\boxed{4,5,6}\boxed{2,4}$ and $\boxed{4,5,6}\boxed{3,4}$ $\boxed{1,2,3}\boxed{1,2}$. And Bob should not know which rk in **Table**(rk) have the form of rk_{00}^{b} or rk_{11}^{b} anymore.

c. When adding two rows in **Table**(rk), we need change the tables with the input wires. We add one row to the tables with the input wire and modify another row, which can be seen in Figure 4. For example, as for the AND gate, We change the two input wires' table of $\boxed{\Phi_1^0 = \mathsf{Enc}(pk_{1,0}, s)}\boxed{1,3}$ $\boxed{\Phi_1^1 = \mathsf{Enc}(pk_{1,1}, s)}\boxed{2,4}$, $\boxed{\Phi_2^0 = \mathsf{Enc}(pk_{2,0}, s)}\boxed{1,2}$ $\boxed{\Phi_2^1 = \mathsf{Enc}(pk_{2,1}, s)}\boxed{3,4}$ to be $\boxed{\Phi_1^0 = \mathsf{Enc}(pk_{1,0}, s)}\boxed{1,3}$ $\boxed{\Phi_1^1 = \mathsf{Enc}(pk_{1,1}, s)}\boxed{2,5}$, $\boxed{\Phi_2^0 = \mathsf{Enc}(pk_{2,0}, s)}\boxed{1,2}$ $\boxed{\Phi_2^1 = \mathsf{Enc}(pk_{2,1}, s)}\boxed{3,6}$ $\boxed{\Phi_1^1 = \mathsf{Enc}(pk_{1,1}, s)}\boxed{4,6}$ $\boxed{\Phi_2^1 = \mathsf{Enc}(pk_{2,1}, s)}\boxed{4,5}$ But this time there are two rows corresponding to input message 1 as the form of $\boxed{\Phi_1^1 = \mathsf{Enc}(pk_{1,1}, s)}\boxed{2,5}$, $\boxed{\Phi_2^1 = \mathsf{Enc}(pk_{2,1}, s)}\boxed{3,6}$ $\boxed{\Phi_1^1 = \mathsf{Enc}(pk_{1,1}, s)}\boxed{4,6}$ $\boxed{\Phi_2^1 = \mathsf{Enc}(pk_{2,1}, s)}\boxed{4,5}$ Alice need to choose the rows which has common SIGNAL for the two inputs. For example, if Alice choose $\boxed{\Phi_1^1 = \mathsf{Enc}(pk_{1,1}, s)}\boxed{2,5}$ for the left input, then she need to choose $\boxed{\Phi_2^1 = \mathsf{Enc}(pk_{2,1}, s)}\boxed{4,5}$ for the right input, for they have the common SIGNAL 5. Otherwise if she choose $\boxed{\Phi_2^1 = \mathsf{Enc}(pk_{2,1}, s)}\boxed{3,6}$, then the evaluator (Bob) who received the garbled circuits and input encodings, he can not do the evaluation for there are no common SIGNAL at all!

d. But be careful! We need consider the following attack scenario: If the garbled gate generator (Alice) choose the input message $x \in \{0,1\}^n$ and its encoding $(\mathsf{Enc}(pk_{i,x_i}, s)(i = 1, 2, \cdots, n))$ uniformly, then for the AND gate, the probability of the inputs are guided to rk_{11}^{1}, rk_{11}^{1a} and rk_{11}^{1b} is 1/4 rather than 1/2, and for the OR gate, the probability of the inputs are guided to rk_{00}^{0}, rk_{00}^{0a} and rk_{00}^{0b} is 1/4 rather than 1/2. Thus the evaluator can from this probability distribution to decide the two inputs of one gate are same or not.

e. So we need some other technique to make this probability distribution uniform. Our technique is to add an additional message which has an opposite output bit for every gate with the input message, we call this technique as "twinning the input message". We can see this technique in Figure 5. For example, if Alice want to let the input to the AND gate be $(0,0)$ with which the output bit should be 0, by using twinning the input message technique, she need to choose another input $(1,1)$ which has the opposite output 1. She then give $\boxed{\Phi_1^0 = \mathsf{Enc}(pk_{1,0}, s)}\boxed{1,\ 3}$ $\boxed{\Phi_1^1 = \mathsf{Enc}(pk_{1,1}, s)}\boxed{2,5}$ as the left input and $\boxed{\Phi_2^0 = \mathsf{Enc}(pk_{2,0}, s)}\boxed{1,\ 2}$ $\boxed{\Phi_2^1 = \mathsf{Enc}(pk_{2,1}, s)}\boxed{4,5}$ as the right input to Bob. Alice apply the "twinning the input message" to every lowest gate, and this will make the SIGNAL's probability distribution be uniform and then he can not no longer derive which inputs of this gate are same by SIGNAL's non-uniform probability distribution.

4. **Same SIGNAL for some gate in two evaluations will imply the gate's inputs in these two evaluations be same**.

   If the two inputs of some input gate are guided to point to 3th row in **Table**(rk) in one evaluation, and in another evaluation the two inputs of the same gate are guided to also the 3th row in **Table**(rk), then the adversary can easily derive the inputs of this gate in the two evaluations are the same, which will break Alice's input privacy. How to solve this problem? Our solution is the technique of "substituting the input message", that is, Alice can choose another input pair instead of the original input pair, as long as this another input pair has the same output bit with the original input pair. And this will make the adversary be difficult to derive the inputs in the two evaluation are same for the inputs can have been substituted! Figure 6 shows a concrete example of "substituting the input message" technique.
   For example, if Alice want to let the input to the AND gate be $(0,0)$ with which the output bit should be 0, by using substituting the input message technique, she need to choose another input $(1,0)$ which has the same output 0 as the input. She then give $\boxed{\Phi_1^1 = \mathsf{Enc}(pk_{1,1},s)}\boxed{2,5}$ as the left input and $\boxed{\Phi_2^0 = \mathsf{Enc}(pk_{2,0},s)}\boxed{1,2}$ as the right input to Bob instead of $\boxed{\Phi_1^0 = \mathsf{Enc}(pk_{1,0},s)}\boxed{1,3}$ as the left input and $\boxed{\Phi_2^0 = \mathsf{Enc}(pk_{2,0},s)}\boxed{1,2}$ as the right input. Alice apply the "twinning the input message" to every lowest gate, and this will make the SIGNAL's probability distribution be uniform and then he can not no longer derive which inputs of this gate are same by SIGNAL's non-uniform probability distribution.

5.1 Security Analysis

We can prove our new proposal can achieve ind-privacy for the circuit and the input message.

Theorem 1. *Our proposal* $\mathsf{RGC}-\mathsf{IP}-\mathsf{CP}$ *can achieve ind-privacy notion for Alice's input and wIND-privacy circuit, if the underlying* TOR *is secure and probabilistic, and Alice honestly follows the protocol described in* $\mathsf{RGC}-\mathsf{IP}-\mathsf{CP}$.

Proof. 1. For the circuit privacy, the intuition is that Bob can only get the SIGNALS and tables, and the distribution of them is uniform, and the structure of AND gate and OR gate is same, the structure of XOR gate and NXOR gate is same. Thus given two circuits of same structure chosen by Bob(but different in some position like one is a AND gate and the other is a OR gate, but can not be one is a AND gate and the other is a XOR gate which is the reason why the proposal can only achieve wIND-privacy instead of IND-privacy), Alice choose one of them and can garble it to be a new circuit C, and this C can not be distinguished by Bob corresponding to which one, for the transformed circuits of these two original circuits are the same.

2. For the input privacy, the intuition is that by using the twinning and substituting the input message techniques, Bob can no longer find the clue to derive any useful information on the input message. Thus given two inputs chosen by Bob, Alice choose one of them and using twinning technique on it by adding a twinning message which has opposite output bit for every lowest layer (input) gate, and then by using substituting technique on these two input messages, and finally output the two encoding sequences and the garbled circuit. From these two encoding sequence and the garbled circuit, Bob can not find it corresponding to which input. Thus it can achieve IND-privacy.

6 Conclusion

In this paper, we propose an efficient way to construct a $\mathsf{RGC}-\mathsf{IP}-\mathsf{CP}$ protocol based on two-to-one recoding (TOR) scheme without using FHE. But we point out that the whole computation workload of the $\mathsf{RGC}-\mathsf{IP}-\mathsf{CP}$ protocol is still huge, this owned to two reasons: One is that we using the "twinning the input message" technique, which will double the computation workload, the other is that TOR primitive now heavily relies on public key cryptography (either lattice cryptography or multilinear map), which is very heavy compared with Yao's original garbled circuits while it only uses block cipher or hash function. But we point out the only $\mathsf{RGC}-\mathsf{IP}-\mathsf{CP}$ protocol proposed in [5] until now also using TOR and FHE which are both heavy primitives, especially the latter. Thus our work can be seen as an intermediate step for practical $\mathsf{RGC}-\mathsf{IP}-\mathsf{CP}$ protocol.

Acknowledgements This work is supported by the Natural Science Foundation of Shaanxi Province (Grant No. 2014JM8300).

References

[1] Bellare, M., Hoang, V.T., Rogaway, P.: Foundations of garbled circuits. In: T. Yu, G. Danezis, V.D. Gligor (eds.) ACM CCS 12, pp. 784–796. ACM Press, Raleigh, NC, USA (2012)

[2] Boneh, D., Gentry, C., Gorbunov, S., Halevi, S., Nikolaenko, V., Segev, G., Vaikuntanathan, V., Vinayagamurthy, D.: Fully key-homomorphic encryption, arithmetic circuit ABE and compact garbled circuits. In: P.Q. Nguyen, E. Oswald (eds.) EUROCRYPT 2014, *LNCS*, vol. 8441, pp. 533–556. Springer, Berlin, Germany, Copenhagen, Denmark (2014). DOI 10.1007/978-3-642-55220-5_30

[3] Garg, S., Gentry, C., Halevi, S., Sahai, A., Waters, B.: Attribute-based encryption for circuits from multilinear maps. In: R. Canetti, J.A. Garay (eds.) CRYP-

TO 2013, Part II, *LNCS*, vol. 8043, pp. 479–499. Springer, Berlin, Germany, Santa Barbara, CA, USA (2013). DOI 10.1007/978-3-642-40084-1_27

[4] Garg, S., Gentry, C., Sahai, A., Waters, B.: Witness encryption and its applications. In: D. Boneh, T. Roughgarden, J. Feigenbaum (eds.) 45th ACM STOC, pp. 467–476. ACM Press, Palo Alto, CA, USA (2013)

[5] Goldwasser, S., Kalai, Y.T., Popa, R.A., Vaikuntanathan, V., Zeldovich, N.: Reusable garbled circuits and succinct functional encryption. In: D. Boneh, T. Roughgarden, J. Feigenbaum (eds.) 45th ACM STOC, pp. 555–564. ACM Press, Palo Alto, CA, USA (2013)

[6] Gorbunov, S., Vaikuntanathan, V., Wee, H.: Attribute-based encryption for circuits. In: D. Boneh, T. Roughgarden, J. Feigenbaum (eds.) 45th ACM STOC, pp. 545–554. ACM Press, Palo Alto, CA, USA (2013)

[7] Hazay, C., Lindell, Y.: Efficient secure two-party protocols: Techniques and constructions. Springer, ISBN 978-3-642-14303-8 (2010). http://u.cs.biu.ac.il/~lindell/efficient-protocols.html

[8] Lindell, Y.: The yao construction and its proof of security. 1st Bar-Ilan Winter School on Cryptography: Secure Computation and Efficiency (2011). http://u.cs.biu.ac.il/~lindell/winterschool2011

[9] Lindell, Y.: Secure two-party computation in practice. 2nd TCE Summer School on Computer Security (2013). http://events-tce.technion.ac.il/files/2013/07/YehudaLindell2.pdf

[10] Naor, M., Pinkas, B., Sumner, R.: Privacy preserving auctions and mechanism design. In: Proceedings of the 1st ACM Conference on Electronic Commerce, EC '99, pp. 129–139. ACM, New York, NY, USA (1999). DOI 10.1145/336992.337028. URL http://doi.acm.org/10.1145/336992.337028

[11] Pandey, O., Ramchen, K., Waters, B.: Relaxed two-to-one recoding schemes. In: SCN 14, LNCS, pp. 57–76. Springer, Berlin, Germany (2014). DOI 10.1007/978-3-319-10879-7_4

[12] Yao, A.C.C.: Theory and applications of trapdoor functions (extended abstract). In: 23rd FOCS, pp. 80–91. IEEE Computer Society Press, Chicago, Illinois (1982)

[13] Yao, A.C.C.: How to generate and exchange secrets (extended abstract). In: 27th FOCS, pp. 162–167. IEEE Computer Society Press, Toronto, Ontario, Canada (1986)

A Heuristically Optimized Partitioning Strategy on Elias-Fano Index

Xingshen Song[1], Kun Jiang[2], and Yuexiang Yang[1]

[1] College of Computer, National University of Defense Technology, Changsha, China
{songxingshen, yyx}@nudt.edu.cn
[2] School of Electronic and Information Engineering, Xian Jiaotong University, Xian, China
jk_365@126.com

Abstract. Inverted index is the preferred data structure for various query processing in large information systems, its compression techniques have long been studied to mitigate the dichotomy between space occupancy and decompression time. During compression, partitioning posting list into blocks aligning to its clustered distribution, can effectively minimize the compressed size while keeping partitions separately accessed. Traditional partitioning strategies using fixed-sized blocks trend to be easy to implement, but their compression effectiveness is vulnerable to outliers. Recently researchers begin to apply dynamic programming to determine optimal partitions with variable-sized blocks. However, these partitioning strategies sacrifice too much compression time.
In this paper, we first compare performances of existing encoders in the space-time trade-off curve, then we present a faster algorithm to heuristically compute optimal partitions for the state-of-the-art Partitioned Elias-Fano index, taking account of compression time while maintaining the same approximation guarantees. Experimental results on TREC GOV2 document collection show that our method makes a significant improvement against its original version.

Keywords: Inverted Index, Index Compression, Partitioning Strategy, Approximation Algorithm

1 Introduction

Due to its simplicity and flexibility, inverted index gains much popularity among modern IR systems since 1950s. Especially in large scale search engines, inverted index is now adopted as their core component to maintain billions of documents and respond to enormous queries. In its most basic and popular form, an inverted index is a collection of sorted sequences of integers[9, 21, 24]. Growing size of data and stringent query processing efficiency requirement have appealed a large amount of research, with the aim to compress the space occupancy of the index and speed up the query processing.

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_9

The compression techniques can be roughly divided into two categories,namely the *integer-oriented encoders* and the *list-oriented encoders*[4, 15]. The integer-oriented encoders assign an unique codeword to each integer of the posting list and are hard to decode because of plenty of bitwise operations. Recent work makes progress by optimizing byte/word-aligned encoders using SIMD instructions[16, 17]. The list-oriented encoders are designed to exploit the cluster of neighboring integers and are much faster to decode, however, inferior in compression ratio.

While state-of-the-art techniques do obtain excellent space-time trade-offs, compression time is always neglected as one evaluating criterion. This can be attributed to the fact that index is always preprocessed offline before deployment, and once being taken into effect, update can be committed in an asynchronous and parallel manner. However, timely update for unexpected queries is becoming more and more stringent in search engine, traditional methods do not fit this scenario any more. Until recently researchers begin to use breakthroughs in compact data structure to bear on problems of optimizing inverted index[10, 12, 13].

Our contribution. We show that many encoders are far from optimal in terms of compression time while partitioned Elias-Fano index (PEF) yields a better trade-off than others, and we provide a heuristic partitioning strategy on PEF which computes partitions in a single pass with only one sliding window. We perform an experiment that shows index partitioned by our method is comparable to the original one.

2 Partitioning Strategies

One thing to be noted is that list-oriented encoders may cost equivalent time to compress the posting list as the integer-oriented encoders even they are designed to compress a list of integers at the same time. Before compressing, a partitioning strategy needs to traverse the whole posting list to decide an optimal partition for compression ratio and decompression time, even after that, an uniform bit width is also needed to encode every element inside each block. Compression techniques like the Simple Family[2, 3], enumerates all the possible partitioning cases to choose the suitable decision; OptPFOR[22] needs an additional computation to choose the optimal proportion of exceptions in each block in order to achieve a better space efficiency. In the last few years there is a surge of partitioning schemes to accelerate the compression procedure[8, 12].

Directed Acyclic Graph. Blocks with fixed length are likely to be suboptimal as integers in posting list will not be evenly distributed.Works from literature[1, 5, 15]give another perspective on partitioning. The posting list $\mathcal{S}[0, n-1]$ is considered a particular directed acyclic graph $\mathcal{G}$, each integer is represented by a vertex v, edge denoted as (v_i, v_j) is an exact correspondence of a partition in the posting list $\mathcal{S}[i, j]$, edge has also associated its cost $c(v_i, v_j) = |\mathcal{E}(\mathcal{S}[i, j-1])|$ that corresponds to the size in bits of the partition compressed by encoder $\mathcal{E}$.

Thus, the problem of optimally partitioning a list is reduced to the problem of Single-Source Shortest Path(SSSP) Problem.

However, a trivial traversal may not suffice to obtain an efficient solution for this problem, as the graph $\mathcal{G}$ is complete with $\Theta\left(n^2\right)$ edges, even partitioning posting lists with thousands of integers wi'll be intolerable. Anh et al.[1] and Delbru et al.[5] adopt a greedy mechanism to partition the lists, the difference between them is that the former partitions the lists under a static table driven approach while the latter uses a set of fixed-sized sliding windows to determine. Fabrizio[15] finds the optimal partition under a *dynamic programming* approach but with limited options for partition lengths (say h), reducing its time complexity from $O(n^2)$ to $O(nh)$, but still barely satisfying in practice.

Since dynamic programming is inefficient and greedy mechanism is too coarse, a more feasible way will be using an elaborate approximation algorithm to find slightly suboptimal solutions, which reduce the time and space complexities to linear.

3 Method Evaluation Criteria

Compression techniques are optimized under two criteria, namely *decompression time* and *space occupancy*, different encoders yield different space-time trade-offs, generally, integer-oriented encoders tend to be more space-efficient while list-oriented encoders focus on time efficiency. *Compression time* stays a low profile in the literature, as it does not make a difference for practical use, many encoders pursue a better performance at the expense of prolonging compression time, recently it begins to gain attention as one criterion to evaluate a compression technique[8, 11].

Pareto-Optimal Compression. Taking the compression time into account, we get an extended tri-criteria to evaluate the performance of compression techniques. In this respect, we recall the concept of *Pareto-optimal* compression to define the notion of "best" compression in a principled way, i.e., encoders achieve different points in the space-time trade-off curve, optimal ones are those extreme points where performances are not worse than others in one dimension, and if to be optimized in any dimension, performances in the other two will get impaired accordingly. Absolutely there exists a set of Pareto-Optimal Compressions with different considerations. Figure 1 shows performances of different encoders[3].

From the fitting curve and the marginal rugs, we can observe that points, which approach to limit of one dimension, begins to cluster and performances of the other two dimension drops intensely. Points beyond the curve are either superior or inferior to the average (i.e., VB and AFOR).

Partitioned Elias-Fano Index. We also emphasize PEF by two straight lines in Figure 1, as it reaches a space occupancy competitive with integer-oriented

[3] The installation of these implementations is same with the Experiments Section, even quite different from results in other work, but they are directly comparable with each other.

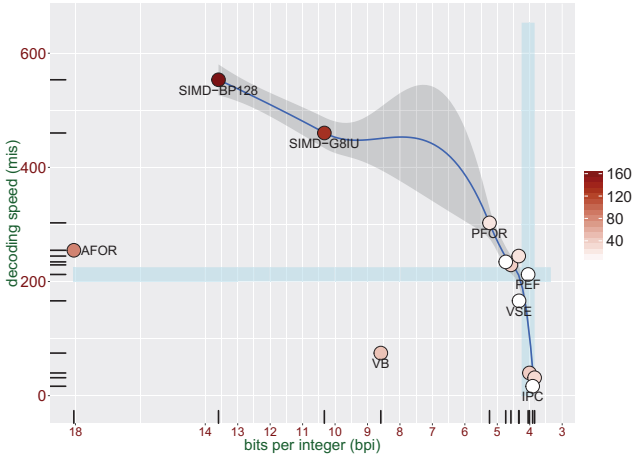


Fig. 1. Performance of different encoders under tri-criteria. The x-axis (bpi) is arranged in a reverse order and color of each point indicates the encoding speed, the deeper, the faster.

encoders while keeping a reasonable decoding speed. It is a two-level data structure which partitions the posting list into Elias-Fano (EF) represented *chunks* to adapt to its local statistics and stores pointers to these chunks in the upper EF sequence. Study shows that it is efficient in random access and search operations but slow in sequential decoding[11, 19].

PEF describes a suboptimal partitioning strategy which finds the solution by generating a pruned DAG on the posting list in $O\left(n \log_{1+\varepsilon} \frac{1}{\varepsilon}\right)$ time and $O\left(\log_{1+\varepsilon} \frac{1}{\varepsilon}\right)$ space, and it is an $(1+\varepsilon)$-approximation algorithm to the optimal one, for any given $\varepsilon \in (0, 1)$. However, it is still prohibitive in compression time, next section we describe a faster algorithm preserving the same approximation guarantee.

4 Linear Partitioning Strategy

4.1 Partitioned Elias-Fano Index

Given a monotonically increasing sequence $\mathcal{S}[0, n-1]$ of n integers drawn from the range $[0, u]$, and u is called an *universe* of $\mathcal{S}$. The Elias-Fano representation is a efficient quasi-succinct representation for monotone sequences used in [6, 19]. $\mathcal{S}$ is represented using two bit arrays, namely *the higher bits* and *the lower bits*. For a given integer ℓ, each element of $\mathcal{S}$ is split into the lower ℓ bits and the higher $\lceil \log u \rceil - \ell$ bits. The lower bits are stored explicitly and contiguously in the lower bits array L, the higher bits are stored in the higher bits array H as a sequence of unary-coded gaps ([12] uses unary-coded buckets instead, but they both reach the same space complexity). Thus, representing L needs exactly $n\ell$ bits, H needs $n + \frac{\lceil u \rceil}{2^\ell}$ bits. It has been shown that setting $\ell = \lfloor \log \frac{u}{n} \rfloor$ minimizes

the overall space of $\mathcal{S}$, namely, at most $n\lceil \log \frac{u}{n} \rceil + 2n$ bits. Elements in Elias-Fano represented sequence are directly accessible by performing unary read on H and direct read on L, then merging them together.

However, Elias-Fano fails to exploit the distribution of clusters in the sequence as it treats $\mathcal{S}$ as a whole. We can actually expect for a better space occupancy when the sequence is formed by clusters of integers which are very close to each other. This observation motivates the introduction of a two-level Elias-Fano, namely PEF, in which $\mathcal{S}$ is first partitioned into chunks with variable lengths, then pointers to the head of each chunk are grouped together using Elias-Fano representation. Partitioning the sequence aligning to its clusters will definitely narrow down the average distance of consecutive elements inside, thus a smaller compressed size is possible. An optimal partition must satisfy the following two requirements: on one hand, the chunks should be as large as possible to minimize the number of pointers in upper level; on the other hand, the chunks should be as small as possible so as to narrow down the average distance. However, traditional dynamic programming methods like [15] turn out to be very costly when finding the optimal solution. A more feasible way is adopting an approximation algorithm to prune the complete graph $\mathcal{G}$ under some criteria, retaining its shortest path while cutting out edges which are impossible to be space-efficient.

Recall the approximation algorithm adopted in PEF, whose core idea is to generate a pruned subgraph $\mathcal{G}_\varepsilon(\mathcal{S})$ of the original $\mathcal{G}(\mathcal{S})$, the pruning strategy works by sparsifying edges in $\mathcal{G}(\mathcal{S})$ into a geometric sequence with common ratio equals $(1+\varepsilon_2)$, where $\varepsilon_2 \in (0,1)$, that is, for each vertex v_i keeping the edge that approximates at the best the value $F(1+\varepsilon_2)^k$ from below for each integer k (F is the fixed cost of one partition). By setting an upper bound U (say, $U = \frac{F}{\varepsilon_1}$ by a predefined $\varepsilon_1 \in (0,1)$), the total edges, addressed as $\varepsilon - maximal$ edges, in $\mathcal{G}_\varepsilon(\mathcal{S})$ are $O\left(n \log_{1+\varepsilon} \frac{1}{\varepsilon}\right)$.

4.2 Heuristically Finding Optimal Partition

These $\varepsilon - maximal$ edges can be found by keeping $k = O\left(\log_{1+\varepsilon} \frac{1}{\varepsilon}\right)$ sliding windows over $\mathcal{S}$. During the algorithm visits posting list, the sliding windows are actually potential partitions, which start at the same position v_i but have different lengths. At first, all these windows are docked at vertex 0 with 0 length. Each time as the algorithm visits a subsequent vertex, these windows expand their sizes by appending this vertex to the end, once cost of the vertexes within current window, say ω_i, exceeds its threshold $F(1+\varepsilon)^i$, ω_i stops expanding its size and we get one $\varepsilon - maximal$ edge of class i. Windows $\omega_{i+1}, \ldots, \omega_k$ will keep repeating this procedure until all the $\varepsilon - maximal$ edges out going from the starting vertex are figured out. Then algorithm repeats the above procedure to find $\varepsilon - maximal$ edges for next vertex until reaches the end. Given the fact that cost calculation can be done in constant time, it is easy to prove that returning an optimal partition only needs almost $2n \log_{1+\varepsilon} \frac{1}{\varepsilon}$ calls to cost calculation, namely in $O\left(n \log_{1+\varepsilon} \frac{1}{\varepsilon}\right)$ time.

However, it is still inefficient in practice since there can be tens of them, actually we need only one window to accommodate as many as possible integers and stop expanding as soon as an *exception* is encountered. The following lemma states a crucial property of the path over $\mathcal{G}(\mathcal{S})$ to base our partitioning strategy on: given any triple of indexes i, j and k in $\mathcal{G}$ with $0 \leqslant i < j < k \leqslant n$, we have $c(v_i, v_j) \leqslant c(v_i, v_k)$ and $c(v_j, v_k) \leqslant c(v_i, v_k)$.

We denote edges that start from v_i over $\mathcal{G}_\varepsilon(\mathcal{S})$ as (v_i, v_{j_0}), (v_i, v_{j_1}), ..., (v_i, v_{j_k}), and for any adjacent two edges the ratio between them is $(1 + \varepsilon_2)$. We speculate the lengths between them should also follow some regulations, an exception can enlarge the average distance inside an edge in a sudden, to keep its weight proportional to its former one, current edge's incremental length has to be comparatively small. However, edges without exceptions should have their edge lengths growing proportionally. A chunk that contains exception can be found by sequentially traversing these edges and comparing the incremental lengths. Intuitively, an exception always leads to a short interval, and once we find this kind of interval a chunk can be built inside it, thus omitting exhaustive traversing all edges from one vertex to figure it out. We can further shift the head of sliding window to the end of current chunk to skip more calculations. As shown in Figure 2.

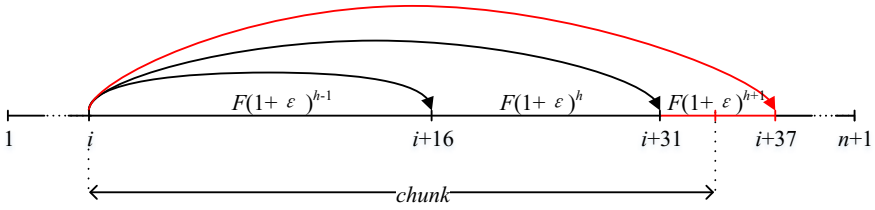


Fig. 2. Finding $\varepsilon - maximal$ edges for position i using sliding windows. Here we show three edges with cost upper bound as $F(1+\varepsilon)^{h-1}$, $F(1+\varepsilon)^h$, $F(1+\varepsilon)^{h+1}$, and their edge lengths are 16, 15, 6, respectively. Then we certain that an exception is encountered inside the third edge, and we can build a chunk right before the exception to save more space and reduce calculation.

It remains to describe how to locate the chunk which contains exceptions, given a chunk (v_i, v_{j_h}), we only consider comparing with its former one, namely $(v_i, v_{j_{h-1}})$, inspired by Markov process. Thus we have their partition lengths and partition universes as follow:

$$n_h = j_h - i, \qquad u_h = v_{j_h} - v_i \tag{1}$$

$$n_{h-1} = j_{h-1} - i, \qquad u_{h-1} = v_{j_{h-1}} - v_i \tag{2}$$

and their bit costs

$$c(v_i, v_{j_h}) / c(v_i, v_{j_{h-1}}) = 1 + \varepsilon_2 \tag{3}$$

However, different encodings are used in PEF to overcome the space inefficiencies of Elias-Fano in representing dense chunks. A chunk $\mathcal{P}_h$ is called a dense chunk if it covers a large fraction of the elements in the universe, that is n_h is close to u_h. As stated before, it will cost Elias-Fano $n_h\lceil\log\frac{u_h}{n_h}\rceil + 2n_h$ bits to represent $\mathcal{P}_h$, nearly $2u_h$ for a dense chunk. In this case, a bitvector with a length u_h bits is a better choice, $\mathcal{P}_h$ is represented by setting positions where integers happen to one. And traditional *rank/select* are also easy to implemented. another extreme case is that chunk covers all the elements of universe u_h, then we just need to store the partition length and partition universe while leaving the chunk non-encoded. In spite of it is a fairly rare occurrence, once occurs it does sharply reduce the space occupancy.

To sum up, there are three representations $\mathcal{E}$ for different chunks. $\mathcal{E} = 0$: *non-encoding* with $c(v_i, v_{j_h}) = 0$, if partition length n_h equals to the universe u_h; $\mathcal{E} = 1$: *bitvector* with $c(v_i, v_{j_h}) = u_h$, if $n_h \geqslant \frac{u_h}{4}$ and $\mathcal{E} = 2$: EF with $c(v_i, v_{j_h}) = n_h\lceil\log\frac{u_h}{n_h}\rceil + 2n_h$, if $n_h < \frac{u_h}{4}$. When we are about to identify chunks contain exceptions, we have to enumerate all the possible combinations of two consecutive chunks to find out corresponding reasonable incremental lengths. That is, given Eq. (1) (2) (3), we need to calculate $\frac{n_h}{n_{h-1}}$ for $\mathcal{E}[h-1]$ and $\mathcal{E}[h]$, $\mathcal{E} \in (0, 1, 2)$. Actually there are only 4 combinations to consider since *non-encoding* is unlikely to be leaded by the other two and more sensitive to exceptions.

Take $\mathcal{E}[h-1] = 1$ and $\mathcal{E}[h] = 2$ for example, then

$$1 + \varepsilon_2 = \frac{n_h\lceil\log\frac{u_h}{n_h}\rceil + 2n_h}{u_{h-1}} \geqslant \frac{n_h}{n_{h-1}} \cdot \frac{2 + \log\frac{u_h}{n_h}}{4} > \frac{n_h}{n_{h-1}}.$$

Others get the same conclusion by deduction , so we set

$$\frac{u_h}{n_h} = 1 + \varepsilon_2 \tag{4}$$

as the same approximation factor with that of $\varepsilon - maximal$ edges. Thus the whole procedure, which can be done within a single pass, is recast into using one window to expand its size from a starting vertex, building a chunk when current $\varepsilon - maximal$ edge is $(1 + \varepsilon_2)$ times larger than the previous one and moving the starting vertex to $v_{h-1} + 1$. A more space-efficient way is to traverse the interval (v_h, v_{h-1}) to determine a better cut-off vertex, which increases time complexity by no more than $(1 + \varepsilon_2)$, and a minimum partition length is set to 8 to avoid slow-start and fragmentation of chunks.

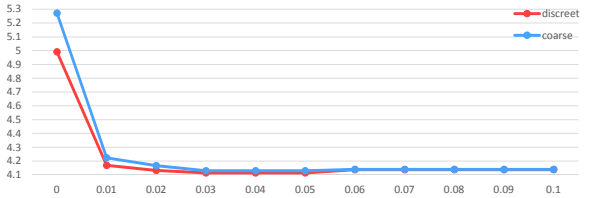
5 Experiments

We use the posting lists extracted from the TREC GOV2 collection, which consists of 25.2 million documents and 32.8 million terms. All the terms have the Porter stemmer applied, stopwords are removed, and docids are reordered by the lexicographic order of URLs. All the implementations are carried out on an 8 core Intel(r) Xeon(r) E5620 processor running at 2.40 GHz with 128GB of

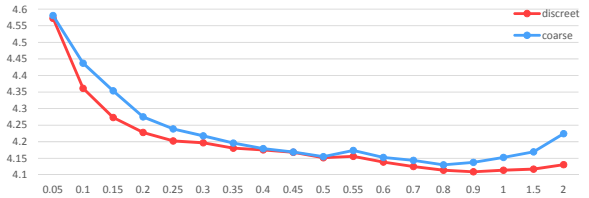
RAM and 12,288KB of cache. Our algorithms are implemented using C++ and compiled with GCC 4.8.1 with -O3 optimizations. In all our runs, the whole inverted index is completely loaded into main memory, executions are reported as the mean of 4 consecutive replications.

We test performances of two strategies on EF index, *coarse* which finishes partitioning in a single pass, and *discreet* which works in a more space-efficient way, against the original methods, namely uniform-partitioned PEF (uniform) and ε-partitioned PEF (optimal).

First of all, We experiment differences caused by predefined parameters, namely the approximation ε_2 and the upper bound parameter ε_1. As coarse and discreet partition a posting list in a single pass, relying only on the total length. Construction time has little relevance to these parameters, and in practice it stays stable to them unsurprisingly. Figure 3 shows the influences of them on index size, from which we can gain an insight into these two parameters: ε_2 determines the sensitivity to partitions which contains exceptions, and ε_1 determines the largest partition length which contains no exceptions, we can see index size is insensitive to ε_1 when it is nonzero, demonstrating the fact that most partitions cannot reach the largest length when encounter an exception. ε_2 used to be approximation bound, however index size is independent on it, as a small value will increase the number of partitions and a large one will yield too many long partitions. We set $\varepsilon_2 = 0.9$ and $\varepsilon_1 = 0.03$ in terms of performance.



(a) $\varepsilon_2 = 0$, varying ε_1 from 0 to 0.1



(b) $\varepsilon_1 = 0.03$, varying ε_2 from 0.05 to 2

Fig. 3. Influences of the parameters ε_2 and ε_1 on two partitioning strategies.

Compression. Table 1 shows the performances of different partitioning strategies, adding OptPFOR-compressed index (OptPFOR) as a baseline. It is clear that both proposed strategies outperform the original ones by a large margin over

the compression time, making construction of PEF competitive with OptPFOR, at the cost of a slight larger space occupancy. Discreet has obvious superiority compared with coarse, sacrificing 4.7% construction time in exchange for 6% smaller index size, however, it still gets a better compression ratio than the last two uniformly partitioned indexes. The last column denotes time used for a sequential decompression of the index, discreet and coarse tends to use 10% longer time to traverse the index than optimal and uniform, delayed by suboptimal partitions. Although all EF-based methods decompress much slower than OptPFOR, their advantages are embodied in random access when applied to query processing.

Table 2 shows comparisons of average partition lengths for different indexes, optimal has larger partition lengths but smaller index size than coarse and discreet, proving the fact it can better exploit the local clustering of the posting list. Discreet and coarse do not differ much from each other, as they can only approximately locate an exception and build a chunk in front of it. Next we show these lengths also influence query processing speed.

Table 1. Comparison of construction time, and average bits per element of each component

methods	compress **sec**	docid **bpi**	freq **bpi**	decompress **sec**
discreet	870.70	4.40	2.32	99.46
coarse	785.94	4.66	2.41	99.09
optimal	2638.91	4.05	2.19	86.11
uniform	1615.94	4.63	2.40	88.61
OptPFOR	688.90	4.55	2.34	49.43

Table 2. Average partition lengths of different indexes for each component

	optimal	coarse	discreet	uniform	OptPFOR
docid	208	194	187	128	128
freq	496	331	319	128	128

Query efficiency. In order to explore the differences caused by partitioning strategies, the experiment adopts four widely-used methods of ranked query to find the top-20 results, DAAT_AND, DAAT_OR, WAND and Maxscore. We randomly select 4000 terms from the lexicon, any of which has a document frequency within the range $(10^5, 10^6)$, and we regroup them into 1000 queries

every 4 terms, avoiding the bias brought by short posting lists while highlighting advantage of random access.

As shown in Figure 4, query time has been enlarged by involvements of multiple long posting lists, outliers can be as large as 10^5 msec, however, most queries can be processed within hundreds of msec. Performances vary a lot among different processing methods,DAAT traverses the index in an exhaustive way, relying only on the throughput of postings. Needless to say, OptPFOR performs the best in DAAT and other indexes rank in accord with average partition lengths, the larger the length, the faster the speed. In detail, DAAT_AND is the most efficient one as list intersection can cut off large of invalid postings, DAAT_OR comes at a much higher cost as it cut nothing off, its rank also coincides with decompression speed shown in Table 1. When it comes to dynamic pruning, PEF begins to gain an advantage over OptPFOR, also performance gaps among indexes again become unapparent, discreet and coarse even outperform optimal in WAND comparing both outliers and median. In general, results of different indexes do not differ much except those using DAAT_OR, performances of discreet and coarse rank between optimal and uniform.

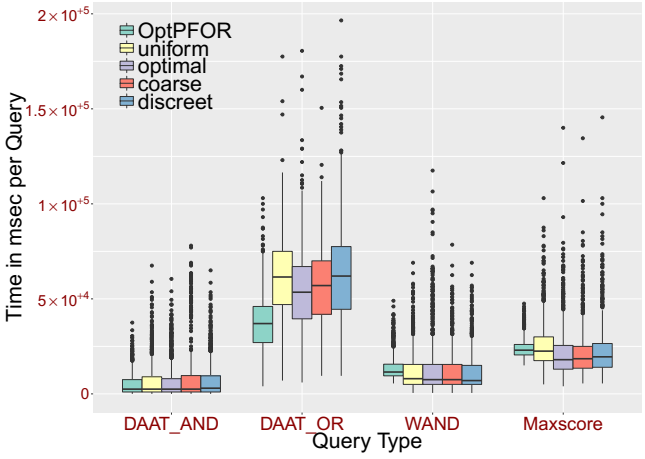


Fig. 4. Query time distribution using different ranked query processing methods on candidate indexes.

6 Conclusion and Future Work

We have introduced the notion of Pareto-Optimal compression techniques taking compression time as one criterion and compared performances of different compressions. Our heuristic partitioning strategy over PEF gains lower time complexity while preserving same approximation guarantees, despite its simplicity, it does work well in practice when applied on GOV2. Inspired by research

from literature [20, 23, 14, 7, 18], future work will focus on formulating the notion of Pareto-Optimal Compression and exploring linear-time algorithms with different trade-offs under its criteria.

References

1. Anh, V.N., Moffat, A.: Index compression using fixed binary codewords. In: Proc. ADC. pp. 61–67 (2004)
2. Anh, V.N., Moffat, A.: Inverted index compression using word-aligned binary codes. Inform. Retrieval 8(1), 151–166 (2005)
3. Anh, V.N., Moffat, A.: Index compression using 64-bit words. Soft. Prac. & Exp. 40(2), 131–147 (2010)
4. Catena, M., Macdonald, C., Ounis, I.: On inverted index compression for search engine efficiency. In: Proc. ECIR. pp. 359–371 (2014)
5. Delbru, R., Campinas, S., Tummarello, G.: Searching web data: An entity retrieval and high-performance indexing model. J. Web Sem 10, 33–58 (2012)
6. Elias, P.: Efficient storage and retrieval by content and address of static files. Journal of the ACM (JACM) 21(2), 246–260 (1974)
7. Lai, C., Moulin, C.: Semantic indexing modelling of resources within a distributed system. International Journal of Grid and Utility Computing 4(1), 21–39 (2013)
8. Lemire, D., Boytsov, L.: Decoding billions of integers per second through vectorization. Soft. Prac. & Exp 45(1), 1–29 (2015)
9. Manning, C.D., Raghavan, P., Schütze, H., et al.: Introduction to information retrieval. Cambridge university press (2008)
10. Navarro, G., Puglisi, S.J.: Dual-sorted inverted lists. In: Proc. SPIRE. pp. 309–321 (2010)
11. Ottaviano, G., Tonellotto, N., Venturini, R.: Optimal space-time tradeoffs for inverted indexes. In: Proc. WSDM. pp. 47–56 (2015)
12. Ottaviano, G., Venturini, R.: Partitioned elias-fano indexes. In: Proc. SIGIR. pp. 273–282 (2014)
13. Petri, M., Moffat, A., Culpepper, J.S.: Score-safe term-dependency processing with hybrid indexes. In: Proc. SIGIR. pp. 899–902 (2014)
14. Shorfuzzaman, M., Graham, P., Eskicioglu, R.: Allocating replicas in large-scale data grids using a qos-aware distributed technique with workload constraints. International Journal of Grid and Utility Computing 3(2-3), 157–174 (2012)
15. Silvestri, F., Venturini, R.: Vsencoding: efficient coding and fast decoding of integer lists via dynamic programming. In: Proc. CIKM. pp. 1219–1228 (2010)
16. Stepanov, A.A., Gangolli, A.R., Rose, D.E., Ernst, R.J., Oberoi, P.S.: Simd-based decoding of posting lists. In: Proc. CIKM. pp. 317–326 (2011)
17. Trotman, A.: Compression, simd, and postings lists. In: Proc. ADCS. p. 50 (2014)
18. Tudor, D., Macariu, G., Schreiner, W., Cretu, V.I.: Experiences on grid shared data programming. International Journal of Grid and Utility Computing 1(4), 296–307 (2009)
19. Vigna, S.: Quasi-succinct indices. In: Proc. WSDM. pp. 83–92 (2013)
20. Wang, Y., Ma, J., Lu, X., Lu, D., Zhang, L.: Efficiency optimisation signature scheme for time-critical multicast data origin authentication. International Journal of Grid and Utility Computing 7(1), 1–11 (2016)
21. Witten, I.H., Moffat, A., Bell, T.C.: Managing gigabytes: compressing and indexing documents and images. Morgan Kaufmann (1999)

22. Yan, H., Ding, S., Suel, T.: Inverted index compression and query processing with optimized document ordering. In: Proc. WWW. pp. 401–410 (2009)
23. Zhang, T., Cui, L., Xu, M.: A lns-based data placement strategy for data-intensive e-science applications. International Journal of Grid and Utility Computing 5(4), 249–262 (2014)
24. Zobel, J., Moffat, A.: Inverted files for text search engines. ACM Comp. Surv. 38(2), 6 (2006)

Smart Underground: Enhancing Cultural Heritage Information Access and Management through Proximity-Based Interaction

Giuseppe Caggianese and Luigi Gallo

Abstract This paper describes the Smart Underground system, the main aim of which is to enhance the access to cultural heritage information for the visitors. The system provides a more interactive visiting experience based on a proximity interaction with the artefacts in an exhibition, which allows an easy access to a new level of cultural information proposed by the exhibition curators. For this reason, the system also offers a set of tools for the curators with the aim of simplifying the organization and updating of the cultural content information. Finally, the system, by integrating modern technologies with the real works of art in the exhibition, proposes a possible solution to the emerging problem of the management and dissemination of cultural heritage digitalized content, leading to an improvement in the experiences of both visitors and curators.

1 Introduction

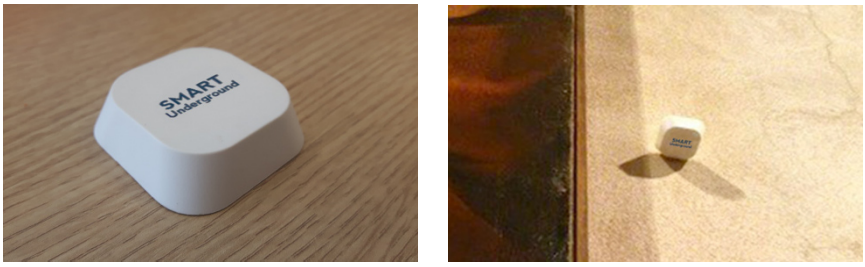
In the last few years, many research activities have been undertaken in order to improve the access to cultural heritage information. The main focus of all these works has been related to an improvement in the visitor experience, achieved by exploiting the technologies available at that time. Most of the initial works were related to the use of web systems to provide a general description of the museum or historical site, together with information about visiting times and special events. Soon afterwards, web systems were also used to reproduce virtually a visit to a museum, giving to the

Giuseppe Caggianese
Institute for High Performance Computing and Networking
National Research Council of Italy (ICAR-CNR), Naples, Italy, e-mail: giuseppe.caggianese@icar.cnr.it

Luigi Gallo
Institute for High Performance Computing and Networking
National Research Council of Italy (ICAR-CNR), Naples, Italy, e-mail: luigi.gallo@cnr.it

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_10



(a) (b)

Fig. 1 A Bluetooth beacon device and its possible installation close to an artefact.

visitor the opportunity to navigate a computer generated exhibition and enjoy the cultural artefact in the same way that she/he was used to doing during a real visit. In line with this trend, cultural heritage institutions that have recently made their own version of a virtual exhibition include prestigious museums such as the Smithsonian National Museum of Natural History, the Louvre, the Metropolitan Museum, and the Archaeological Museum in Naples. However, in all these proposed virtual tours, the visitor's enjoyment is still characterized by a passive visualization that does not allow any active interaction with the artefacts and, moreover, only a limited use of the virtual information proposed is made during a real visit to the site.

Nowadays, the rise of new mobile technologies has provided the means to enhance this experience, offering the possibility of supporting the visitor, principally during a real visit. In fact, since the early stages of their development, mobile technologies have been employed to realize electronic guides with the aim of facilitating the access to cultural artefact information. However, common digital guides present many drawbacks such as: the requirement for the visitor to follow a predefined path; the need for structural interventions in the exhibition in order to clearly indicate to the visitor her/his position during the visit; the lack of any flexibility for the curator in updating the cultural information and finally, and not least importantly, the time wasted by the visitor waiting in line to obtain such a digital guide.

In this paper, we present a cultural system which tries to overcome the aforementioned problems of the classic digital guides. The proposed system exploits the visitor's smartphone and a proximity-based interaction in order to offer a more attractive way of visiting a cultural site through a fully-featured, interactive exploration of the cultural artefacts in an exhibition. At the same time, the system is extremely flexible, enabling the exhibition curators to more easily organize and manage any additional cultural content information proposed to the visitor.

Thanks to the proposed platform, visitors are not required to wait in a line for their digital guide, but can immediately start their visit using their smartphone. They can freely move in the environment without concentrating on their position in the exhibition because the system will be able to perform that task automatically. Finally, on the other hand, exhibition organizers can focus their attention on the cultural information to be proposed, exploiting a flexible and dynamic means of providing and

updating such information without any need to spend time in preparing predetermined pathways around the exhibition.

The rest of the paper is structured as follows. In Section 2, we give an overview of related work summarizing the current use of mobile technologies in the cultural heritage domain. Sections 3 presents goals and explains the targets and introduces the concept of proximity interaction and the main components of the architecture. Afterwards, Section 4 focuses on the user interface of the proposed system describing the user experience for both visitors and curators. Finally, in Section 5, we present our conclusions.

2 Related Work

In the last few decades even more cultural institutions have increasingly recognized the requirement to better promote the diffusion of culture and education to a wider public. Moreover, the methods of cultural heritage information provision have changed, shifting from approaches based on showing collections of items to an expert and culturally prepared audience, to approaches based on entertainment and education aimed at capturing the interest of people of different ages, with different levels of education, and from different cultural backgrounds. Such considerations have led to a significant increase in the employment of non-invasive technology in the cultural heritage domain in order to improve the visitor's experience during, for example, a visit to a museum or historical site [4, 9, 16, 5].

The first proposed system employing mobile technology to interactively support the visitor during a museum visit made use of personal digital assistants (PDAs) provided with location-awareness based on the crossing of predefined gates [22, 7, 23]. Afterwards, a greater effort was made to provide cultural guides with context-awareness capabilities [17], by placing, for instance, RFID tags [13], smaller IrDA beacons [10] or Wi-Fi sensors [6] near to every artefact in the exhibition. The latest generation of visiting guides surpasses this capability, becoming a visitor's multimedia supports, designed not only to provide a contextual data about the artefacts, but also to improve the visitor's experience by offering a more engaging access to historical and cultural information [8, 18, 3].

In order to achieve these objectives a cheap technology, pervasively available on today's mobile devices, has been exploited. Very inexpensive Bluetooth-based beacons can be deployed in an exhibition to be precisely identified by a visitor's mobile device. The achieved proximity information can then be used to show the most appropriate information to the visitor and also to personalize her/his visit, thanks to an environment that becomes active [15]. The success of these small devices rests in the fact that they represent a trade-off between energy consumption, latency, coverage area and battery life [11]. Moreover, they prove to be suitable for museum installations due to their reduced dimensions and the limited time required for their deployment on site.

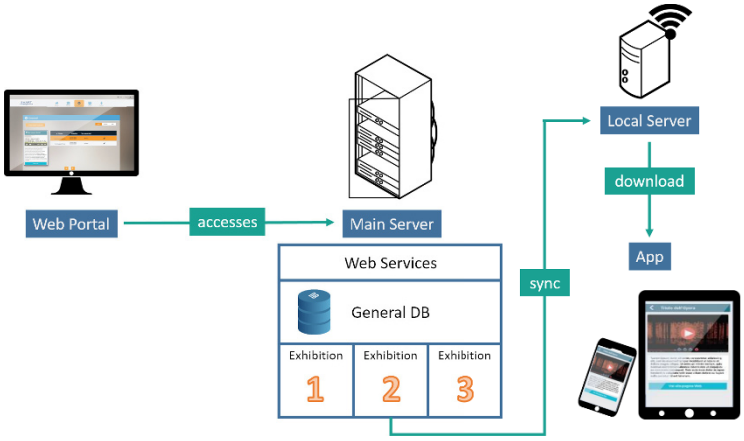


Fig. 2 Smart Underground system architecture.

Finally, until now, researchers have focused more on visitor needs with only a few attempts undertaken to support curators in making an effective use of exhibit-based information [12]. This aspect becomes all the more surprising considering the fact that curators nowadays, in almost all developed countries, are increasingly immersed in electronic archives of digitalized Cultural Heritage content [20] asking for an ICT support to manage growing-up complexity and privacy problems [1, 2].

3 System Description

3.1 Targets and Rationale

Most current cultural guides include the requirement for the visitor to follow a predefined path without allowing any free navigation around the exhibition area.

On the contrary, the proposed system has been designed to offer to the visitors an engaging experience of the cultural heritage allowing them to enjoy a more active interaction. The objective is to allow the visitors to freely move around the site and directly interact with all the works of art exhibited. By directly using her/his smartphone, the visitor will be able to dynamically choose for which artefact she/he wishes to access the proposed additional information. At the same time, the system represents an innovative working instrument also for the curators. In fact, the system was designed to be flexible responding to the needs of the organizers who, by using the system, do not need to make any physical and invasive modifications to the exhibition area each time a variation of the pathway is required. On the contrary, the flexibility of the system and the lack of any requirement to create marked pathways around the physical environment allow the curators to frequently update the information related to the works of art.

In more detail, the proposed system allows a visitor to:

- move freely around the exhibition or following proposed cultural pathway directly on her/his smartphone;
- receive a notification on her/his smartphone each time she/he is in the proximity of an artefact;
- choose for which of the nearest works of art to visualize more information; and
- store that information on her/his smartphone in order to continue the enjoyment of the visit even after leaving the exhibition.

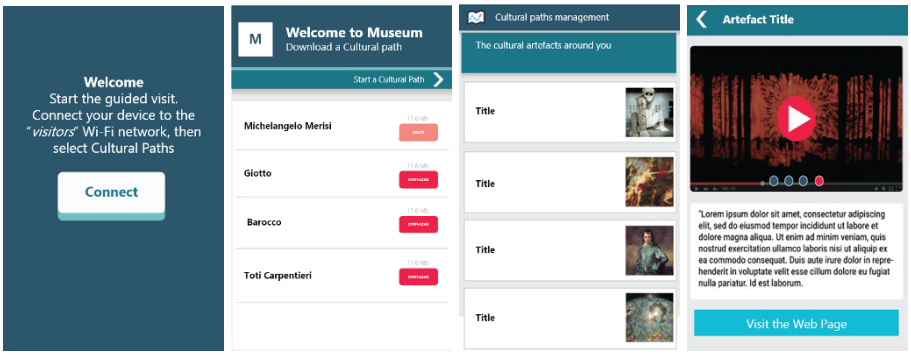
And at the same time, the system allows a curator to:

- organize the exhibition virtually by placing specific points of interest (POI) on a map;
- populate each POI with appropriate cultural contents;
- map each POI to a specific hardware component, a wireless beacon, easily installed in the museum;
- organize a set of POIs in thematic, temporal or special pathways within the exhibition; and
- organize multiple exhibitions at the same time, managing them from a single interface.

3.2 Proximity-Based Interaction

The realization of a system of this type has become possible thanks to the recent availability of a new generation of wireless beacons. These devices exploit an emerging low-power wireless technology called Bluetooth low energy (BLE) to broadcast their identifier to nearby mobile devices. BLE is the distinctive feature of the Bluetooth 4.0 specification [19], which enables a smartphone or any other mobile device to perform an action when it is in the proximity of a beacon. Moreover, the widespread use of Bluetooth technology in almost all current devices (e.g., wearable devices, mobile phones, and laptops) suggests that BLE is expected to be used in billions of devices in the near future [21], becoming important also in relation to the Internet of Things paradigm [14].

In an exhibition, each time a smartphone comes within the Bluetooth range of a beacon, the distance of the smartphone from the beacon can be determined and used to propose contextualized content to the visitor. Moreover, since these devices are very small and relatively cheap (see figure 1a), the system proves to be inexpensive to deploy and, more importantly, not invasive for a museum installation (see figure 1b). Because of the limited range of the Bluetooth reception, each beacon is physically placed close to the artefact to which it refers so that each time the visitor with her/his smartphone comes within the range of the beacon it will notify the proximity to the associated artefact. In this way, the physical integration between artefacts and digital devices modifies both the visitor's and curator's experience, enabling the device to adapt to personal preferences and needs.



(a) (b) (c) (d)

Fig. 3 Visitor's Experience. (a) When the visitor is near to the Local Server the mobile application on her/his device shows a welcome message and an invitation to connect to the server in order to download or update the cultural pathways available for the cultural site visited. (b) When the mobile app is connected to the Local Server all the available cultural pathways are shown to the visitor. (c) During the visit, without any need for a connection with the Local Server, the visitor's app is updated with a list of the nearest artefacts. (d) When the visitor selects an item from the proposed list, all the multimedia content provided by the curator becomes available to the visitor.

3.3 System Architecture

In order to achieve the aforementioned goals, the proposed system has been designed in the form of different components organized in a client/server architecture (see figure 2).

The Main Server: this server collects the cultural content relating to any artefact of each exhibition often organized in different places. Moreover, the server stores both the association of wireless beacons with cultural content and the collection of beacons used to define a specific cultural pathway.

The Web Portal: this represents the curator's interface with the system and allows her/him to access and modify the content of the Main Server. By using the Web Portal the curator can upload and update the multimedia content for each exhibition, modify the association of wireless beacons with cultural content, and finally, acting on a group of beacons to create different cultural pathways.

The Local Server: this server is physically present in each cultural site which hosts an exhibition. The server is used to create an ad-hoc Wi-fi network that allows the visitor to download the cultural pathways of the exhibition locally on her/his smartphone. The Local Server is synchronized with the content of the Main Server in order to propose to the visitor always the most recent content organized by the curator.

The Mobile App: this represents the visitor's interface with the proposed system and it is used to access the cultural information. The mobile app, when connected to the Local Server, allows the visitor to download and update her/his cultural pathways of interest. More importantly, during the visit, the mobile app allows an inter-

active enjoyment of the cultural information by exploiting the proximity sensing of the BLE technology.

4 User Interface Description

This section describes the user interface of the proposed system and explains how the experiences of both visitors and curators are improved.

4.1 Visitor's Experience Description

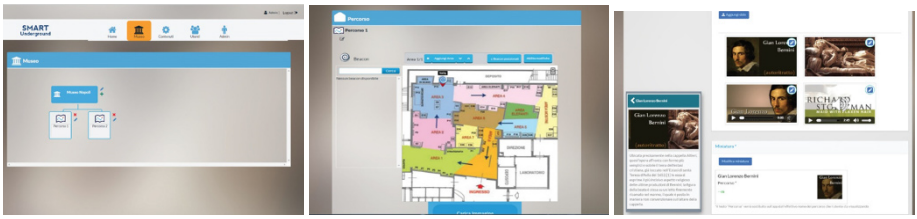
The cultural experience of the visitor starts at the entrance to the exhibition. In fact, at the entrance, through the ad-hoc Wi-fi of the Local Server, any visitor with the mobile app installed on her/his mobile device receives a welcome message to the exhibition and an invitation to connect to the system in order to download the cultural content or check for updates (see Figure 3a).

When the mobile device is connected to the Local Server the Smart Underground app shows all the predefined cultural pathways uploaded onto the server by the curator. The app allows the visitor to choose which one (if any) she/he wishes to download. Moreover, for any pathways already downloaded but subsequently modified by the curator, the app notifies the visitor of the availability of an update. Also in this case the visitor may choose whether or not to download the update (see Figure 3b).

With all the pathways of interest downloaded onto her/his mobile device, the visitor starts to move within the exhibition area. Each time the visitor comes within the proximity of a beacon the mobile app notifies her/him of the presence of a cultural content. Of course, there will sometimes be more than one cultural content close to the visitor's position so that in most cases a list of content is shown (see Figure 3c). In order to visualize all the multimedia information uploaded by the curator for each of the works of art in the surrounded area, the visitor needs to select the desired item from the list (see Figure 3d).

4.2 Curator's Experience Description

As already mentioned, the proposed system constitutes a support for all curators who wish to realize a more engaging experience for visitors to their exhibition. The most important functionality is the simplified management of multiple exhibitions. In fact, the architecture of the proposed system allows the management of multiple exhibitions from a single component with which the curator needs to interact, the Web Portal.



(a) (b) (c)

Fig. 4 Curator's Experience. (a) In order to enhance the enjoyment of a cultural visit the curator creates different cultural pathway for each exhibition. (b) By using a simple drag and drop interface the curator places the POIs on the exhibition map. (c) For each POI the curator uploads and updates all the multimedia contents to be proposed to the visitor.

Through this component the authorized curator can create for each exhibition many different cultural pathways, for instance based on the iconographic features of the works of art (see Figure 4a). Effectively, the creation of a cultural pathway corresponds to the design of an experience that the curator plans to show to the visitor through an additional layer of information. Therefore, the next step requires the curator to populate the cultural pathway with a number of POIs placed near the related artefacts. In order to facilitate this step, the system proposes a drag and drop solution with which the user picks up a beacon and places it on the map representing the exhibition area (see Figure 4b). Whenever a beacon is placed on the map it is considered active. Finally, the curator can select any of the active beacons on the map on which to upload the desired multimedia content (see Figure 4c).

In this way, the proposed system greatly simplifies the design of a cultural pathway, which becomes flexible and easy to change every time it is necessary. In fact, the curator can delete, modify or create new pathways whenever she/he wants. Moreover, a single POI can be assigned to different cultural pathways in order to address those situations in which one POI is relevant to several different artefacts. Additionally, different multimedia content can be assigned to the same POI in order to offer to the visitor the same content in different languages without the need to use multiple beacons for the same artefact. Finally, a temporal validity can be assigned to each POI after which deadline the information will no longer be accessible to the visitors, in order to facilitate the management of temporary exhibitions or daily events.

Obviously, each of the activated beacons in the system needs to be effectively placed in the physical environment immediately after the end of the design phase performed on the Web Portal.

5 Conclusions

In this paper, we have described a cultural system that, while enhancing the access to cultural heritage information for the visitors, intends also to be an innovative instrument for museum curators. The system exploits the visitorâĂŹs smartphone and BLE technology for a fully-featured, proximity-based exploration of the works of art in an exhibition. At the same time, it reduces the effort required of the curator in managing a cultural exhibition in that, by using a single interface, she/he can continuously create new cultural content in order to offer always a different experience to the visitors. In particular, the proposed system modifies both the visitor's and the curator's experiences, making them more involved respectively in the assimilation and promotion of cultural heritage content.

In our future research we will focus on exploiting this system to collect data about the travelling routes made and the time spent near each artefact by the visitors in order to improve the support for exhibition curators. Finally, the system is going to be installed in the cultural site "'Basilica di Santa Maria Maggiore"' called "'La Pietrasanta"' in Naples and, at that time, we will start an evaluation of the system interfaces in order to collect both visitors' and curators' user feedback.

Acknowledgments

The proposed system has been developed together with the ICT company QUICON within the project Smart Underground, founded by the Regione Campania through P.O.R. CAMPANIA 2007-2013, under the call Sportello dell'Innovazione - Progetti Cultural and Creative Lab.

References

1. Amato, F., Moscato, F.: A model driven approach to data privacy verification in e-health systems. Transactions on Data Privacy 8(3), 273–296 (2015)
2. Amato, F., Moscato, F.: Exploiting cloud and workflow patterns for the analysis of composite cloud services. Future Generation Computer Systems (jul 2016), http://dx.doi.org/10.1016/j.future.2016.06.035
3. Ardito, C., Costabile, M.F., Lanzilotti, R., Simeone, A.L.: Combining multimedia resources for an engaging experience of cultural heritage. In: Proceedings of the 2010 ACM workshop on Social, adaptive and personalized multimedia interaction and access. pp. 45–48. ACM (2010)
4. Buzzi, M., Buzzi, M., Leporini, B., Marchesini, G.: Smartweet: A location-based smart application for exhibits and museums. In: Proceedings of the IADIS - Interfaces and Human Computer Interaction. pp. 327–331 (2013)
5. Caggianese, G., Gallo, L., De Pietro, G.: Design and preliminary evaluation of a touchless interface for manipulating virtual heritage artefacts. In: Signal-Image Technology and Internet-Based Systems (SITIS), 2014 Tenth International Conference on. pp. 493–500. IEEE (2014)

6. Chianese, A., Marulli, F., Moscato, V., Piccialli, F.: Smartweet: A location-based smart application for exhibits and museums. In: Signal-Image Technology & Internet-Based Systems (SITIS), 2013 International Conference on. pp. 408–415. IEEE (2013)
7. Ciavarella, C., Paternò, F.: Visiting a museum with an handheld interactive support. Demo at Mobile HCI (2002)
8. Costabile, M.F., De Angeli, A., Lanzilotti, R., Ardito, C., Buono, P., Pederson, T.: Explore! possibilities and challenges of mobile learning. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 145–154. ACM (2008)
9. Dini, R., Paternò, F., Santoro, C.: An environment to support multi-user interaction and cooperation for improving museum visits through games. In: Proceedings of the 9th International Conference on Human Computer Interaction with Mobile Devices and Services. pp. 515–521. MobileHCI '07, ACM, New York, NY, USA (2007), http://doi.acm.org/10.1145/1377999.1378062
10. Fleck, M., Frid, M., Kindberg, T., Rajani, R., OâĂŹBrien-Strain, E., Spasojevic, M.: From informing to remembering: Deploying a ubiquitous system in an interactive science museum. IEEE pervasive computing 1(2), 13–21 (2002)
11. Gomez, C., Oller, J., Paradells, J.: Overview and evaluation of bluetooth low energy: An emerging low-power wireless technology. Sensors 12(9), 11734–11753 (2012)
12. Hsi, S.: I-guides in progress: two prototype applications for museum educators and visitors using wireless technologies to support science learning. In: Wireless and Mobile Technologies in Education, 2004. Proceedings. The 2nd IEEE International Workshop on. pp. 187–192. IEEE (2004)
13. Hsi, S., Fait, H.: Rfid enhances visitors' museum experience at the exploratorium. Communications of the ACM 48(9), 60–65 (2005)
14. Hui, J.W., Culler, D.E.: Extending ip to low-power, wireless personal area networks. IEEE Internet Computing 12(4), 37–45 (2008)
15. Kuflik, T., Stock, O., Zancanaro, M., Gorfinkel, A., Jbara, S., Kats, S., Sheidin, J., Kashtan, N.: A visitor's guide in an active museum: Presentations, communications, and reflection. J. Comput. Cult. Herit. 3(3), 11:1–11:25 (Feb 2011), http://doi.acm.org/10.1145/1921614.1921618
16. Ott, M., Pozzi, F.: Towards a new era for cultural heritage education: Discussing the role of ict. Computers in Human Behavior 27(4), 1365–1371 (2011)
17. Schilit, B., Adams, N., Want, R.: Context-aware computing applications. In: Mobile Computing Systems and Applications, 1994. WMCSA 1994. First Workshop on. pp. 85–90. IEEE (1994)
18. Stock, O., Zancanaro, M.: PEACH-Intelligent interfaces for museum visits. Springer Science & Business Media (2007)
19. The Bluetooth Special Interest Group: Kirkland, WA, U.: Specification of the bluetooth system, covered core package, version: 4.0 (2010)
20. Unesco: Information and communication technologies in schools - a handbook for teachers. (2005)
21. West, A.: Smartphone, the key for bluetooth low energy technology. Dostępne w Internecie: www. bluetooth. com/Pages/Smartphones. aspx (2014)
22. Woodruff, A., Aoki, P.M., Hurst, A., Szymanski, M.H.: Electronic guidebooks and visitor attention. In: ICHIM (1). pp. 437–454 (2001)
23. Yatani, K., Onuma, M., Sugimoto, M., Kusunoki, F.: Musex: A system for supporting children's collaborative learning in a museum with pdas. Systems and Computers in Japan 35(14), 54–63 (2004)

Ciphertext-Policy Attribute Based Encryption with Large Attribute Universe

Siyu Xiao, Aijun Ge, Fushan Wei and Chuangui Ma

Abstract Ciphertext-policy attribute-based encryption(CP-ABE) has become a crucial technical for cloud computing in that it enables one to share data with users under the access policy defined by himself. Generally, the universe of attributes is not fixed before the system setup in practice. So in this paper, we propose a CP-ABE scheme with large attribute universe based on the scheme presented by Chen et al. The number of attributes is independent of the public parameter in our scheme, and it inherents the excellent properties of both constant ciphertext and constant computation cost.

1 Introduction

With the development of cloud computing technology, more and more clients are willing to store and distribute their large scale of data on a cloud server. Meanwhile, there has already emerged many well-known service providers such as Google storage cloud, Amazon' S3 and so on. Despite the fact that such cloud service offers great convenience to users, it has indeed introduced some non-negligible threats. For example, cloud storage system is fully public to which everyone can have access, so the data privacy seems impossible in this way.

One method to figure this out is encrypting the data before it being outsourced to the cloud. Thus, malicious clients won't gain any useful information about the data even if they corrupt the service provider. Nevertheless, this will make it diffi-

Siyu Xiao, Aijun Ge, Fushan Wei
State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou, China
e-mail: siyuxiao32@163.com, geaijun@163.com, weifs831020@163.com

Chuangui Ma
Department of Basic Courses, Army Aviation Institute, Beijing, China
State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou, China
e-mail: chuanguima@sina.com

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_11

cult for users to selectively share their encrypted data under a fine-grained policy. Suppose at some point, a user wants to distribute a sensitive encrypted document, and only the "women" in "finance department" of her company have the ability to decrypt. The concept introduced by [1] called Attribute-Based Encryption(ABE) makes some important step to solve this problem. In an ABE scheme, each user's key and each ciphertext are associated with a set of attributes respectively. If and only if there exists a match between the user's attributes and the ciphertext's attributes, he can have the ability to decrypt. Later, many researchers make further efforts to achieve more fine-grained access policy.

ABE can be divided into Ciphertext-Policy ABE(CP-ABE) and Key-Policy ABE(KP-ABE). In CP-ABE scheme, the secret key is associated with a set of attributes while the ciphertext is associated with an access policy. A user then can have the ability to decrypt a ciphertext if and only if his attributes related to the secret key satisfy the policy . In KP-ABE scheme, the ciphertext is associated with a set of attributes and the secret key is associated with an access policy. A user then can have the ability to decrypt a given ciphertext if and only if the underlying set of attributes related to the ciphertext satisfies the policy. In this paper, we mainly consider CP-ABE in which data owners can decide whether or not one have the authority to share his data. [2] proposed a CP-ABE with constant-size ciphertext and constant computation cost, but their scheme is established on small attribute universe. That is to say, the number of attributes is fixed before the system setup, which is not satisfying current tendency, for example, in big data sharing, where the user authority is decided by his attributes. In this paper, we aim to figure out a solution on large attribute universe.

When it comes to the security of ABE, the most important thing we consider is to resist collusion. A group of members cannot decrypt a ciphertext if neither of them can. For example, if an access policy associated with a ciphertext is "cryptography AND doctor", then a cryptography master and a economics doctor cannot decrypt this ciphertext even though they can get attributes "cryptography" and "doctor" via collusion. How to avoid collusion attacks is always a hot research area, and also a difficult research area.

Our Results. We propose a CP-ABE scheme supporting large attribute universe. In this scheme, the number of attributes is independent of the public parameter as long as each user's personal number of attributes is less than the upper bound. Furthermore, it inherits the excellent properties in [2] and has constant ciphertext length as well as consant computation cost. The security of our scheme can be proved under the n-BDHE assumption.

Related Work. Attribute-Based Encryption(ABE) is actually an extension of Identity-Based Encryption(IBE) to improve the flexibility of users sharing their data. It is first introduced by [1] and be further classified into CP-ABE and KP-ABE by [3]. The first KP-ABE scheme proposed by [3] achieves monotonic access policy. Later, [4] propose another KP-ABE scheme supporting non-monotone key polices to increase the expressiveness. In 2007, [5] present the first construction of CP-ABE realizing

tree-based access policy, but its security is proved in generic group models. Then [6] propose a CP-ABE scheme with security in the standard model, however, it can just support AND gate operation. Until now, research on realizing fine-grained access policy ABE with security under the standard model is still a hot area, also a challenging area.

Besides the expressiveness and security, there also exists another point deserved our attention. That is the computation cost of the scheme, both in encryption and decryption. [7] initiate the study of CP-ABE with constant-size ciphertext but it supports just simple AND gate operation. Subsequent results [2] and [8] are the same. Afterwards, [9] propose a threshold CP-ABE scheme with constant-size ciphertext and can be extended to realize resistance against Chosen-Ciphertext-Attack(CCA). Fully secure threshold CP-ABE with constant-size ciphertext is achieved by [10] via a universal transformation from Inner Product Encryption(IPE) and can be further extended to a large attribute universe scheme. The only flaw is the foundation of composite-order bilinear groups. [11] propose a CP-ABE scheme with constant-size keys and expressive access policy so that lightweight devices can be used as storage for decryption keys. In this paper, motivated by all the existing results, we propose a construction of CP-ABE with large attribute universe based on prime-order bilinear groups, which inherits the good properties of constant-size ciphertext in [2] at the same time.

Organization. The remainder of the paper is organized as follows. In section 2, some preliminaries are reviewed including the definition of bilinear group and the syntax of CP-ABE. Next, we present our concrete scheme and give necessary proof of the security. Finally, we give conclusion in section 4.

2 Preliminary

2.1 Bilinear Group

Let $\mathscr{G}$ be an algorithm that take s input a security parameter k and outputs a tuple (p, G, G_T, g, e), where G and G_T are cyclic groups of order p for some large prime p, g is a generator of G. The map $e : G \times G \to G_T$ satisfies the following properties:
1. Bilinear: $e(u^a, v^b) = e(u, v)^{ab}$ for all $u, v \in G$ and $a, b \in Z_p$.
2. Non-degenerate: $e(g, g) \neq 1$.

We say G generated in this way is a bilinear group if the group operation in G and the map e are efficiently computable.

Let G be a bilinear map of prime order p defined above, g, h be two independent generators of G. Denote $\overrightarrow{y}_{g,\alpha,n} = (g_1, g_2, \ldots, g_n, g_{n+2}, \ldots, g_{2n}) \in G^{2n-1}$, where $g_i = g^{\alpha^i}$. For an adversary $\mathscr{A}$, we define $Adv^{\text{n-BDHE}}_{G,\mathscr{A}}(k)$ as follows:

$$|Pr[\mathscr{A}(g, h, \overrightarrow{y}_{g,\alpha,n}, e(g_{n+1}, h)) = 0] - Pr[\mathscr{A}(g, h, \overrightarrow{y}_{g,\alpha,n}, Z) = 0]|$$

where $Z \in G_T$ and $\alpha \in Z_p$ are all randomly chosen. We say that the decision n-BDHE assumption holds in G if $Adv_{G,\mathscr{A}}^{\text{n-BDHE}}(k)$ is negligible for arbitrary polynomial adversary $\mathscr{A}$.

The security proof of our scheme is based on the above decision n-BDHE assumption.

2.2 Ciphertext Policy ABE

A CP-ABE system consists of four probabilistic polynomial-time algorithms **Setup**, **KeyGen**, **Encrypt**, **Decrypt** as follows:

Setup(1^k): Takes input the security parameter k and outputs the system public parameter PP and master private key MK. PP is distributed to users while MK kept secret.
KeyGen(PP, MK, S): Takes input the private key MK and a attribute set S, outputs SK_S as the secret key for the user associated with S.
Encrypt(PP, M, Ω): Takes input a message M under the access policy Ω, the algorithm outputs a ciphertext C_Ω using the public parameter PP.
Decrypt(PP, SK_S, C_Ω): Takes input the users' secret key SK_S and ciphertext C_Ω associated with access policy Ω, outputs the message M if S satisfies Ω and $\perp$ otherwise.

For an access policy Ω, we mean it is a rule that returns either 0 or 1 given a set of attributes S. If S satisfies Ω, it will return 1. Otherwise, it will return 0. Actually, arbitrary boolean functions, threshold trees can be an access policy. In this paper, we mainly consider AND gate.

2.2.1 Security Model

The selective security model against chosen plaintext attacks for CP-ABE can be defined via the following IND-sCP-CPA game. In this game, a challenge access policy Ω is supposed to be chosen before **Setup** and the adversary is allowed to query keys for any attribute set S that is not satisfied by Ω.
1). The adversary $\mathscr{A}$ chooses a challenge access policy Ω and gives it to the challenger $\mathscr{C}$.
2). $\mathscr{C}$ runs the algorithm **Setup** to generates public parameter PP and master secret key MK. Then it gives PP to $\mathscr{A}$.
3). $\mathscr{A}$ adaptively queries keys for any attribute set S that is not satisfied by Ω. $\mathscr{C}$ runs **KeyGen**(PP, MK, S) and returns SK_S to the adversary.
4). At some point, $\mathscr{A}$ outputs two equal length messages M_0 and M_1. The challenger randomly chooses a bit $b \in \{0,1\}$ and computes **Encrypt**(PP, M_b, Ω). It then sends CT_Ω to the adversary $\mathscr{A}$.

5). $\mathscr{A}$ can additionally make key queries for attribute sets not satisfying Ω and $\mathscr{C}$ responds the same as above.
6). $\mathscr{A}$ outputs a guess bit $b' \in \{0,1\}$ and wins the game if $b' = b$.

The advantage of an adversary in the above game is defined as follows:

$$Adv_{\mathscr{A}}^{\text{IND-sCP-CPA}}(k) = |Pr[b' = b] - \frac{1}{2}|$$

Definition 1. A CP-ABE scheme is said to be IND-sCP-CPA secure if no probabilistic polynomial-time adversary can have non-negligible advantage in the above game.

3 Our Scheme

In this paper, the access policy we mainly consider is AND gate $\bigwedge_{A_i \in U} A_i$. Actually, if denoting $\neg A_i$ as an individual attribute in the system, NOT gate is supported as well. Here, we just omit this part for simplicity.

Setup(k,n): Takes as input the security parameter k and the maximum number of attributes n in the system, the algorithm first runs $\mathscr{G}(1^k)$ and generates bilinear maps (p,G,G_T,g,e). Then it randomly chooses two polynomials $p_1(x)$, $p_2(x)$ in Z_p with order $n-1$, and sets $R_i = g^{-p_1(r_i)}$, $U_i = e(g^{p_2(r_i)},g)$ where $r_i \in Z_p$ are randomly chosen for $i = 1,\ldots,n$.

The public parameter is

$$PP = \{g, < r_i, R_i, U_i >_{i=1,\ldots,n}\}$$

The private key is

$$MK = \{p_1(x), p_2(x)\}.$$

KeyGen(PP,MK,S): Takes an attribute set S as input, the algorithm randomly chooses $V \in G$, and computes $\sigma_j = g^{p_2(j)} V^{p_1(j)}$ for $j \in S$.

The secret key for the user is $SK_S = \{V, \{\sigma_j\}_{j \in S}\}$.

Encrypt(PP,M,Ω): The encryption algorithm encrypts a message $M \in G_T$ under the AND policy $W = \bigwedge_{j \in \Omega} j$. It chooses random element $t \in Z_p$, then computes $C_0 = M \cdot (\prod_{j \in \Omega} U_j)^t, C_1 = (\prod_{j \in \Omega} R_j)^t, C_2 = g^t$. Where $U_j = e(g,g)^{p_2(j)}$ and $R_j = g^{-p_1(j)}$ for $j \in \Omega$ can be interpolation calculated using the public parameter.

The ciphertext for message M is $CT_\Omega = \{\Omega, C_0, C_1, C_2\}$.

Decrypt(PP,SK_S,CT_Ω): The decryption algorithm first checks whether $\Omega \subseteq S$. If

not, return $\perp$. Otherwise, computes $\sigma = \prod_{j\in\Omega} \sigma_j$ and outputs $M = \frac{C_0}{e(V,C_1)\cdot e(\sigma,C_2)}$ as the plaintext.

Theorem 1. *Suppose the decisional n-BDHE assumption holds in G. Then no polynomial time adversary can win the IND-sCP-CPA game defined in section 2.2 with non-negligible probability.*

Proof. Our proof of the security is almost the same with that of [2] except some little difference in secret key generation.

Suppose there exists a simulator $\mathscr{S}$ with n-BDHE inputs $(g, g^s, \overrightarrow{y}_{g,\alpha,n}, T)$, then $\mathscr{S}$ can simulate the IND-sCP-CPA game via following steps:

Initiation. The adversary $\mathscr{A}$ sends $\mathscr{S}$ a challenge access policy $\mathbb{W} = \bigwedge_{i\in\Omega} i$.

Setup. The simulator randomly chooses $i^* \in \Omega$ and random elements $r_k, a_k \in Z_p$, $k = 1,\ldots,n$. Then it computes

$$(R_{i^*}, U_{i^*}) = (g^{r_{i^*}}(\prod_{k\in\Omega, k\neq i^*} g_{n+1-k}), e(g,g)^{a_{i^*}} e(g,g)^{\alpha^{n+1}})$$

For $i \in \Omega \setminus \{i^*\}$

$$(R_i, U_i) = (g^{r_i} g_{n+1-i}^{-1}, e(g,g)^{a_i})$$

Then $n - |\Omega|$ other random elements are chosen and for every element i in it, computes

$$(R_i, U_i) = (g^{r_i}, e(g,g)^{a_i})$$

Let U denotes all the attributes i mentioned above and sends $\mathscr{A}$ the public parameter $< i, R_i, U_i >_{i\in U} = < i, g^{-p_1(i)}, e(g^{p_2(i)}, g) >_{i\in U}$.

Key Queries. The adversary can query keys for attribute set $w(\Omega \subsetneq w)$. The simulator chooses $i' \in \Omega \setminus w$ and random $r \in Z_p$, computes $V = g_{i'} g^r$.

For attribute in U, just computes $\sigma_i = g^{p_2(i)} V^{p_1(i)}$.

For attribute not in U, the secret key can be computed utilizing interpolation calculation $\sigma_i = g^{\sum_{j\in U} l_j(i) p_2(j)} V^{\sum_{j\in U} l_j(i) p_1(j)} = \prod_{j\in U} (g^{p_2(j)} V^{p_1(j)})^{l_j(i)} = \prod_{j\in U} \sigma_j^{l_i(j)}$, where $l_j(i) = \prod_{k\in U, k\neq j} \frac{i-k}{j-k}$.

Challenge. At some point, $\mathscr{A}$ outputs two equal length messages M_0 and M_1, the simulator chooses random bit $b \in \{0,1\}$ and computes

$$CT = (C_0 = M_b \cdot Te(g,g^s)^{a_\Omega}, C_1 = g^{s r_\Omega}, C_2 = g^s)$$

where $a_\Omega = \sum_{i\in\Omega} a_i$, $r_\Omega = -\sum_{i\in\Omega} r_i$.

Guess. Finally, if $\mathscr{A}$ guess $b' = b$, $\mathscr{S}$ outputs 0; otherwise, he outputs 1.

We can known if $T = e(g_{n+1}, g^s)$, CT is a valid encryption of message M_b; otherwise, it's a ciphertext for a random message. Thus, if the adversary wins the game with probability ε, then the simulator will attack the problem of n-BDHE with probability $1/2\varepsilon$.

4 Conclusion and Future Work

In this work, we construct a CP-ABE scheme with large attribute universe based on the results of [2]. The number of attributes in the proposed system is independent of the public parameter. That is to say, once any user gains a new attribute, he can add it to the system as long as his personal number of attributes less than the upper bound. It is crucial in cloud storage system for not necessary to fix how many attributes all together at the beginning, thus is more flexible. Though the added payment are just two interpolation operations for every encryption, our scheme only supports a restricted access policy, which is AND gate. How to achieve ABE with more expressive access policy with large attribute universe while maintaining constant-size ciphertext is what we will continue to investigate in the future.

Acknowledgements The authors would like to thank the anonymous referees for their helpful comments. This work is supported by the National Natural Science Foundation of China (Nos. 61309016, 61379150,61602512).

References

1. Sahai A. and Waters B.: Fuzzy identity based encryption. In: Proc. Advances in Cryptology-Eurocrypt, pp.457-473(2005)
2. Chen C., Zhang Z., and Feng D.: Efficient ciphertext-policy attribute-based encryption with constant-size ciphertext and constant computation-cost. In: Proc. ProveSec'11, pp.84-101(2011)
3. Goyal V., Pandey O., Sahai A. and Waters B.: Attribute-based encryption for fine-grained access control of encrypted data. In: Proc. CCS'06, pp.89-98(2006)
4. Ostrovsky R., Sahai A. and Waters B.: Attribute-based encryption with nonmonotonic access structures. In: Proc. ACM Conference on Computer and Communication Security, pp.195-203(2007)
5. Bethencourt J., Sahai A. and Waters B.: Ciphertext-policy attribute-based encryption. In: Proc. IEEE Symposium on Security and Privacy, pp.321-334(2007)
6. Cheung L. and Newport C.: Provably secure ciphertext policy abe. In: Proc. ACM Conference on Computer and Communication Security, pp.456-465(2007)
7. Emura K., Miyaji A., Nomura A., et al.: A ciphertext-policy attribute-based encryption scheme with constant ciphertext length. In: Proc. ISPEC'09, pp.13-23(2009)
8. Zhou Z., and Huang D.: On efficient ciphertext-policy attribute-based encryption and broadcast encryption. In: Proc. CCS'10, pp.753-755(2010)
9. Ge A., Zhang R., Chen C., Ma C.and Zhang,Z.: Threshold ciphertext-policy attribute-based encryption with constant-size ciphertexts. In: Proc. ACISP'12, pp.336-349(2012)

10. Chen C., Chen J., Lim H., et al.: Fully secure attribute-based systems with short ciphertexts/signatures and threshold access structures. In: Proc. CT-RSA'13, pp.50-67(2013)
11. Guo F., Mu Y., Susilo W., Wong D.and Varadharajan,V.: CP-ABE with constant-size keys for lightweight devices. In: IEEE Trans. Inf. Forensics Security, vol.9, no.5, pp.763-771(2014)

Asymmetric Searchable Encryption from Inner Product Encryption

Siyu Xiao, Aijun Ge, Jie Zhang, Chuangui Ma and Xu'an Wang

Abstract Asymmetric searchable encryption(ASE) enables one to retrieve encrypted data stored on an untrusted server without revealing the contents. Now, beyond single keyword search, more and more attention have already been paid to the problem of multi-keyword search. However, existing schemes are mainly based on composite-order bilinear groups. In this paper, we propose a public key encryption with conjunctive and disjunctive keyword search(PECDK) scheme which can simultaneously support conjunction and disjunction within each keyword field for cloud storage. It is based on prime-order bilinear groups, and can be proved fully secure under the standard model.

1 Introduction

Kamara et al.[1] point out that there has been a tendency for clients with limited resources, to store and distribute their data in a public cloud. Instead of building and maintaining data centers of their own, now they just remotely store their data and then enjoy benefits. This is what we called cloud storage service and there actually exists many well-known service providers such as Amazon's S3, Google Cloud Storage and so on. Simultaneously, there also exists many works [3, 4, 5, 6] concen-

Siyu Xiao, Aijun Ge, Jie Zhang
State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou, China
e-mail: siyuxiao32@163.com, geaijun@163.com, zhangjie902@sina.cn

Chuangui Ma
Department of Basic Courses, Army Aviation Institute, Beijing, China
State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou, China
e-mail: chuanguima@sina.com

Xu'an Wang
Engineering University of CAPF, Xi'an, China
e-mail: wangxazjd@163.com

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_12

trated on research about the cloud. Despite all the advantages it brings, it has indeed introduced some new treats and challenges due to the fact that data is uploaded to a public platform to which everyone can have access [2].

For confidentiality, data are usually encrypted before being sent to the cloud. Nevertheless, this will make it difficult when the client wants to download only parts of the documents needed. Song et al.[7] initiate the investigation on searchable encryption, which can be divided into symmetric searchable encryption and asymmetric searchable encryption, enabling clients to search encrypted data but still guarantees the privacy.

Public key encryption with keyword search(PEKS), proposed by Boneh et al.[8], is one of the variants of asymmetric searchable encryption. In a PEKS scheme, a sender encrypts a keyword w' under Alice's public key and sends the ciphertext to the server. Alice can then provide the server with a trapdoor T_w(computed as a function of her private key) for any keyword w of her choice, and enables the latter to learn whether or not $w' = w$ but nothing else about w'.

Later, several solutions are presented in [10, 13, 15, 17, 18] to improve the efficiency of the PEKS system or provide stronger security. Nonetheless, schemes above mainly consider the situation of single keyword search. That means, once Alice is actually interested in documents containing several keywords, we must either use set-intersection or meta-keywords. However, both these two solutions are not appropriate, in that the former will reveal the relationship between each document and each keyword, and the latter will consume an exponentially growing storage for the number of keywords. It is [9] that initiate the research on public key encryption with conjunctive keyword search(PECK). Then Lee et al.[12] present a PECK scheme with a short ciphertext size and one private key, but it is established in the random oracle.

Later, Boneh and Waters[11] propose hidden vector encryption(HVE) supporting conjunctive, subset and range queries over encrypted data, however it's based on composite-order bilinear groups and can only achieve selective security. In [14], Katz et al. present the concept of predicate encryption that can also be used to achieve conjunctive keyword search. In predicate encryption, secret keys are corresponding to predicates and ciphertexts are associated with attributes; the secret key SK_f corresponding to a predicate f can be used to decrypt a ciphertext associated with attribute I if and only if $f(I) = 1$. When the predicate f is corresponding to the evaluation of inner products over Z_N, we call it inner product encryption(IPE). Recently, a new idea called expressive and secure asymmetric searchable encryption(ESASE) is presented by [19] to realize adaptively secure multi-keyword search in the standard model. Their scheme can simultaneously support conjunctive, disjunctive and negation search operation. But, it is also based on bilinear groups with composite order N which is the multiplication of four primes.

As we can see, HVE, IPE, ESASE all can achieve fairly expressive search queries. However, all of them are concentrated on disjunctions between keyword fields rather than disjunctions within every keyword field, i.e search queries with restriction of the first keyword field "A_1 or B_1". To deal with this dilemma, we propose the concept public key encryption with conjunctive and disjunctive key-

word search(PECDK), and present a concrete scheme based on prime-order bilinear groups which is fully secure under the standard model.

Our results. Based on the excellent properties of IPE, we present a scheme supporting conjunctions and disjunctions within each keyword field named PECDK. It is based on prime-order bilinear groups and can be proved to be computational consistent as well as fully secure in the standard model.

Organization. The remainder of this paper is organized as follows. In section 2, we introduce some preliminaries. In section 3, we give our concrete scheme and provide necessary proofs for consistency and security. Conclusion is given in section 4.

2 Preliminary

2.1 Dual Paring Vector Spaces

Definition 1. "Dual Paring Vector Spaces(DPVS)" $(q, \mathbf{V}, G_T, \mathbf{A}, \hat{e})$ by a direct product of bilinear groups (q, G, G_T, P, e) are a tuple of prime q, N-dimensional vector space $\mathbf{V} = \overbrace{G \times \cdots \times G}^{N}$ over Z_q, canonical basis $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_N)$ of $\mathbf{V}$, where $\mathbf{a}_i = (\overbrace{0, \ldots, 0}^{i-1}, P, \overbrace{0, \ldots, 0}^{N-i})$, and paring $\hat{e} : \mathbf{V} \times \mathbf{V} \to G_T$.

The paring is defined by $\hat{e}(\mathbf{x}, \mathbf{y}) = \prod_{i=1}^{N} e(x_i P, y_i P)$, where $\mathbf{x} = (x_1 P, \ldots, x_N P) \in \mathbf{V}$ and $\mathbf{y} = (y_1 P, \ldots, y_N P) \in \mathbf{V}$, $x_i, y_i \in Z_q$ for all i. Obviously, $\hat{e}$ satisfies following properties:

1. Bilinear: $\hat{e}(s\mathbf{x}, t\mathbf{y}) = \hat{e}(\mathbf{x}, \mathbf{y})^{st}$ for all $s, t \in Z_q$ and $\mathbf{x}, \mathbf{y} \in \mathbf{V}$.
2. Non-degenerate: If $\hat{e}(\mathbf{x}, \mathbf{y}) = 1$ for all $\mathbf{y} \in \mathbf{V}$, then $\mathbf{x} = \mathbf{0}$.
3. Computable: There is a polynomial time algorithm to compute $\hat{e}(\mathbf{x}, \mathbf{y})$ for all $\mathbf{x}, \mathbf{y} \in \mathbf{V}$.

Therefore, we can naturally view DPVS as an extension of bilinear groups. Besides, DPVS can do linear transformation on canonical basis $\mathbf{A}$. Let $\mathbf{G}_{dpvs}$ be an algorithm that takes input security parameter and $N \in \mathbf{N}$,and outputs a DPVS description $(q, \mathbf{V}, G_T, \mathbf{A}, \hat{e})$. We describe random dual orthogonal bases generator $\mathbf{G}_{ob}$ as follows.

$\mathbf{G}_{ob}(1^k, N)$: $param_{\mathbf{V}} = (q, \mathbf{V}, G_T, \mathbf{A}, \hat{e}) \xleftarrow{R} G_{dpvs}(1^k, N)$,

$X = (x_{i,j}) \xleftarrow{U} GL(N, F_q)$, $(v_{i,j}) = (X^T)^{-1}$,

$\mathbf{b}_i = \sum_{j=1}^{N} x_{i,j} \mathbf{a}_j$, $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_N)$,

$\mathbf{b}_i^* = \sum_{j=1}^{N} v_{i,j}\mathbf{a}_j$, $\mathbf{B}^* = (\mathbf{b}_1^*, \ldots, \mathbf{b}_N^*)$,
return $(param_{\mathbf{V}}, \mathbf{B}, \mathbf{B}^*)$.

2.2 Inner Product Encryption

2.2.1 Syntax

An IPE scheme for the attribute space $\Sigma = Z_N^n / \{0^n\}$ consists of four probabilistic polynomial-time algorithms as follows:
Setup$(1^k, n)$: Takes a security parameter 1^k and n as input and outputs the public key pk and master secret key sk.
Enc$(pk, \mathbf{x}, m)$: Takes the public key pk, attribute vector $\mathbf{x} \in \Sigma$ and message m as input and outputs a ciphertext $C_{\mathbf{x}}$.
KeyGen$(sk, \mathbf{y})$: Takes the master secret key sk and attribute vector $\mathbf{y} \in \Sigma$ as inputs and outputs a secret key $sk_{\mathbf{y}}$.
Dec$(pk, C_{\mathbf{x}}, sk_{\mathbf{y}})$: Takes a ciphertext associated with attribute vector $\mathbf{x}$ and the secret key for $\mathbf{y}$ as input, outputs plaintext m if and only if $< \mathbf{x}, \mathbf{y} > \equiv 0 (modN)$. Otherwise outputs $\perp$.
Correctness. For all $\mathbf{x}, \mathbf{y} \in \Sigma$, the corresponding ciphertext $C_{\mathbf{x}} \leftarrow \text{Enc}(pk, \mathbf{x}, m)$ for arbitrary plaintext m from the message space and corresponding secret key $sk_{\mathbf{y}}$, it holds that $m = \text{Dec}(pk, C_{\mathbf{x}}, sk_{\mathbf{y}})$ if $< \mathbf{x}, \mathbf{y} >= 0$. Otherwise, it holds with negligible probability.

2.2.2 Security Model

Reference [16] defines an IPE scheme to be weakly attribute-hiding against chosen plaintext attacks (AH-CPA) via the following game.
1. The challenger $\mathscr{C}$ runs the algorithm *Setup* to get pk, sk, and pk is given to the adversary $\mathscr{A}$.
2. $\mathscr{A}$ can adaptively query secret keys for polynomial number of attribute vectors $\mathbf{y}$, $\mathscr{C}$ responds with $sk_{\mathbf{y}} \leftarrow \text{KeyGen}(sk, \mathbf{y})$.
3. At some point, $\mathscr{A}$ outputs two attribute vectors $\mathbf{x}_0, \mathbf{x}_1 \in \Sigma$ and two challenge messages $m_0, m_1 \in M$. The only restriction is that $< \mathbf{y}, \mathbf{x}_0 > \neq 0, < \mathbf{y}, \mathbf{x}_1 > \neq 0$ for all queried vectors $\mathbf{y}$.
4. $\mathscr{C}$ randomly chooses a bit $b \in \{0, 1\}$, and returns $C_b \leftarrow \text{Enc}(pk, \mathbf{x}_b, m_b)$ to $\mathscr{A}$.
5. The adversary $\mathscr{A}$ can additionally issue key queries for vectors $\mathbf{y}$ and get $sk_{\mathbf{y}} \leftarrow \text{KeyGen}(sk, \mathbf{y})$ as response provided that $< \mathbf{y}, \mathbf{x}_0 > \neq 0, < \mathbf{y}, \mathbf{x}_1 > \neq 0$.
6. Finally, $\mathscr{A}$ outputs a bit $b' \in \{0, 1\}$, and succeeds if $b' = b$.

The advantage of $\mathscr{A}$ in this game is defined as

$$Adv_{\mathscr{A}}^{\mathrm{AH-CPA}}(1^k) = |Pr[b' = b] - \frac{1}{2}|$$

Definition 2. We say an IPE scheme is $(t, q_t, \varepsilon(k))$ weakly AH-CPA secure if for any t-time adversary $\mathscr{A}$ making at most q_t attribute vector key queries, we have $Adv_{\mathscr{A}}^{\mathrm{AH-CPA}}(1^k) < \varepsilon(k)$ in the above game.

2.3 PECDK Scheme

2.3.1 Syntax

A non-interactive public key encryption with conjunctive and disjunctive keyword search(PECDK) scheme consists of four probabilistic polynomial time algorithms as follows:
KeyGen(1^k,n): Takes the security parameter 1^k and n as input, outputs the users' public and private keys (pk, sk).
PECDK(pk, W): Encrypts a keyword set W with the public key pk to produce a searchable index C_W.
Trapdoor(sk, W'): Given a private key sk and a keyword predicate W' ,outputs a trapdoor $T_{W'}$.
Test$(pk, C_W, T_{W'})$: Outputs 1 iff W satisfies W' and 0 otherwise.
Correctness. For all possible W and W', the corresponding ciphertext $C_W = \mathrm{PECDK}(pk, W)$ and corresponding trapdoor $T_{W'} = \mathrm{Trapdoor}(sk, W')$, it holds that $1 = \mathrm{Test}(pk, C_W, T_{W'})$ if W satisfies W'. Otherwise, it holds with negligible probability.

2.3.2 Security Model

The security of PECDK can be defined by the following IND-CC-CTA game.
1. The challenger $\mathscr{C}$ takes a security parameter 1^k and runs KeyGen algorithm. The public key pk is given to the adversary $\mathscr{A}$. The secret key sk is kept by the challenger $\mathscr{C}$ itself.
2. $\mathscr{A}$ adaptively queries trapdoors for a polynomial number of search predicate W'. The challenger runs Trapdoor(sk, W') and responds the trapdoor $T_{W'}$ to $\mathscr{A}$.
3. $\mathscr{A}$ selects two keyword sets W_0 and W_1, and sends them to the challenger $\mathscr{C}$. The only restriction is that W_0 and W_1 cannot satisfy any of W' queried in previous phase. The challenger picks a random bit $b \in \{0, 1\}$. And it sets $C_b = \mathrm{PECDK}(pk, W_b)$ and sends it to $\mathscr{A}$.
4. $\mathscr{A}$ additionally queries search predicates W' to trapdoor oracle. The challenger runs Trapdoor(sk, W') and responds the trapdoor $T_{W'}$ to $\mathscr{A}$ if W' cannot be satisfied by W_0 and W_1.

5. $\mathscr{A}$ finally outputs a guess $b' \in \{0,1\}$. It wins the game it $b' = b$.
The advantage of $\mathscr{A}$ against IND-CC-CTA is

$$Adv_{\mathscr{A}}^{\text{IND-CC-CTA}}(1^k) = |Pr[b' = b] - \frac{1}{2}|$$

Definition 3. We say a PECDK scheme is $(t, q_t, \varepsilon(k))$ secure if for any t-time adversary $\mathscr{A}$ making at most q_t trapdoor queries in the above game, we have $Adv_{\text{PECDK},\mathscr{A}}^{\text{IND-CC-CTA}}(1^k) < \varepsilon(k)$.

3 Our Scheme

Suppose the system has n different keyword fields $X_1, \ldots, X_n$, and the maximum number of disjunctions within each keyword field is 2 (it can be any integer, we just for simplicity here). To use good properties of the IPE scheme in [16], we encode our keywords to be elements of Z_q.

KeyGen(1^k,n) Given the security parameter 1^k and n, it runs algorithm $\mathbf{G}_{ob}(1^k, 4n+5)$ to get $(param_{\mathbf{V}}, \mathbf{B}, \mathbf{B}^*)$, where $param_{\mathbf{V}} =< q, \mathbf{V}, G_T, \mathbf{A}, e >$, $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_{4n+5})$, $\mathbf{B}^* = (\mathbf{b}_1^*, \ldots, \mathbf{b}_{4n+5}^*)$. Return public/private keys

$$sk = \mathbf{B}^*,\ pk = (param_{\mathbf{V}}, \hat{\mathbf{B}})$$

$$where\ \hat{\mathbf{B}} = (\mathbf{b}_1, \ldots, \mathbf{b}_{2n+1}, \mathbf{b}_{4n+3}, \mathbf{b}_{4n+5})$$

PECDK(pk, W) Let $W = (w_1, \ldots, w_n) \in Z_q^n$ be the keyword set related to a message, the encryption algorithm first computes vector $\mathbf{v} = (w_1, w_1^2, \ldots, w_n, w_n^2, 1)$, Then chooses random elements $\delta_1, \delta_2, \zeta \in Z_q$, outputs $C_W = (\mathbf{c}_1, c_2)$.

$$\mathbf{c}_1 = \delta_1 (\sum_{i=1}^{2n+1} v_i \mathbf{b}_i) + \zeta \mathbf{b}_{4n+3} + \delta_2 \mathbf{b}_{4n+5},\ c_2 = g_T^{\zeta}$$

Trapdoor(sk, W') Suppose the user wants to retrieve emails with restrictions $\bigwedge_{i=1}^{n} (X_i = w'_{i1} \vee X_i = w'_{i2})$, where $w'_{ij} \in Z_q$ or $w'_{ij} = \perp$ ($\perp$ denotes none). As for $W' = (w'_{11}, w'_{12}, \ldots, w'_{n1}, w'_{n2})$, first chooses random elements $r_2, \ldots, r_n \in Z_q$ and computes polynomial $f(x_1, \ldots, x_n) = \sum_{i=1}^{n} (x_i^2 + a_i x_i) + a_{n+1} = (x_1 - w'_{11})(x_1 - w'_{12}) + r_2(x_2 - w'_{21})(x_2 - w'_{22}) + \ldots + r_n(x_n - w'_{n1})(x_n - w'_{n2})$. If there exists any $w'_{ij} = \perp$, just omit the corresponding term. Let $\mathbf{u} = (1, a_1, r_2, a_2, \ldots, r_n, a_n, a_{n+1})$ and then chooses random elements $\sigma, \eta \in Z_q$, outputs $\mathbf{T}_{W'}$.

$$\mathbf{T}_{W'} = \sigma(\sum_{i=1}^{2n+1} u_i \mathbf{b}_i^*) + \mathbf{b}_{4n+3}^* + \eta \mathbf{b}_{4n+4}^*$$

Test$(pk, C_W, \mathbf{T}_{W'})$ When the server receive a trapdoor $T_{W'}$ from the user, it tests whether ciphertext C_W corresponds to $T_{W'}$. The algorithm outputs 1 iff

$$c_2 / e(\mathbf{c}_1, \mathbf{T}_{W'}) = 1$$

Consistency. When expressed by bases of the DPVS, $\mathbf{c}_1 = (\delta_1 \mathbf{v}, 0^{2n+1}, \zeta, 0, \delta_2)_{\mathbf{B}}$ and $\mathbf{T}_{W'} = (\sigma \mathbf{u}, 0^{2n+1}, 1, \eta, 0)_{\mathbf{B}^*}$. Hence, we can get $c_2 / e(\mathbf{c}_1, \mathbf{T}_{W'}) = g_T^{\delta_1 \sigma < \mathbf{u}, \mathbf{v} >}$.

If W satisfies W' , we can get

$$\begin{aligned}
&(w_1 - w'_{11})(w_1 - w'_{12}) + \sum_{i=2}^{n} r_i (w_i - w'_{i1})(w_i - w'_{i2}) = 0 \\
\Rightarrow &\sum_{i=1}^{n} (w_i^2 + a_i w_i) + a_{n+1} = 0 \\
\Rightarrow &< \mathbf{u}, \mathbf{v} >= 0 \\
\Rightarrow &\textbf{Test}(pk, C_W, \mathbf{T}_{W'}) = 1.
\end{aligned}$$

If W does not satisfy W', we can get

$$\begin{aligned}
&Pr[f(w_1, \ldots, w_n) = 0] = 1/q \\
&\Rightarrow Pr[\textbf{Test}(pk, C_W, \mathbf{T}_{W'}) = 1] = 1/q
\end{aligned}$$

To sum up, our scheme is computational consistent.

3.1 Security

Theorem 1. Since the corresponding IPE scheme in [16] is $(t, q_t, \varepsilon'(k))$ weakly AH-CPA secure, the PECDK scheme we proposed is $(t, q_t, \varepsilon'(k) + \frac{q_t}{q^2} + \frac{2q_t}{q})$ IND-CC-CTA secure.

Proof. Let $\Pi' = (\text{KeyGen}', \text{PECDK}', \text{Trapdoor}', \text{Test}')$ denotes the PECDK scheme above, $\Pi = (\text{Setup}, \text{KeyGen}, \text{Enc}, \text{Dec})$ denotes the corresponding IPE scheme. Let $\mathscr{A}$ be a probabilistic polynomial-time adversary, and define

$$\varepsilon(k) = Adv_{\mathscr{A}}^{IND-CC-CTA}(1^k)$$

In an execution of experiment IND-CC-CTA, let $\mathbf{u}_0$ and $\mathbf{u}_1$ be the intermediate vectors generated while encrypting W_0 and W_1 respectively. Let $\mathbf{v}_i$ be the intermediate vectors generated in the trapdoor algorithm for W_i'. At any point during its execution, $\mathscr{A}$ queries W_i' to the oracle Trapdoor, where both W_0 and W_1 cannot satisfy W_i'. Let Query1, Query2 and Query 3 be three events disjoint from each other but make up all the possibilities in the execution

Query1: there exists an $i \in [1, q_t]$ such that $< \mathbf{u}_0, \mathbf{v}_i >=< \mathbf{u}_1, \mathbf{v}_i >= 0$.
Query2: there exists an $i \in [1, q_t]$ such that $< \mathbf{u}_0, \mathbf{v}_i >= 0, < \mathbf{u}_1, \mathbf{v}_i > \neq 0$ or $< \mathbf{u}_0, \mathbf{v}_i > \neq 0, < \mathbf{u}_1, \mathbf{v}_i >= 0$
Query3: the event otherwise.

Then, we can denote the probability above as follows.

$$\begin{aligned}\varepsilon(k) &= \sum_{i=1}^{3} Pr[b' = b \wedge Query i] - \tfrac{1}{2} \\ &\leq \sum_{i=1}^{2} Pr[Query i] + Pr[b' = b \wedge Query 3] - \tfrac{1}{2}\end{aligned}$$

where all probabilities are taken over the randomness used in experiment IND-CC-CTA. We can show that the advantage of $\mathscr{A}$ is negligible through three claims below.

Claim 1. If the adversary is allowed to query q_t times, then $Pr[Query1] \leq q_t/q^2$.
Claim 2. If the adversary is allowed to query q_t times, then $Pr[Query2] \leq 2q_t/q$.
Claim 3. If the corresponding IPE scheme Π' is $(t, q_t, \varepsilon'(k))$ AH-CPA secure, then $Pr[b' = b \wedge\ Query3] - \frac{1}{2} \leq \varepsilon'(k)$.

3.2 Efficiency

We compare our scheme with the work of [12] and [19] in Table 1. As we can see , the size of a ciphertext and trapdoor is linear with the number of keyword fields n in all three algorithms. Though our scheme do not support negation of keyword when compared with [19], it can achieve disjunctions within each keyword field that the other two cannot do. Furthermore, PECDK is based on prime-order bilinear groups rather than composite-order ones. [20] points out that to achieve AES-128 security level, the minimum bit length of composite-order elliptic curves is 2644 bits, which is far more bigger than 256-bit in prime-order setting.

Table 1: Comparisons of existing multi-keyword search schemes

Schemes	HL07[12]	LH14[19]	PECDK
Ciphertext size	O(n)	O(n)	O(n)
Trapdoor seze	O(n)	O(n)	O(n)
Expressiveness	AND	AND,OR,NOT	AND,OR
Security	full	full	full
Security model	random oracle	standard model	standard model
Bilinear groups	prime-order	composite-order	prime-order
Disjunctions within keyword field	no	no	yes

4 Conclusion and Future Work

Considering the large scale of users and documents in cloud, it is crucial for servers to support multi-keyword search with both efficient computation and strong security. However, the disjunctive search ability of existing works mainly concentrated on different keyword fields. In this paper, we propose a PECDK scheme based on the excellent properties of existing IPE scheme to achieve both conjunction and disjunction within each keyword field. It is constructed in prime-order bilinear groups and has full security in the standard model. What still needs to be improved, however, is that our scheme has a double expansion of the parameter compared to the work of IPE by Lewko et al.. There still exists much for us to do in the future to further improve the efficiency and expressiveness of our scheme.

Acknowledgements The authors would like to thank the anonymous referees for their helpful comments. This work is supported by the National Natural Science Foundation of China (Nos. 61309016, 61379150,61602512).

References

1. Kamara S. and Lauter K.: Cryptographic cloud storage. In: Proc. of Financial Cryptography and Data Security, pp.136-149(2010)
2. Feng D., Zhang M., Zhang Y. and Xu Z.: Study on cloud computing security. In: Ruan Jian Xue Bao/Journal of Software, vol.22, no.1, pp.71-83(2011)
3. Yuriyama M. and Kushida T.: Integrated cloud computing environment with IT resources and sensor devices. In: International Journal of Space-Based and Situated Computing, vol.1, no.2/3, pp.163-173(2011)
4. Ronaldc P., Stephan S. and Christoph S.: A privacy-friendly architecture for future cloud computing. In: International Journal of Grid and Utility Computing, vol.4 no.4, pp.265-277(2013)
5. Ma K. and Zhang L.: Bookmarklet-triggered unified literature sharing services in the cloud. In: International Journal of Grid and Utility Computing, vol.5, no.4, pp.217-226(2014)
6. Yang W., Zhang C. and Mu B.: Towards mashup optimisation with global constraints in the cloud. In: International Journal of Grid and Utility Computing, vol.5, no.4, pp.227-235(2014)
7. Song D., Wagner D. and Perrig A.: Practical techniques for searches on encrypted data. In: Proc. of SP'00, pp.44-55(2000)
8. Boneh D., Crescenzo G., Ostrovsky R., et al.: Public key encryption with keyword search. In: Proc. of Advances in Cryptology-EUROCRYPT, pp.506-522(2004)
9. Park J., Kim K., and Lee P.: Public key encryption with conjunctive field keyword search. In: LNCS 3325, pp.73-86(2004)
10. Khader D.: Public key encryption with keyword search based on K-resilient IBE. In: LNCS 4707, pp.298-308(2006)
11. Boneh D. and Waters B.: Conjunctive, subset, and range queries on encrypted data. In: Proc. of Theory of cryptography, pp.535-554(2007)
12. Yong H. and Lee P.: Public key encryption with conjunctive keyword search and its extension to a multi-user system. In: Proc. of the First international conference on Pairing-Based Cryptography, pp.2-22(2007)
13. Baek J., Safavinaini R. and Susilo W.: Public key encryption with keyword search revisited. In: LNCS 2005, pp.1249-1259(2008)

14. Katz J., Sahai A., and Waters B.: Predicate encryption supporting disjunctions, polynomial equations, and inner products. In: Proc. of Advances in Cryptology-EUROCRYPT, pp.146-162(2008)
15. Rhee H., Park J., Susilo W., et al.: Trapdoor security in a searchable public-key encryption scheme with a designated tester, Journal of Systems and Software, vol.83, no.5, pp.763-771(2010)
16. Lewko A., Okamoto T., Sahai A., et al.: Fully secure functional encryption: attribute-based encryption and (hierarchical) inner product encryption. In: Proc. of Advances in Cryptology-EUROCRYPT, pp.62-91(2010)
17. Yang H., Xu C., and Zhao H.: An efficient public key encryption with keyword scheme not using pairing. In: Proc. of First International Conference on Instrumentation, pp.900-904(2011)
18. Zhao Y., Chen X., Ma H., et al.: A new trapdoor-indistinguishable public key encryption with keyword search. In: Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, vol.3, no.1/2, pp.72-81(2012)
19. Lv Z., Hong C., Zhang M., et al.: Expressive and secure searchable encryption in the public key setting. In: Proc. of Information Security, pp.364-376(2014)
20. Guillevic A.: Comparing the pairing efficiency over composite-order and prime-order elliptic curves. In: Proc. of Applied Cryptography and Network Security, pp.357-372(2013)

Design of a Reconfigurable Parallel Nonlinear Boolean Function Targeted at Stream Cipher

Su Yang
Engineering University of CAPF, Xi'an 710086, China
wj_suyang@126.com

Abstract. Nonlinear Boolean function plays a pivotal role in the stream cipher algorithms and trusted cloud computing platform, Based on the analysis of multiple algorithms, this paper proposes a hardware structure of reconfigurable nonlinear Boolean function. This structure can realize the number of variables and AND terms less than 80 arbitrary nonlinear Boolean function in stream cipher algorithms. The entire architecture is verified on the FPGA platform and synthesized under the 0.18μm CMOS technology, the clock frequency reaches 248.7MHz, the result proves that the design is propitious to carry out the most nonlinear Boolean functions in stream ciphers which have been published, compared with other designs, the structure can achieve relatively high flexibility, and it has an obvious advantage in the area of circuits and processing speed.

1 Introduction

With the research on cloud computing security based on trusted computing currently, the TPM (Trusted Platform Module) becomes more and more important which directly determines the performance and security of entire cloud computing platform [1-3]. Considering the nonlinear Boolean function is one of the most time consuming units in TPM, we must find an effective way to solve it. Through the analysis of a variety of stream cipher algorithms, we find that the feedback function, the filtering function and the clock control function in stream cipher algorithms all can be achieved by using nonlinear Boolean function [4-5]. Therefore, it has great practical significance to study the reconstruction of nonlinear Boolean function for the stream cipher algorithms.

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_13

At present, the research on reconfigurable hardware structure of nonlinear Boolean function is mainly divided into two parts: on the one hand, the hardware architecture is designed for each Boolean function, through selecting the configuration information it can implement one or several specific cryptographic algorithms [6], but these methods are not extended very well for lack of the reconfigurable feature of Boolean function; On the other hand, the hardware is realized by using FPGA (Field Programmable Gate Array) based on LUT (Look up Table) or CPLD (Complex Programmable Logic Device) based on AND-OR array [7], however, these methods do not take into account the characteristics of the stream cipher algorithms, and the efficiency is not high.

Based on the analysis of the structure characteristics of nonlinear Boolean function in stream cipher algorithms, this paper designs the reconfigurable structure of nonlinear Boolean function which can adapt to a variety of known stream cipher algorithms and greatly improve the adaptability and efficiency of nonlinear Boolean function.

2 Analysis of Characteristics of Nonlinear Boolean Function in Stream Cipher

According to the analysis of structure of stream ciphers in NESSIE project and ECRYPT project, the nonlinear feedback function and filtering function in stream ciphers can be summarized almost as nonlinear Boolean function. In order to design the reconfigurable architecture better, this paper analyzes the characteristics of nonlinear Boolean functions in stream ciphers as shown in Table 1.

Table 1. Characteristics of nonlinear Boolean functions in stream ciphers

Stream ciphers	Type of function	Number of variables	Maximum times	Number of AND terms
A5-2	Filtering function	12	2	13
Grain-80	Filtering function	12	3	17
Grain-80	Feedback function	14	6	23
W7	Filtering function	28	3	12
Mikey	Feedback function	7	1	7

Grain-128	Filtering function	17	3	13
Grain-128	Feedback function	20	2	13
Trivium	Feedback function	5	2	4
pomaranch	Filtering function	6	3	9
Decim	Filtering function	14	2	92

In order to use the nonlinear Boolean function in stream ciphers, they must fit the corresponding cryptographic characteristics. According to the analysis above, we can summary some characteristics of nonlinear Boolean function as follows:

(1) When the number of variables is less, the number of AND terms is more. In order to increase the complexity of nonlinear Boolean function, when the number of variables is few, the expression is inevitable complex. This characteristic increases the difficulty of code breaking and improves the security of cryptographic algorithms.

(2) When the number of variable is more, the number of AND terms is less. In the operation of Boolean function, the calculation of high order AND terms is always the bottleneck, in order to improve the speed of algorithm, on the basis of ensuring the security, we can decrease the number of AND terms as far as possible.

(3) High order AND terms and low order AND terms have a relationship with the inclusion. When the number of input variables is already determined, high order AND terms and low order AND terms must have a inclusion relationship, in the process of calculation, we can utilize the inclusion relationship designing the hardware to improve the processing efficiency of algorithms.

3 Design of Reconfigurable Hardware of Nonlinear Boolean Function

According to the calculation characteristics of nonlinear Boolean function analyzed above, the reconfigurable nonlinear Boolean function of stream cipher algorithms can be designed and realized with three parts: a kind of improved ALM (Adaptive Logic Module) is used to realize the reconfigurable hardware of low order AND terms which account for a large proportion; Tree-like network structure is used to realize the reconfigurable design of high order AND terms and output XOR network; The linear part of nonlinear Boolean function can be accomplished parallel with nonlinear part. Among them the improved ALM circuit is designed on the basis of characteristics (1) of Boolean function, the

Tree-like network is designed on the basis of characteristics (2) and (3) of Boolean function. The reconfigurable hardware structure of nonlinear Boolean function is shown as Fig. 1.

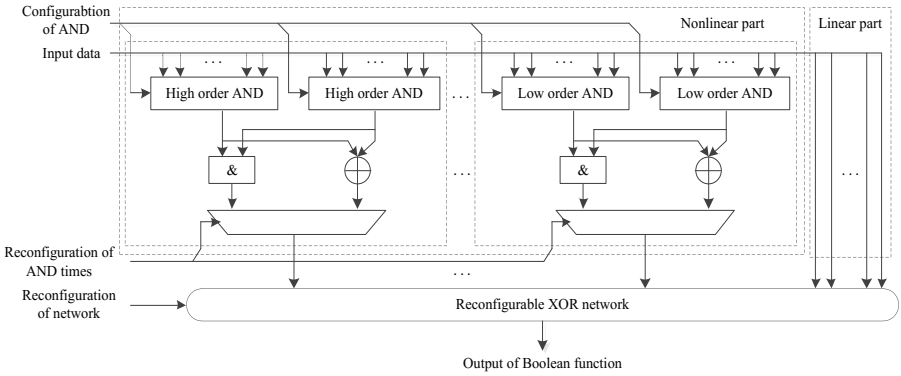


Fig. 1. Reconfigurable hardware structure of nonlinear Boolean function

3.1 Reconfigurable Design of Low Order AND Terms

Through the statistical research on the public cryptographic algorithms, we find that the times of AND terms are not more than 10 times in many stream cipher algorithms, so how to design and realize the low order AND terms has realistic significance.

For any expression of nonlinear Boolean function, the transformation from arbitrary form of nonlinear Boolean function to standard algebraic form is a complicated process by using programmable AND-OR array, for example, as to n input random nonlinear Boolean function, if we transform it to standard algebraic form, it needs to calculate 2^n modulus, with the increase of n, the calculation will be very complex, and the storage resources occupied by modulus will grow exponentially. So we consider using LUT to realize the low order AND terms. For LUT can realize any N input logic function, the time delay is small and each input is logically equivalent, so it is advantageous to realize the mapping algorithms and we just need to consider the requirements of input and output terminals. However, the LUT is actually a memory, for N input LUT requires 2^N storage units. With the increase of the input, the scale of LUT increases exponentially and area becomes larger. Therefore, in the actual design, we need to value the number of LUT input and take a more reasonable value.

Combined with the structure characteristics of nonlinear Boolean function in stream cipher algorithms and idea of programmable logic module in the circuit of FPGA, this paper proposes an improved ALM structure with 5 input variables LUT to realize low order AND terms. The structure of improved ALM is shown as Fig. 2.

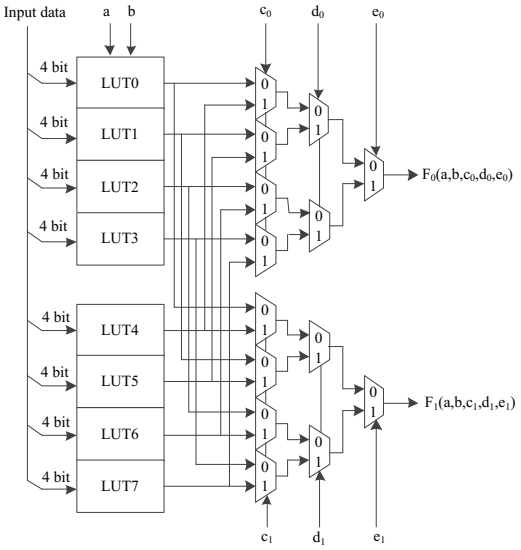


Fig. 2. The structure of improved ALM

The improved ALM circuit designed in the paper can realize reconfigurable nonlinear Boolean function with strong adaptation ability by changing the configuration information. The reconstruct ability is as shown in Table 2.

Table 2. Reconstruct ability of improved ALM circuit

Type of function	ALM_Config	Output of function
4 variables	$c_0 = 0$	ALM_Dataout0=$F_{40}(a,b,d_0,e_0)$
	$c_1 = 1$	ALM_Dataout1=$F_{41}(a,b,d_1,e_1)$
5 variables	$c_0=c_0$	ALM_Dataout0=$F_{50}(a,b,c_0,d_0,e_0)$
	$c_1=c_1$	ALM_Dataout1=$F_{51}(a,b,c_1,d_1,e_1)$

This structure has these reconfigurable characteristics:

(1) It can realize a Boolean function of any one of the five input variables, for example ALM_Dataout0=$F_{50}(a,b,c_0,d_0,e_0)$ or ALM_Dataout1=$F_{51}(a,b,c_1,d_1,e_1)$. The storage resources are monopolized by Boolean function.

(2) It can simultaneously achieve two Boolean functions of five input variables, but the function needs to have two identical variables, and the other three variables have the same expression, such as ALM_Dataout0=$F_{50}(a,b,c_0,d_0,e_0)$ and ALM_Dataout1=$F_{51}(a,b,c_1,d_1,e_1)$. These two Boolean functions reuse the storage unit.

(3) It can realize two Boolean functions of four input variables, through choosing the corresponding terminal, the expression has some flexibility, such as ALM_Dataout0=$F_{40}(a,b,d_0,e_0)$ and ALM_Dataout1=$F_{41}(a,b,d_1,e_1)$. Each Boolean function monopolizes four LUT units.

(4) According to the requirements of algorithms, we can reconstruct reconfigurable circuit with better adaptation ability by increasing the number of LUT units and the steps of MUX.

For two Boolean functions of five variables with the same structure, the realization of FPGA needs two 32 bit LUT units and 64 MUX units, while our structure just needs one 32 bit LUT units and 38 MUX units, the area savings rate reaches 50%, and the time delay has not changed. So our design has a good applicability for the nonlinear Boolean function with few variables and high repetition rate.

3.2 Reconfigurable Design of High Order AND Terms

Statistical analysis shows that realization of the high order AND terms are the critical path and bottleneck problem in the nonlinear Boolean function. Through the choice of configuration information, our design is to calculate the relationship between the AND terms in advance, then we adopt tree like structure to generate the high order AND terms based on the configuration information.

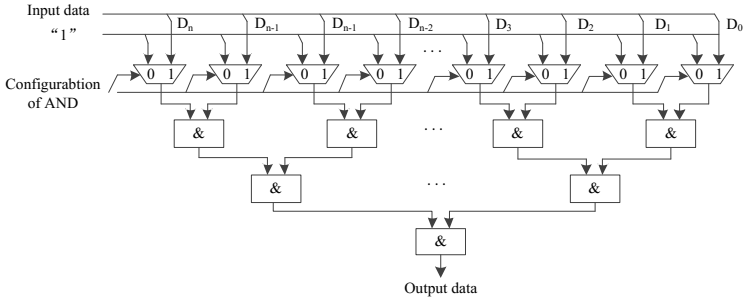


Fig. 3. The structure of reconfigurable high order AND terms

The structure of reconfigurable high order AND terms is shown as Fig. 3. By setting the data selector logic, the structure can accomplish any AND logic with arbitrary variables, when the input data which may come from the state value of shift register is not the effective variable in the AND logic, the data selector will select constant "1" entering to the next level circuits under the control of configuration information. Due to the constant "1" does not change the output of AND logic, so it will not affect the transmission of effective variables down to the next level circuits, then we can achieve any AND logic with arbitrary variables in the shift register and complete the refactoring operation of AND logic in the overall XOR logic. Through the control of configuration information, the structure can reuse the logical resources and time delay, and finally achieve the goal of improving the utilization ratio of resources and computing efficiency.

3.3 Reconfigurable Design of Output Network

To obtain the output of the final function operation, the reconfigurable output network of nonlinear Boolean function is to XOR each AND terms, for different algorithms, the number of the XOR terms is different, so through reconfigurable design, we can improve the computing speed of nonlinear Boolean function based on realization of the reconfigurable output network. It is assumed that the nonlinear Boolean function has p XOR terms, in the traditional implementations they set p as controller node and use the p-1 XOR gate cascade output, the overall time delay of the output network is a level of AND gate and p-1 levels of XOR gate, the logic resources of the design are p AND gates and p-1 XOR gates. With

the increase of the number of AND terms, the time delay will increase very obviously.

Based on the analysis of the characteristics of the above implementations, this paper proposes an optimized implementation method based on tree structure. As shown in Fig. 4, it is assumed that the nonlinear Boolean function has p XOR terms, the first level of tree structure has $p/2$ XOR terms, the second level has $p/4$ XOR terms, the n-th level has $p/2^n$ XOR terms, then the logic resources finally are p AND gates and p-1 XOR gates, the output delay of the circuit is a level of AND gate and $log_2 p$ levels of XOR gate.

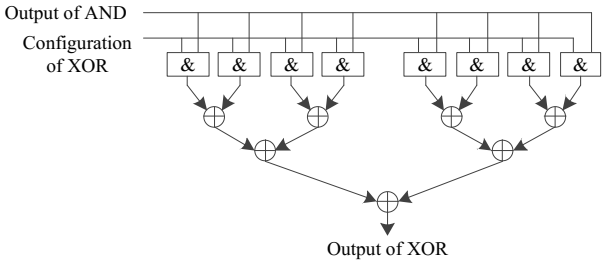


Fig. 4. The structure of reconfigurable output network

Compared with the computing result of traditional implementation way, the reconfigurable tree output network proposed in this paper can reduce the time delay from p-1 levels of XOR gate to $log_2 p$ levels of XOR gate under the constant of the logic resources and configuration information, and the optimization effect will be more obvious when the number of terms is more.

4 Performance and Analysis

4.1 Performance of This Design

Based on the analysis above, the prototype has been accomplished with RTL description using Verilog language and synthesized by Quartus II 10.0 form Altera Corporation, the prototype has been verified successfully, the result shows that our design can realize the nonlinear Boolean function of random variables and times in the 80 levels of cipher algorithms, Table 3 gives the clock frequency and resource occupancy when the number of variables are 40, 60 and 80.

Furthermore, our design has been synthesized under 0.18μm CMOS process using Synopsys Design Compiler to evaluate performance more accurately, the performance result shows in Table 4.

Table 3. The performance of reconfigurable nonlinear Boolean function based on FPGA

Device	Number of variables	Maximum clock frequency	ALUT
EP2S180F1020I4	40	233 MHz	172
	60	158 MHz	326
	80	125 MHz	498

Table 4. The performance of reconfigurable nonlinear Boolean function based on ASIC

Number of variables	Constraint	Area		Delay	Slack
		Combinational	Non combinational		
40	5 ns	228734	6896	3.22 ns	+0.87
60	5 ns	447468	10032	3.89 ns	+0.66
80	5 ns	603218	14783	4.02 ns	+0.36

4.2 Contrasts with Other Designs

Based on the synthesis result above, we make a comparison with the structure of reconfigurable nonlinear Boolean function with the structure of CPLD and FPGA which can realize the nonlinear Boolean function too, as to there are two critical parameters including area and latency in the synthesis result, so we list the area and latency of these three structures as shown in Fig. 5 and Fig. 6.

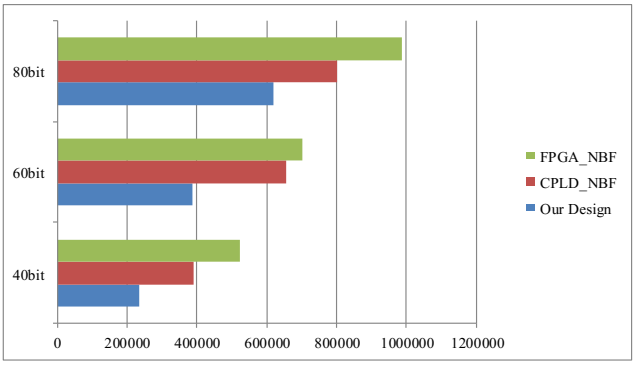


Fig. 5. The area comparison with other designs

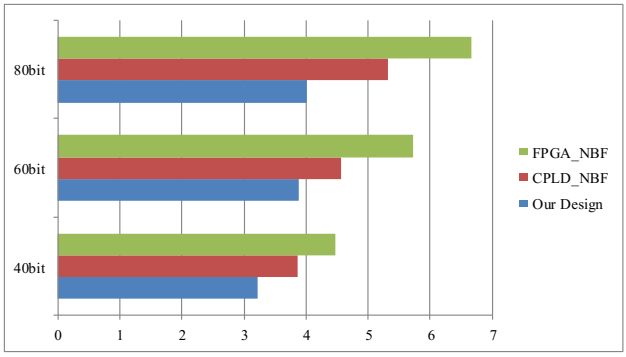


Fig. 6. The latency comparison with other designs

The comparison result shows that when the number of variables is 40, the area resources occupied of reconfigurable nonlinear Boolean function are 230 thousand gates, and the latency is 3.22 ns, which has been improved greatly compared with other designs. Meanwhile, with the increase of the number of variables, the advantages of our design are more obvious.

5 Conclusion

This paper presents a realization of high speed reconfigurable nonlinear Boolean function, which can satisfy random level, arbitrary variables and any forms of nonlinear function of stream cipher algorithms. In view of the low order AND terms, the optimization scheme is proposed based on the implementation of

LUT structure, which makes it more suitable for the structural characteristics of the nonlinear function; In the light of high order AND terms, an optimization scheme based on tree network is proposed; The final output network uses the tree like structure to improve the computing speed. Synthesis, placement and routing of reconfigurable design have accomplished on 018mm CMOS process. Compared with other designs, the result proves our design has an obvious advantage at the area and latency.

Acknowledgments. This work was supported in part by open project foundation of State Key Laboratory of Cryptology; National Natural Science Foundation of China (NSFC) under Grant No. 61202492, No. 61309022 and No. 61309008;

References

1. Barenghi A, Pelosi G, Terraneo F. Secure and efficient design of software block cipher implementations on microcontrollers [J]. International Journal of Grid & Utility Computing, 2013, 4(2/3):110-118.
2. Chengyu Hu, Bo Yang, Pengtao Liu:Multi-keyword ranked searchable public-key encryption. IJGUC 2015, 6(3/4): 221-231.
3. Tian H. A new strong multiple designated verifiers signature [J]. International Journal of Grid & Utility Computing, 2012(3):1-11.
4. Yuriyama M, Kushida T. Integrated cloud computing environment with IT resources and sensor devices[J]. International Journal of Space-Based and Situated Computing, 2011, 5(7): 11-14.
5. Iguchi N. Development of a self-study and testing function for NetPowerLab, an IP networking practice system [J]. International Journal of Space-Based and Situated Computing, 2014, 8(1): 22-25.
6. Xueyin Zhang, Zibin Dai, Wei Li, etc. Research on reconfigurable nonlinear Boolean funcitons hardware structure targeted at stream cipher [C]. 2009 2nd International Conference on Power Electronics and Intelligent Transportation System. 2009: 55-58.
7. Ji Xiangjun, Chen Xun, Dai Zibin etc. Design and Realization of an Implementation hardware with Non-Linear Boolean Function [J]. Computer Application and Software, 2014, 31(7): 283-285.

Temporally Adaptive Co-operation Schemes

Jakub Nalepa and Miroslaw Blocho

Abstract Selecting an appropriate co-operation scheme in parallel evolutionary algorithms is an important task and it should be undertaken with care. In this paper, we introduce the temporally adaptive schemes, and apply them in our parallel memetic algorithm for solving the vehicle routing problem with time windows. The experimental results revealed that this approach allows for retrieving better solutions in much shorter time compared with other co-operation schemes. The analysis is backed up with the statistical tests, which gave the clear evidence that the results are important. We report one new world's best solution to the benchmark problem obtained using our adaptive co-operation scheme.

Key words: Parallel algorithm; co-operation; memetic algorithm; VRPTW

1 Introduction

Solving rich vehicle routing problems (VRPs) is a vital research topic due to their practical applications which include delivery of food, beverages and parcels, bus routing, delivery of cash to ATM terminals, waste collection, and many others. There exist a plethora of variants of rich VRPs reflecting a wide range of real-life scheduling scenarios [6, 19]—they usually combine multiple realistic constraints which are imposed on feasible solutions. Although exact algorithms retrieve the optimum routing schedules, they are

Jakub Nalepa
Institute of Informatics, Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland e-mail: jakub.nalepa@polsl.pl

Miroslaw Blocho
Institute of Informatics, Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland e-mail: blochom@gmail.com

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_14

still very difficult to exploit in practice, because of their unacceptable execution times for massively-large problems. Therefore, approximate algorithms became the main stream of research and development—these approaches aim at delivering high-quality (however not necessarily optimum) schedules in significantly shorter time. In our recent work [14], we showed that our parallel memetic algorithm (PMA–VRPTW)—a hybrid of a genetic algorithm and some local refinement procedures—elaborates very high-quality schedules for the vehicle routing problem with time windows (VRPTW). Although PMA–VRPTW was very efficient, selecting the appropriate co-operation scheme (defining the co-operation topology, frequency and strategies to handle emigrants/immigrants) is extremely challenging and time-consuming—the improper selection can easily jeopardize the PMA–VRPTW capabilities.

1.1 Contribution

We propose two temporally adaptive co-operation schemes in PMA–VRPTW. In these schemes, the master process samples several time points during the execution, and monitors the search progress. Based on this analysis, the scheme is dynamically updated to balance the exploration and exploitation of the solution space, and to guide the search process as best as possible.

Our experiments performed on the well-known Gehring and Homberger's benchmark (in this work, we consider all 400-customer tests with wide time windows, large truck capacities, and random positions of the customers, which appeared very challenging [14]), revealed that the new temporally adaptive co-operation schemes allow for retrieving better solutions quickly (the differences are statistically important), compared with other means of co-operations. We report one new world's best solution elaborated using the new scheme. It is worth mentioning that such temporally adaptive strategies of establishing the desired co-operation schemes have not been intensively studied in the literature so far, and they may become an immediate answer to the problems which require the parallel processes to co-operate efficiently to guide the search process towards high-quality solutions quickly.

1.2 Paper Structure

This paper is structured as follows. Section 2 describes the VRPTW. In Section 3, we review the state of the art on the VRPTW. PMA–VRPTW is briefly discussed in Section 4. In the same section, we present the temporally adaptive co-operation schemes, which are the main contribution of this work. Section 5 contains the analysis of the experimental results. Section 6 concludes the paper and serves as the outlook to the future work.

2 Problem Formulation

The VRPTW is an NP-hard optimization problem of delivering goods to C customers using K homogeneous trucks. The main objective is to minimize the fleet size, and the secondary one is to optimize the total travel distance.

The VRPTW is defined on a complete graph $G = (V, E)$ with vertices $V=\{v_0,v_1,\ldots,v_C\}$ (representing the travel points), and edges $E=\{(v_i, v_j): v_i, v_j \in V, i \neq j\}$ (travel connections). The node v_0 is the depot (there is only one depot, i.e., the start and the finish travel point of all trucks). Each v_i defines its non-negative demand q_i (there is no depot demand, thus $q_0 = 0$), service time s_i ($s_0 = 0$), and time window $[e_i, l_i]$ (the service must be *started* within this slot, however it may finish after the time window has been closed). Every edge (v_i, v_j) has a travel cost c_{ij} (given in the Euclidean metric). A feasible solution is a set of K routes such that: (i) each route starts and ends at the depot, (ii) the truck loads do not exceed Q, (iii) the service of each v_i begins between e_i and l_i, (iv) each truck returns to the depot before l_0, and (v) each customer is served in exactly one route. If any of the constraints is violated, then the solution becomes unacceptable.

Let (K_α, T_α) and (K_β, T_β) represent two feasible VRPTW solution, denoted as α and β, respectively. The solution β is of a higher quality than the solution α, if $(K_\beta < K_\alpha)$ or $(K_\beta = K_\alpha$ and $T_\beta < T_\alpha)$. Hence, the solution β encompasses a lower number of routes, or—if the numbers of trucks are equal for both α and β—the total distance traveled during the service is smaller.

An exemplary solution σ of the VRPTW instance containing 25 customers is visualized in Fig. 1. This solution consists of three routes (r_1, r_2, and r_3): $r_1 = \langle v_0, v_8, v_{10}, v_{21}, v_{12}, v_{22}, v_{23}, v_{24}, v_{25}, v_{17}, v_{14}, v_0\rangle$ (10 customers are visited), $r_2 = \langle v_0, v_{11}, v_{15}, v_{19}, v_{20}, v_{18}, v_{16}, v_9, v_{13}, v_7, v_0\rangle$ (9 customers), and $r_3 = \langle v_0, v_6, v_2, v_1, v_4, v_3, v_5, v_0\rangle$ (6 customers). It is easy to see that each customer v_i, $i \in \{1, \ldots, 25\}$, is served exactly once (i.e., in one route). Assuming that the vehicle loads do not exceed the capacity in any route, and the time window constraints are not violated, this routing schedule is feasible.

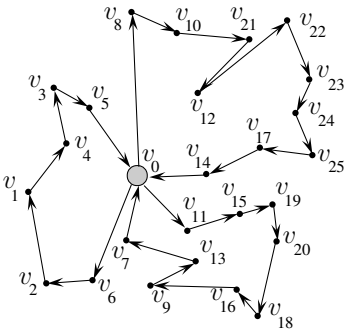


Fig. 1 An exemplary solution to the VRPTW instance with 25 clients served in 3 routes.

3 Related literature

Due to its wide practical applicability, the VRPTW attracted research attention. Exact algorithms aim at delivering the optimum solutions, however they are still difficult to apply in practice, because of their unacceptable execution times. These approaches encompass branch-and-cut, branch-and-bound, dynamic programming solutions, along with a plethora of various VRPTW formulations [1]. Exact algorithms were summarized and thoroughly discussed in numerous interesting surveys and reviews [2,8]. It is worth mentioning that in a majority of such approaches, minimizing the total distance is considered as the single objective.

The approximate methods include construction (creating solutions from scratch [20]) and improvement (which boost the quality of initial, usually very low-quality solutions [5, 11]) heuristics, and various meta-heuristics (very often allowing for the temporary deterioration of the solution quality during the optimization process) [5], including ant colony optimization techniques [7], particle swarm-based approaches [9], neighborhood searches [10], and many others [3]. In genetic algorithms (GAs), a population of solutions (chromosomes) undergoes the evolution in search of well-fitted individuals representing high-quality feasible solutions [21].

Memetic algorithms (MAs) combine EAs for exploring the entire search space, with intensive refinement procedures applied to exploit solutions already found [17] (they are often referred to as hybrid GAs). Such approaches have been successfully applied for solving a wide spectrum of optimization and pattern recognition problems [23]. A number of sequential and parallel MAs have been proposed for tackling the VRPTW [12,16,22], as well as other challenging rich VRPs [13, 18].

In our recent work [14], we showed that the co-operation scheme has a tremendous impact on the quality of final VRPTW solutions, and on the convergence time in our co-operative parallel MA. Its appropriate selection is not trivial and should respond to the search state. Also, we showed that dividing the search space across the co-operating processes (referred to as *islands*) helps significantly improve the exploration capabilities of the parallel algorithm [4, 15]. In this work, we tackle the problem of retrieving the appropriate co-operation schemes on the fly. This should allow for responding to the current search progress, and for choosing the best-fitted co-operation scheme (either explorative or exploitative). Such approaches have not been intensively explored in the literature so far.

4 Parallel Algorithm

In PMA–VRPTW (Algorithm 1)—which is a homogeneous island model parallel MA, since each island (a parallel process) runs the same MA to minimize

T—each individual p_i, where $i \in \{1, 2, \ldots, N\}$, corresponds to a VRPTW solution with K routes in a population of N solutions (on each island). The initial populations are generated using the parallel guided search [14] (it minimizes K at first, and then is used to create initial populations for each island). These populations evolve to optimize the distance T (lines 2–16).

Algorithm 1 Parallel memetic algorithm (PMA–VRPTW).

```
 1: Minimize K and find populations for each island;
 2: parfor P_i ← P_1 to P_n do
 3:     while not finished do
 4:         Determine N pairs (p_a, p_b);
 5:         for all (p_a, p_b) do
 6:             GenerateChild(p_a, p_b);                    ▷ Fig. 2
 7:         end for
 8:         Form the next population of size N;
 9:         if (can co-operate) then
10:             Determine and send emigrant(s);
11:             Receive and handle immigrant(s);
12:         end if
13:         Verify termination condition;
14:     end while
15: end parfor
16: return best solution among all islands;
```

The evolution involves selecting pairs of individuals for crossover, recombining them using the edge-assembly operator [12], and restoring the feasibility of children if it is necessary, using local edge-exchange moves (Fig. 2). Then, the children are *educated* (this is a memetic operator, thus it is rendered in light red), and mutated. Both operations involve applying edge-exchange and edge-relocate moves. The islands co-operate (Algorithm 1, lines 9–12), to propagate the best solutions found up to date, and to guide the search towards better routing schedules. The best individual (across all processes) is finally returned (line 16). For more details on PMA–VRPTW, see [14].

4.1 Temporally Adaptive Co-operation

In the temporally adaptive co-operation schemes (which are based upon our previous knowledge synchronization and ring schemes [14]), we monitor the dynamic changes of the total distance T of the best solution in the master island. During each co-operation phase (which occurs after finishing each generation), we calculate the differences:

$$\Delta T_i = \left| G_{(c-i)}(T) - G_c(T) \right|, \tag{1}$$

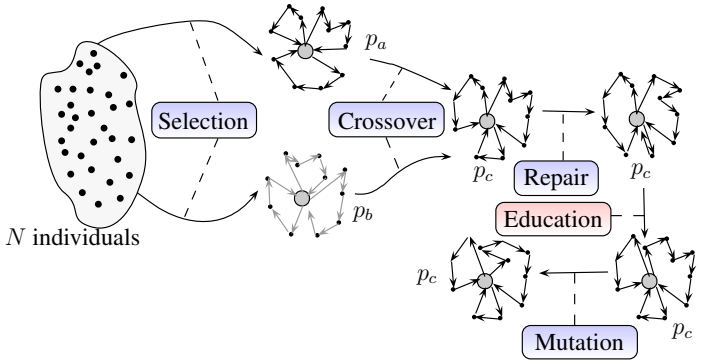


Fig. 2 Creation of a child in PMA–VRPTW.

where $G_{(c-i)}(T)$ denotes the best travel distance in the G_{c-i} generation (let G_c be the current generation). The ΔT_i values are found for three time points in the past—it is visualized in Fig. 3 (the differences are found for the second, fifth, and tenth generation before G_c, shown in blue).

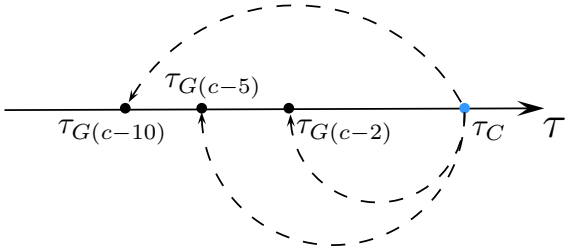


Fig. 3 Sampling several T values during the evolution.

Certain differences are compared with the expected improvements in the travel distances (ΔT_i^e). These comparisons are exploited to adapt the scheme—if the current co-operation is explorative, then it may be appropriate to switch it to the more exploitative one (and vice versa). In both phases (minimizing K and T), the more exploitative version of the scheme (either ring or KS) is used at first (we exploit only 10% of the closest customers to the one being affected in the edge-exchange moves).

In each co-operation phase, we calculate ΔT_2, ΔT_5, and ΔT_{10} (the last increment is found only in the exploitation mode, whereas the first—in the exploration), along with the expected improvements. For the exploitative co-operations, we have: ΔT_5^e and ΔT_{10}^e, where $\Delta T_5^e = \alpha_5 G_{(c-5)}(T)$, and $\Delta T_{10}^e = \alpha_{10} G_{(c-10)}(T)$, and the α coefficients are given in %, whereas for the explorative ones (i.e., ring or KS with the search space partitioning [15]) we additionally have the lower bounds of these measures ($\beta_2 \Delta T_2^e$ and $\beta_5 \Delta T_5^e$,

where β's are in %). These expected improvements are thus dependent on the travel distance in the $G_{(c-i)}$ generation, denoted as $G_{(c-i)}(T)$, and on the current co-operation mode (exploration or exploitation). Note that the α's may differ for both co-operation modes.

Algorithm 2 Temporal adaptation of the co-operation.

```
1: if (exploitation mode) then
2:     if (ΔT_2 = 0 or ΔT_5 ≤ ΔT_5^e or ΔT_10 ≤ ΔT_10^e) then
3:         Switch to explorative co-operation;
4:     end if
5: else
6:     if (ΔT_5 = 0 or
           ΔT_5 ≥ ΔT_5^e or ΔT_5 ≤ β_5 ΔT_5^e or
           ΔT_2 ≥ ΔT_2^e or ΔT_2 ≤ β_2 ΔT_2^e) then
7:         Switch to exploitative co-operation;
8:     end if
9: end if
```

Algorithm 2 presents the adaptation process. If the changes in the best T value are relatively small in the exploitation mode, then it is switched to the explorative one (line 3). On the other hand, if these changes are significant during the exploration, it indicates that this part of the solution space should be more intensively exploited, hence the co-operation toggles its mode (line 7). Also, if they are very small (less than the lower bounds), then the further exploration may not help find new high-quality solutions, and the mode becomes exploitative (this often happens when the high-quality solutions have already been retrieved).

5 Experimental Validation

5.1 Settings

PMA–VRPTW was implemented in the C++ programming language using the Message Passing Interface (MPI). The computations were carried out on the cluster equipped with Intel Xeon Quad Core 2.33 GHz processors, each with 12 MB level 3 cache. The nodes were connected by the Infiniband DDR fat-free network (throughput 20 Gbps, delay 5 μs). The source code was compiled using Intel 10.1 compiler and MPICH v. 1.2.6 MPI library.

We compared the proposed temporally adaptive schemes with our previous (best) ones [14] (in Table 1, we gather the co-operation schemes which have been investigated in this work). For the exploitation mode, we have: $\alpha_5 = 0.2\%$, and $\alpha_{10} = 0.5\%$, whereas for the exploration: $\alpha_2 = 1\%$, $\alpha_5 = 2\%$, and $\beta_2 = \beta_5 = 25\%$—the α and β parameter values were tuned experimen-

Table 1 Investigated co-operation schemes.

(a)	Ring
(b)	Ring with partitioned neighborhoods
(c)	Ring with partitioned routes
(d)	Ring with both partitioning strategies
(e)	Knowledge synchronization
(f)	Knowledge synchronization with partitioned neighborhood
(g)	Knowledge synchronization with partitioned routes
(h)	Knowledge synchronization with both partitioning strategies
(**i**)	**Adaptive knowledge synchronization**
(**j**)	**Adaptive ring**

tally, using test instances of various characteristics and structures. However, while selecting the appropriate α and β values, it is necessary to analyze the underpinning ideas of the current co-operation scheme (note that the α parameters affect the change from the exploitative to the explorative mode, whereas the β coefficients—from the explorative to the exploitative one). In the exploitative mode, the changes in T's are most often notably smaller compared with those retrieved in the explorative mode. This observation may become a good starting point in the tuning process of these parameters, however it requires further research attention. In all experiments, the number of processes was $n = 24$, the maximum evolution time was set to $\tau_E = 2000$ seconds, and the maximum time of minimizing K was $\tau_K = 60$ seconds (the first phase took approximately 10 seconds in all cases).

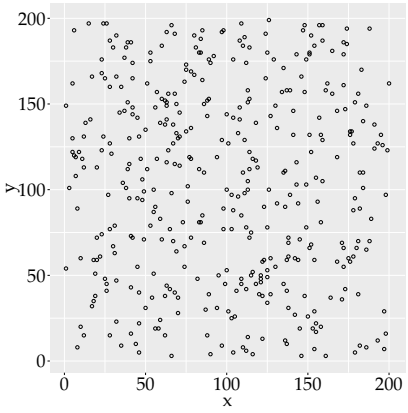


Fig. 4 An exemplary structure of a 400-customer Gehring and Homberger's test instance with the customers randomly scattered around the map.

In this work, we focus on 400-customer Gehring and Homberger's tests with random positions of travel points, wide time windows, and relatively large truck capacities (class r2). An exemplary structure of a test belonging to this class of benchmark instances is visualized in Fig. 4.

5.2 Analysis and Discussion

The results obtained using PMA–VRPTW with various co-operations are gathered in Table 2. We sampled and averaged the best T's (across all islands) in several time points (PMA–VRPTW was executed 10× using each scheme for each—out of 10—problem instance). Our new adaptive schemes significantly outperformed other ones. Importantly, PMA–VRPTW with the new schemes converged to very high-quality schedules quickly—the average T in $\tau = 30$ minutes is reduced by approx. 0.7% and 0.3% compared with $\tau = 5$ minutes for the adaptive KS and ring, respectively. Hence, this decrease is negligible. Therefore, the algorithm could have been terminated much earlier, since acceptable solutions had already been retrieved. It is worth noting that we have beaten the world's best solution (we decreased T from 7129.03 to 7128.93 for $K = 8$) for the r2_4_5 test using the adaptive KS[1].

Table 2 The average travel distances T (the best results out of 10 independent executions of PMA–VRPTW with each co-operation scheme applied are averaged for 10 instances in the r2 class). The best T's (in each sampled time point) are boldfaced.

Scheme	$\tau = 5$ min.	$\tau = 10$ min.	$\tau = 15$ min.	$\tau = 20$ min.	$\tau = 25$ min.	$\tau = 30$ min.
(a)	6265.73	6200.77	6190.64	6189.97	6189.91	6189.91
(b)	6205.87	6195.02	6191.54	6189.57	6189.25	6188.80
(c)	6284.24	6219.63	6199.34	6193.09	6189.83	6187.72
(d)	6199.87	6193.27	6191.77	6191.30	6163.37	6190.35
(e)	6353.40	6257.54	6218.06	6199.36	6191.10	6186.31
(f)	6195.94	6185.86	6183.25	6181.51	6180.35	6179.72
(g)	6329.54	6247.26	6212.84	6197.85	6192.12	6189.56
(h)	6195.58	6180.52	6178.54	6177.81	6177.56	6177.01
(i)	6171.40	6168.65	**6167.72**	**6167.37**	**6167.13**	**6166.89**
(j)	**6169.91**	**6167.77**	**6167.72**	6167.65	6167.62	6167.62

In Fig. 5, we render the average convergence time (i.e., after which the best solution across all co-operating islands could not be further improved, and may be considered as the *target solution*) of the T optimization phase. Applying the temporally adaptive schemes allowed for decreasing this time notably (also, the retrieved solutions were of a much higher quality—see Table 2). In the average case, PMA–VRPTW converges up to 3.7× faster when the adaptive ring scheme is applied, compared with our previous co-operations. It is quite important in practical applications, in which high-quality routing schedules should be retrieved as fast as possible. The results show that converging to target solutions is significantly faster when the adaptation is applied—see e.g., Fig. 5(j) (adaptive ring) compared with Fig. 5(a,d) (ring and ring with both partitioning strategies—almost 1.9× faster on average), with Fig. 5(b) (ring with partitioned neighborhoods—almost 2× faster), and with Fig. 5(c) (ring with partitioned routes—2.5× faster). Similarly, the adaptive KS is up

[1] The details can be found at: `http://sun.aei.polsl.pl/~jnalepa/3PGCIC16`.

to 2.5× faster than KS (on average). Finally, the best and the worst convergence times are also the lowest in the case of adaptive co-operation schemes (see the orange and gray bars in Fig. 5). Since the routing schedules retrieved using these schemes are of the highest-quality (as shown in Table 2), these schemes outperform the other ones when both the convergence time and the quality of final solutions are considered.

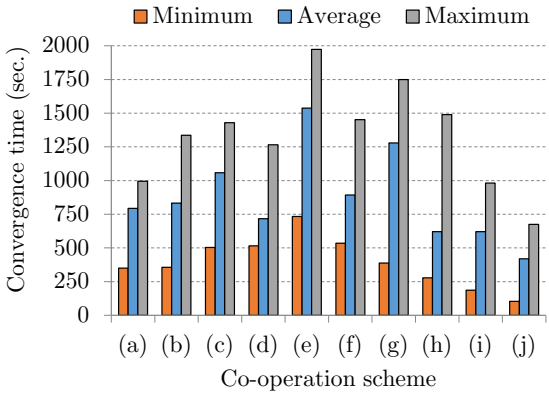


Fig. 5 Convergence time (in seconds) of PMA–VRPTW for various co-operation schemes.

Finally, we performed the two-tailed Wilcoxon tests to verify the null hypothesis saying that "applying different co-operation schemes leads to retrieving solutions of the same quality". The levels of the statistical significance are presented in Table 3—they prove that using our new temporally adaptive schemes allows for elaborating significantly different (better) routing schedules (the null hypothesis can be safely rejected because $p < 0.0001$ in most cases). Although the differences between the schedules obtained using two adaptive schemes (adaptive ring and adaptive KS) are not necessarily statistically important, the adaptive ring should be preferred since it converges faster compared with the adaptive KS (see Fig. 5).

6 Conclusions and Outlook

In this paper, we proposed two temporally adaptive co-operation schemes, and applied them in our parallel algorithm for solving the VRPTW. The adaptation procedure involves monitoring of the search process, and the dynamic selection of the appropriate co-operation strategy—this strategy may exhibit either more explorative or more exploitative behavior, depending on the current optimization state. The experimental study performed on the Gehring and Homberger's benchmark tests with randomized customers re-

Table 3 The level of statistical significance obtained using the two-tailed Wilcoxon tests. The differences which are statistically important (at $p < 0.05$) are boldfaced.

	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)
(a)	0.0949	**0.0061**	0.3173	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**
(b)	—	**0.0019**	0.1556	**0.0003**	**<0.0001**	**0.001**	**0.0001**	**<0.0001**	**<0.0001**
(c)		—	**0.0164**	**0.0001**	**<0.0001**	**0.0016**	**<0.0001**	**<0.0001**	**<0.0001**
(d)			—	**0.0009**	**<0.0001**	**0.0008**	**<0.0001**	**<0.0001**	**<0.0001**
(e)				—	**<0.0001**	**0.0375**	**<0.0001**	**<0.0001**	**<0.0001**
(f)					—	**<0.0001**	0.0767	**<0.0001**	**<0.0001**
(g)						—	**<0.0001**	**<0.0001**	**<0.0001**
(h)							—	**<0.0001**	**<0.0001**
(i)								—	0.332

vealed that utilizing the proposed schemes allows for retrieving solutions of a higher quality (the differences are statistically important) in much shorter time. We reported one new world's best solution elaborated using PMA–VRPTW with the new temporally adaptive scheme.

Our future work is focused on applying our new adaptive schemes for solving other challenging optimization problems (especially, the pickup and delivery with time windows). The presented ideas are quite generic and could be applied in parallel algorithms for other tasks too. Also, we plan to complement the suggested co-operation schemes with the adaptation of the co-operation frequency (similarly, based on the temporal analysis of the search progress). We work on the automatic selection of the most appropriate points to sample the T values in the adaptive schemes, as well as on the adaptation of their parameters. We plan to perform the full scalability tests using the large-scale parallel systems (e.g., computational clusters). Finally, it will be interesting to investigate how the co-operation schemes affect the diversity of the populations (of all islands) during the PMA–VRPTW execution.

7 Acknowledgments

This research was supported by the National Science Centre under research Grant No. DEC-2013/09/N/ST6/03461, and performed using the Intel CPU and Xeon Phi platforms provided by the MICLAB project No. POIG.02.03.00.24-093/13. We also thank the Gdańsk Computer Centre (TASK CI), where the computations were carried out.

References

1. R. Baldacci, A. Mingozzi, and R. Roberti. New route relaxation and pricing strategies for the vehicle routing problem. *Operations Research*, 59(5):1269–1283, 2011.

2. R. Baldacci, A. Mingozzi, and R. Roberti. Recent exact algorithms for solving the vehicle routing problem under capacity and time window constraints. *European J. of Op. Research*, 218(1):1 – 6, 2012.
3. R. Banos, J. Ortega, C. Gil, A. L. Márquez, and F. de Toro. A hybrid meta-heuristic for multi-objective vehicle routing problems with time windows. *Computers & Industrial Engineering*, 65(2):286 – 296, 2013.
4. M. Blocho and J. Nalepa. A parallel algorithm for minimizing the fleet size in the pickup and delivery problem with time windows. In *Proc. EuroMPI*, pages 15:1–15:2, New York, USA, 2015. ACM.
5. O. Bräysy and M. Gendreau. Vehicle routing problem with time windows, part II: Metaheuristics. *Transportation Science*, 39(1):119–139, 2005.
6. J. Caceres-Cruz, P. Arias, D. Guimarans, D. Riera, and A. A. Juan. Rich vehicle routing problem: Survey. *ACM Computing Surveys*, 47(2):32:1–32:28, 2014.
7. D. Coltorti and A. E. Rizzoli. Ant colony optimization for real-world vehicle routing problems. *SIGEVOlution*, 2(2):2–9, 2007.
8. N. A. El-Sherbeny. Vehicle routing with time windows: An overview of exact, heuristic and metaheuristic methods. *J. of King Saud University*, 22(3):123 – 131, 2010.
9. W. Hu, H. Liang, C. Peng, B. Du, and Q. Hu. A hybrid chaos-particle swarm optimization algorithm for the vehicle routing problem with time window. *Entropy*, 15(4):1247–1270, 2013.
10. B. Jarboui, A. Sifaleras, A. Rebai, M. Bruglieri, F. Pezzella, O. Pisacane, and S. Suraci. A variable neighborhood search branching for the electric vehicle routing problem with time windows. *Electronic Notes in Discrete Mathematics*, 47:221 – 228, 2015.
11. Y. Nagata and O. Bräysy. A powerful route minimization heuristic for the vehicle routing problem with time windows. *Operations Res. Letters*, 37(5):333 – 338, 2009.
12. Y. Nagata, O. Bräysy, and W. Dullaert. A penalty-based edge assembly memetic algorithm for the vehicle routing problem with time windows. *Computers & Operations Research*, 37(4):724 – 737, 2010.
13. Y. Nagata and S. Kobayashi. *A Memetic Algorithm for Pickup and Delivery with Time Windows Using Selective Exchange Crossover*, pages 536–545. Springer, 2010.
14. J. Nalepa and M. Blocho. Co-operation in the parallel memetic algorithm. *International Journal of Parallel Programming*, 43(5):812–839, 2015.
15. J. Nalepa and M. Blocho. A parallel algorithm with the search space partition for the pickup and delivery with time windows. In *Proc. 3PGCIC*, pages 92–99, 2015.
16. J. Nalepa and M. Blocho. Adaptive memetic algorithm for minimizing distance in the vehicle routing problem with time windows. *Soft Computing*, 20(6):2309–2327, 2016.
17. J. Nalepa and M. Kawulok. Adaptive memetic algorithm enhanced with data geometry analysis to select training data for SVMs. *Neurocomputing*, 185:113 – 132, 2016.
18. S. U. Ngueveu, C. Prins, and R. W. Calvo. An effective memetic algorithm for the cumulative capacitated vehicle routing problem. *Computers & Operations Research*, 37(11):1877 – 1885, 2010. Metaheuristics for Logistics and Vehicle Routing.
19. E. Osaba, X.-S. Yang, F. Diaz, E. Onieva, A. D. Masegosa, and A. Perallos. A discrete firefly algorithm to solve a rich vehicle routing problem modelling a newspaper distribution system with recycling policy. *Soft Computing*, pages 1–14, 2016.
20. K.-W. Pang. An adaptive parallel route construction heuristic for the vehicle routing problem with time windows. *Exp. Sys. with App.*, 38(9):11939 – 11946, 2011.
21. P. Repoussis, C. Tarantilis, and G. Ioannou. Arc-guided evolutionary algorithm for vehicle routing problem with time windows. *Evol. Comp., IEEE Trans. on*, 13(3):624–647, 2009.
22. T. Vidal, T. G. Crainic, M. Gendreau, and C. Prins. A hybrid genetic algorithm with adaptive diversity management for a large class of vehicle routing problems with time-windows. *Computers & Operations Research*, 40(1):475 – 489, 2013.
23. S. Wrona and M. Pawelczyk. Controllability-oriented placement of actuators for active noise-vibration control of rectangular plates using a memetic algorithm. *Archives of Acoustics*, 38(4):529–536, 2013.

Discovering Syndrome Regularities in Traditional Chinese Medicine Clinical by Topic Model

Jialin Ma[1,2*], Zhijian Wang[1]
[1] College of Computer and Information, Hohai University, Nanjing, China
majl@hyit.edu.cn, 51077061@qq.com
[2] Huaiyin Institute of Technology, Huaian, China

Abstract. Traditional Chinese Medicine (TCM) as one of most important approach for disease treatment in China for thousands of years. Lots of experience of famous experts in TCM is recorded in medical bibliography. The first vital work for TCM doctor is to diagnose the disease by the patients' symptoms, and then predict the syndromes which the patient has. Generally, this process reflects the medical skill of the TCM doctors. Therefore, TCM diagnose is easy to misdiagnose and difficult to master for TCM doctors. In this paper, we proposed a probabilistic model—the symptom-syndrome topic model (SSTM) to explore connected knowledge between symptoms and syndromes. In the SSTM, symptom-syndrome are modeled by generative process. Finally, we conduct the experiment on the SSTM. The results show that the SSTM is effective for mining the syndrome regularities in TCM data.

Keywords: TCM, syndrome, topic model, SSTM

1 Introduction

Traditional Chinese Medicine (TCM) as one of a important and independent medical theoretical and practical system has been existing thousands of years[1]. Furthermore, TCM is considered as one of an important complementary medical system to modern biomedicine. Especially in recent years, TCM is going abroad, more and more foreigners become accept and enjoy the treatment or health care of TCM[2]. Different from modern biomedicine, TCM doctors rarely resort to medical equipments to diagnose diseases. They usually utilize four diagnostic methods(observation, listening, interrogation, and pulse-taking) to understand the pathological conditions. Human body is regarded as a synthetic system in TCM. Therefore, TCM focuses on analysing the macro-level functional information of patients and utilize the much Chinese traditional naive philosophical theories and thoughts to adjust patients' body health and ecological balance. TCM emphasize individualized diagnosis and treatment which is most different from modern biomedicine[3].

In the past thousands years of the Chinese history, a large number of TCM doctor's experience, such as clinical case, ancient textbooks, classical ancient prescriptions etc., have been recorded. These records imply rich TCM knowledge. Mining regularities from TCM clinical records is one of the main approaches for TCM physicians to improve their clinical skills and empirical knowledge. Data mining is a useful computing approach for discovering hidden knowledge from large-scale data[4], and could be a potential solution for this issue. However, the huge scale of ancient TCM record is existing as the style of texts. Those experience and knowledge is described

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_15

by nature language. It is well known understanding semantics doesn't completely break through in the field of artificial intelligence. Fortunately, many researchers have paid attentions on the TCM record mining, such as[2, 4, 5].

The first and fundamental problem for TCM doctors is to diagnose patients' disease by four diagnostic methods(observation, listening, interrogation, and pulse-taking). Misdiagnosing would be lead to fatal results. Mining and learning diagnosing knowledge or experience from TCM record by computer is significative for doctors. It can guide or aid TCM doctors, especially young doctors to master diagnostic knowledge and experience[6]. In this paper, we proposed a probabilistic model—the symptom-syndrome topic model (SSTM) to explore latent knowledge between symptoms and syndromes. In the SSTM, symptom-syndrome is modelled by generative process which can help to acquire diagnosing knowledge or experience from TCM record. Finally, We conduct experiment on the SSTM. The results show that SSTM can extract effective latent semantics information and relation.

The paper is organized as follows: Section 2 reviews the related work about SMS spam filtering technologies. Section 3 presents our method in detail. Section 4 shows the experiments and discussion. Finally, we conclude and discuss further research in Section 5.

2 Related works

Topic model which is based on probability statistics theories, can detect latent semantic structure and information in large-scale documents[7]. Latent Semantic analysis (LSA) is one of the famous representative method in the early time[8]. It depends on capture word co-occurrence in documents. Therefore, LSA can bring semantic dimensionality between the text and words. Moreover, probabilistic latent semantic analysis (PLSA) is the further improvement of the LSA[9]. In PLSA, a document is regarded as a mixture of topics, while a topic is a probability distribution over words. In order to improve the defects of the PLSA, LDA proposed for the first time by Blei in 2003[10], which added Dirichlet priors in the distributions. LDA is a more completely generative model and achieves great successes in text mining and other artificial intelligence domains. With the rapid development of internet and social media, mass of short texts have been produced. Short texts data analysis (such as microblog) become advanced research hotspot. Many researchers are eager to mine social media date by topic model[11]. But the thorny problem is the lack of statistics information about terms in the short texts. Except directly applying the standard LDA, many improved topic model are researched to apt at short texts[12]. The famous researches in this aspect are Author-Topic Model(ATM)[13] and Twitter-LDA[12]. Moreover, Yan etc al. proposed a Biterm Topic Model(BTM), which can learn topics over short texts by directly modeling the generation of biterms in the whole corpus[14,22].

Many researchers have been eager to Knowledge discovering and data mining (KDD) in biomedicine field for a long times[2]. By contrast, TCM is to become a research hotspot in the recent years. The reviewing references of TCM mining are[4, 15-17]. Zhang et al. [1] proposed a data mining method, called the Symptom-Herb-Diagnosis topic (SHDT) model, to automatically extract the common relationships

among symptoms, herb combinations and diagnoses from large-scale CM clinical data. Jiang et al.[2] apply the Link Latent Dirichlet Allocation (LinkLDA), to automatically extract the latent topic structures which contain the information of both symptoms and their corresponding herbs. Yao et al.[18] proposed a framework which mines the treatment pattern in TCM clinical cases by using probabilistic topic model and TCM domain knowledge. The framework can reflect principle rules in TCM and improve function prediction of a new prescription. They evaluate our model on real world TCM clinical cases.

These mentioned studies have devoted to mining knowledge from TCM case data. They focus on knowledge the compatibility of TCM, diagnose law, or the rules of ‘‘Li-Fa-Fang-Yao”[19].

3 Our Work

Different from modern biomedicine, TCM doctors rarely resort to medical equipments to diagnose diseases. They usually utilize four diagnostic methods(observation, listening, interrogation, and pulse-taking) to understand the pathological conditions. Misdiagnosing would be lead to fatal results. Mining and learning diagnosing knowledge or experience from TCM records, literatures, or clinic cases by computer is significative for doctors. Nevertheless, understanding semantics from TCM records isn’t a easy thing in the state of the art.

Conventional topic models, like PLSA[9] and LDA[10], reveal the latent topics within corpus by implicitly capturing the document-level word co-occurrence patterns. We propose a new topic model SSTM to capture the relationship between symptoms and the syndromes from TCM clinical data. In the SSTM, symptoms is distribution on syndrome. Different from LDA, explicit variable symptoms is divided into cardinal symptoms in diagnosis and secondary symptoms. A cardinal symptom is main feature for a specific disease, but the secondary symptoms is subordinate for the disease. This is more accord with the process of TCM diagnose.

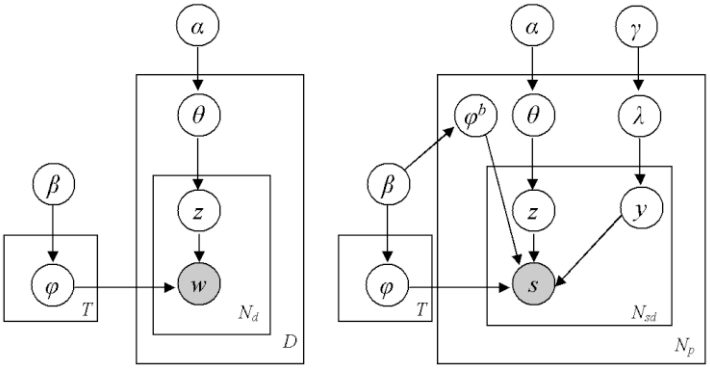


Fig. 1. Graphical models for (a)LDA,, (b) SSTM

Fig. 1(a) shows the graphical model for the “standard topic model”(LDA). D is the number of documents in the corpus and document d has N_d words. The process includes two steps: first, assign a topic number from document-topic distribution θ;

then, draw a word from topic-word distribution φ. All documents share T topics. Document-topic and topic-word distributions all obey the multinomial distributions, and each of them is governed by symmetric Dirichlet distribution. α and β are hyper-parameters of symmetric Dirichlet priors for θ and φ. Parameters θ and φ can be obtained through a Gibbs sampling.

Fig. 1(b) shows the graphical model for the "Symptom-Syndrome Topic Model"(SSTM). N_p is the total number of clinical cases, T is the topic number, z is from the corpus-level topic distribution θ. N_{sd} is total number of syndrome the patient have. Let φ denote the symptom distribution for topics and φ_b denote the symptom distribution for secondary symptoms. Let λ denote the Bernoulli distribution which controls the indicator y for the choice between b cardinal symptom and secondary symptom. Φ, θ, *and* φ_b all obey multinomial distributions, each of term is drawn from symmetric Dirichlet (β), and Dirichlet (α) respectively. λ is drawn from Beta (γ).

The probability of a symptom s is described as the follow(14):

$$p(s) = p(y=0)\sum_{z} p(z)p(s \mid z) + p(y=1)p(s \mid y=1) \tag{1}$$

The generative probability of a *patient* is expressed as:

$$p(patient) = \sum_{n=1}^{N_{sd}} \sum \ (p(y=0)\sum_{z} p(z)p(s \mid z) + p(y=1)p(s \mid y=1)) \tag{2}$$

4 Experiment

In this following, we propose experiment on SSTM in order to verify the effect of mining syndrome regularities in TCM data.

The CTM data come from China National Scientific Data Sharing Platform for Population and Health. We selected one of the databases: TCM Clinical Database of Diabetes. It has been classified into 19 subjects and 4351 records. We just focus on the II Diabetes. The data has the number of 1162 records.

In the experiment, α = 50/ T and β= 0.01, which are common settings in the literature[20]. γ is a prior of Bernoulli distribution, we set γ=0.5, it refers to another similar study[21]. We conducted the experiment on the SSTM and set T=7 by observing the data. We select the first five symptoms for each topic. The results are showed in Table 1.

Table 1. The first five symptoms for each topic

Topics	Symptoms
Topic1	Weak, sweating, shortness of breath, spontaneous perspiration, thirst
Topic2	Thirst, tough, yellow, constipine, polyphagy, feel, upset
Topic3	emaciation, weak, thirst, polydipsia, diuresis
Topic4	dizziness, thirst, palpitation, Shortness of Breath, dry mouth
Topic5	hiccough, pale tongue, prospermia, chilly, backache
Topic6	weak,pectoralgia, dark tongue, umbness of limbs, sluggish pulse
Topic7	thirst, dizziness, weak, exhausted, spontaneous perspiration

We invited experienced TCM doctors to analysis these results about Table 1. They considered those symptoms topic 1-topic 7 are related with some syndromes, for example, topic 1 is similar to qi and yin deficiency, topic 1 is similar to deficiency of yin and excessive heat syndrome, topic 5 is similar to deficiency of both yin and yang, etc.

5 Conclusions and future Work

TCM treat patient according to syndrome differentiation. Therefore, predicting the syndromes is vital work in the diagnosing. TCM is one of important medical branch theory system. In thousands of years, huge number of medical records that is recorded by nature language is accumulated by famous TCM doctors. Our work devote to mining the syndrome rules in these records. The proposed probabilistic model—symptom-syndrome topic model (SSTM) is effective to capture the connected knowledge between symptoms and syndromes. The further work is to continue to prefect SSTM and analysis the detail and complex relationship between symptoms and syndromes.

Acknowledgments.

This work was supported by the Graduate Student Scientific Research and Innovation Project of Jiangsu Province, China (China Central University Basic Scientific Research Business Fund, Hohai university, Grant No.: 2015B38314), and Science and Technology Projects of Huaian (Grant No.: HAS2015033), the University Science Research Project of Jiangsu Province(Grant No.:15KJB520004), and the Science and Technology Projects of Huaian (Grant No.:HAG2015060).

References

1. Zhang, X., Zhou, X., Huang, K., Feng, Q., Chen, S., and Liu, B.: Topic model for Chinese medicine diagnosis and prescription regularities analysis: Case on diabetes, Chinese Journal of Integrative Medicine, 17, 307 (2011).

2. Jiang, Z., Zhou, X., Zhang, X., and Chen, S.: Using link topic model to analyze traditional Chinese Medicine

Clinical symptom-herb regularities, in IEEE International Conference on E-Health Networking, Applications

and Services, pp. 15 (2012).

3. Zhou, X., Chen, S., Liu, B., Zhang, R., Wang, Y., Li, P., & Yan, X.: Development of traditional

Chinese medicine clinical data warehouse for medical knowledge discovery and decision support,

Artificial Intelligence in Medicine, 48, 139 (2010).

4. Zhou, X., Peng, Y., and Liu, B.:Text mining for traditional Chinese medical knowledge discovery: A survey, Journal of Biomedical Informatics, 43, 650 (2010).

5. Liu, C. X., and Shi, Y.: Application of data-mining technologies in analysis of clinical literature on traditional Chinese medicine, Chinese Journal of Medical Library & Information Science (2011).
6. Liu, B., Zhou, X., Wang, Y., Hu, J., He, L., Zhang, R., Chen, S., and Guo, Y.: Data processing and analysis in real-world traditional Chinese medicine clinical data:challenges and approaches, Statistics in Medicine, 31, 653 (2012).
7. Blei, D. M.: Probabilistic topic models, Communications of the ACM, 55, 77 (2012).
8. Thomas K. Landauer, P. W. F., Darrell Laham.:An Introduction to Latent Semantic Analysis, Discourse Processes, 25, 259 (1998).
9. Hofmann, T.: Probabilistic latent semantic indexing, in Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, ACM, pp. 50 (1999).
10. Blei, D. M., Ng, A. Y., and Jordan, M. I.: Latent dirichlet allocation, the Journal of machine Learning research, 3, 993 (2003).
11. Hong, L., and Davison, B. D.: Empirical study of topic modeling in Twitter, Proceedings of the Sigkdd Workshop on Social Media Analytics, 80 (2010).
12. Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E. P., Yan, H., and Li, X.:Comparing Twitter and Traditional Media Using Topic Models, in In ECIR, pp. 338 (2011).
13. Rosen-Zvi, M., Griffiths, T., Steyvers, M., and Smyth, P., The author-topic model for authors and documents: in Proceedings of the 20th conference on Uncertainty in artificial intelligence, AUAI Press, pp. 487 (2004).
14. Yan, X., Guo, J., Lan, Y., and Cheng, X.: A biterm topic model for short texts, in Proceedings of the 22nd international conference on World Wide Web, International World Wide Web Conferences Steering Committee, pp. 1445 (2013).
15. Yi, F., Wu, Z., Zhou, X., Zhou, Z., and Fan, W.: Knowledge discovery in traditional Chinese medicine: State of the art and perspectives, Artificial Intelligence in Medicine, 38, 219 (2006).
16. Lukman, S., He, Y., and Hui, S. C., Computational methods for Traditional Chinese Medicine: A survey, Computer Methods & Programs in Biomedicine, 88, 283 (2007).
17. Wu, Z., Chen, H., and Jiang, X.: 1 – Overview of Knowledge Discovery in Traditional Chinese Medicine 1, 1 (2012).
18. Yao, L., Zhang, Y., Wei, B., Wang, W., Zhang, Y., and Ren, X.: Discovering treatment pattern in traditional Chinese medicine clinical cases using topic model and domain knowledge, in IEEE International Conference on Bioinformatics and Biomedicine, pp. 191 (2014).
19. Liang, Y., Yin, Z., Wei, B., Wei, W., Zhang, Y., Ren, X., and Bian, Y.: Discovering treatment pattern in Traditional Chinese Medicine clinical cases by exploiting supervised topic model and domain knowledge, Journal of Biomedical Informatics, 58, 425 (2015).
20. Heinrich, G.: Parameter estimation for text analysis, Technical Report (2004).
21. Chemudugunta, C., Smyth, P., Steyvers, M.: Modeling General and Specific Aspects of Documents with a Probabilistic Topic Model Vol. 19, MIT Press (2007).
22. Ma J, Zhang Y, Wang Z, et al.: A Message Topic Model for Multi-Grain SMS Spam Filtering. International Journal of Technology and Human Interaction (IJTHI), 2016, 12(2): 83-95.

Fuzzy on FHIR: a Decision Support service for Healthcare Applications

Aniello Minutolo, Massimo Esposito, Giuseppe De Pietro
Institute for High Performance Computing and Networking, ICAR-CNR
via P. Castellino, 111-80131, Napoli, Italy
{aniello.minutolo, massimo.esposito, giuseppe.depietro}@icar.cnr.it

Abstract. In the last years, an explosion of interest has been seen with respect to clinical decision support systems based on guidelines, since they have promised to reduce inter-practice variation, to promote evidence-based medicine, and to contain the cost of health care. Despite this great promise, many obstacles lie in the way of their integration into routine clinical care. Indeed, first, the communication with information systems to collect health data is a very thorny task due to the heterogeneity of data sources. Secondly, the machine-readable representation of guidelines can generate an unrealistic over-simplification of reality, since not able to completely handle uncertainty and imprecision typically affecting guidelines. Finally, a large number of existing decision support systems have been implemented as standalone software solutions that cannot be well reused or transported to other medical scenarios. Starting from these considerations, this paper proposes a standards-based decision support service for facilitating the development of healthcare applications enabling: i) the encoding of uncertain and vague knowledge underpinning clinical guidelines by using Fuzzy Logic; ii) the representation of input and output health data by using the emerging standard FHIR (Fast Healthcare Interoperability Resources). As a proof of concept, a WSDL-based SOAP implementation of the service has been tested on a set of clinical guidelines pertaining the evaluation of blood pressure for a monitored patient.

1 Introduction

In the last years, the advances in information technologies have drastically changed the way healthcare services are provided, by offering the chance to build smart, innovative, and personalized solutions able to support individuals in their daily activities and to improve their quality of life. [1-4]. The seamless integration of information technologies has been more and more investigated for building innovative healthcare facilities able to provision smart and efficient services to the individuals with the goal of promoting self-care and preventive care [2, 5].

In particular, an explosion of interest has been seen with respect to clinical decision support systems based on guidelines. The use of these systems has promised to reduce inter-practice variation, to promote evidence-based medicine, and to contain the cost of health care. Despite the great promise of the clinical use of decision support systems, many obstacles lie in the way of their integration into routine clinical care.

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_16

Firstly, the exploitation of guidelines in decision-support systems requires that health data must be collected from information systems of medical settings and processed in order to generate contextually relevant recommendations. However, the heterogeneity of medical data sources, which may differ in the data models, naming conventions, and level of detail used to represent similar data, has produced either stand-alone solutions or small components directly customized with respect to specific medical information systems.

Several decision support solutions have been presented in literature to face the problem of handling health data residing at different sources, by using standardized representations aimed at supporting the re-usability and interoperability with existing healthcare repositories and services [6-9]. Most of them employ Health Level Seven (HL7) International compliant data, following the prevailing standards at the healthcare enterprise [10]. Although these attempts, the goal of interoperability with medical information systems has remained elusive for clinical decision support systems, due to the complexity and expensiveness of HL7 standards to be implemented in real world implementations.

Secondly, the machine-readable representation of guidelines in decision-support systems can generate an unrealistic over-simplification of reality. Indeed, many knowledge representation formalisms [2, 9] exist, which are able to model medical knowledge by granting expressiveness, upgradability and maintainability. However, they are not able to completely handle uncertainty and imprecision typically affecting guidelines. For instance, hypotension guidelines state that patients with values of systolic blood pressure lower than 90 mmHg could be affected by hypotension. Thus, completely different conclusions could be suggested for patients with values of systolic blood pressure that are close but placed around the threshold (e.g. for patients with values of systolic blood pressure equal to 91 mmHg and 89 mmHg, respectively), so as to lead to possible wrong recommendations.

Finally, a large number of decision support systems have been implemented as standalone software solutions that cannot be well maintained or reused, except by their authors, and therefore cannot be easily transported to other medical scenarios.

Starting from these considerations, this paper proposes Fuzzy on FHIR, a standards-based decision support service for facilitating the development of healthcare applications enabling: i) the representation of input and output health data by using the emerging standard FHIR (Fast Healthcare Interoperability Resources) created by the HL7 Organization; ii) the encoding of uncertain and vague knowledge underpinning clinical guidelines by using Fuzzy Logic. Moreover, by adopting the service-oriented paradigm, this solution takes advantage of important software engineering features such as reuse, ease of maintenance, and cross-platform capabilities.

As a proof of concept, a WSDL-based SOAP implementation of the proposed service has been tested on a set of clinical guidelines pertaining the evaluation of blood pressure for a monitored patient.

The rest of the paper is organized as follows. Section 2 describes some preliminary notions. Section 3 presents the proposed service in detail, whereas Section 4 describes the proof of concept application. Finally, Section 5 concludes the work.

2 Background and preliminaries

Decision support in healthcare applications is typically provided by means of declarative logical frameworks able to represent clinical guidelines, so as to provide smart and case-specific advices to individuals and doctors. Among these declarative logical frameworks, Fuzzy Logic has been profitably used for modelling uncertainty and vagueness of clinical guidelines in the form of if-then rules, aimed at reproducing the reasoning process followed by doctors when health data are evaluated [11-14].

The usage of Fuzzy Logic to build a clinical decision support solution foresees that, on the one hand, the main concepts charactering the medical domain are represented through a fuzzy linguistic model composed of linguistic variables, linguistic terms, and membership functions whose shapes encode uncertainty and vagueness. On the other hand, on top of these linguistic variables, a fuzzy rule model is defined as a set of if-then rules that represent the clinical guidelines. Finally, the fuzzy linguistic and rule models can be used for reasoning on health data in order to provide recommendations based on the modelled medical knowledge. To this aim, input health data have to be converted to some fuzzy sets by applying a fuzzification technique and, afterward, a fuzzy inference can be made on the basis of the defined rule model. Lastly, the resulting inference outputs, i.e. the fuzzy shapes inferred by the rules, and their corresponding defuzzified values, are returned and interpreted for producing the desired feedbacks to the users.

Thus, in order to realize a decision support solution based on Fuzzy Logic, a domain-specific language is required in order to unambiguously specify both the fuzzy linguistic and rule models representing the medical knowledge of interest and the input (output) data that are submitted to (generated from) it at runtime.

To this aim, on the one hand, XML-derived technologies have been profitable used in the past for creating data-oriented markup languages able to unambiguously describe the configuration of fuzzy reasoning models [14]. On the other hand, in order to also enable the integration with medical information systems, facilities for healthcare interoperability have been proposed in the last years, which are intimately connected to the health clinical data standards developed by HL7 [10].

HL7 standards adopt complex data models and architectures that, recently, have been criticized as unsuitable in a variety of scenarios, such as mobile applications and cloud computing, and, thus, the use of web services as a smoother alternative to HL7's messaging protocols has been more and more suggested for the proper exchange of health information and messages [15]. As a consequence, a simplified data model, named Fast Healthcare Interoperability Resources (FHIR), has been proposed, based on HL7's extremely detailed Reference Implementation Model (RIM). FHIR is currently still a draft standard in development but its architecture is clear. FHIR provides a 100-150 JSON or XML objects (the FHIR Resources), each of which contains a group of data items related to a particular clinical concept such as a medication, clinical observation or study [16].

Thus, a proper fuzzy-based decision support solution able to communicate with medical information systems and re-use existing healthcare resources, should admit input data codified according the FHIR data model and produce inference outcomes containing output data compliant with it. Such a way, decision support solutions could easily interoperate with existing healthcare components, repositories and/or

infrastructures in order to acquire FHIR datasets, produce FHIR-compliant recommendations and enable their memorization in medical repositories.

All these considerations represent the rationale of the decision support service proposed in this work and described in the next section.

3 The proposed decision support service

The proposed standards-based decision support service enables the execution of fuzzy reasoning mechanisms on top of FHIR-based data. According to the client-server architectural view shown in Figure 1, a client application can invoke the proposed service for submitting an inference request on the basis of a reasoning model and the input health dataset to evaluate. After receiving the request, the service verifies its validity and the consistency of the reasoning model. In case of invalid request, the service signals the error, otherwise, the service executes the requested inference and returns to the client the produced outcomes. Additionally, in order to reduce the amount of data exchanged when the service is invoked, the inference requests can be also made on the basis of previously configured reasoning models.

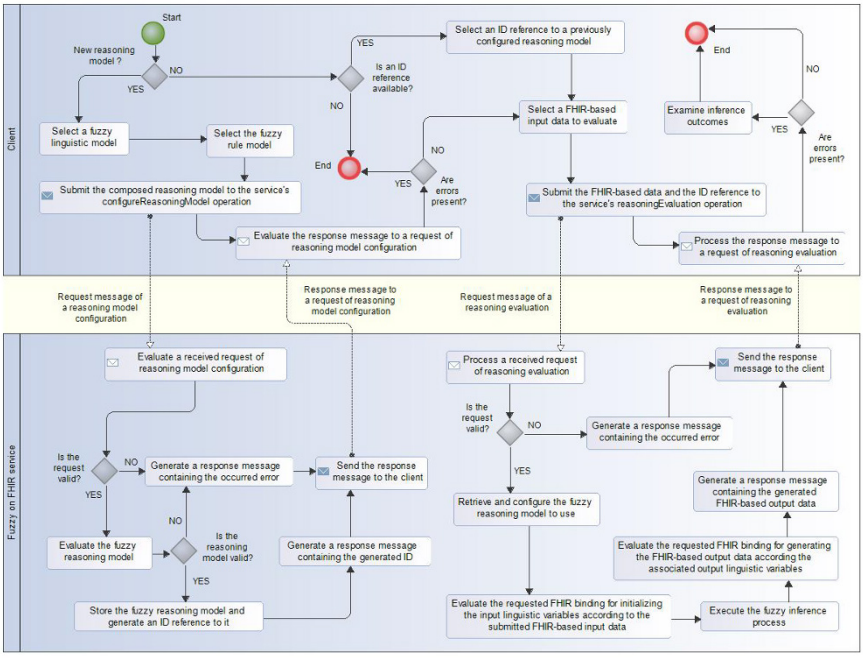


Fig. 1. The client-server architectural view of the decision support service.

As a result, by exploiting and configuring the reasoning model, client applications can personalize the decision support offered by the service and, contextually, decouple the logic about the service behavior and the supported health input and output data, from the current service implementation. In the following, a description

of all the input and output parameters for the service as well as more details about its operations are given.

2.1 Input and output parameters for the service

Health data submitted to the service and recommendations generated by it are represented by FHIR resources characterized by attributes and eventual relations with other FHIR resources. In detail, a FHIR resource *R* is encoded according to an XML-based representation, i.e. a hierarchy of XML nodes where the root represents the resource *R*, and its children $C_R(A_1..A_n)$ represent the *n* attributes $A_1..A_n$ existing for *R*. A child node $C_R(A_i)$ can model a relation with another FHIR resource and, thus, it contains other child nodes modelling that resource. Alternatively, a child node $C_R(A_i)$ can model an attribute whose value is a primitive value, i.e. textual or numerical, specified through the *value* attribute of $C_R(A_i)$.

As an example, Figure 2 reports the FHIR resource named *Observation*, encoding values of systolic and diastolic blood pressures for a patient.

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<ns3:observation xmlns:ns2="http://www.w3.org/1999/xhtml"
                 xmlns:ns3="http://hl7.org/fhir"
                 xmlns:ns4="http://www.ws.ehealthnet.neatec.it">
    <ns3:id value="blood-pressure"/>
    <ns3:identifier>
        <ns3:system value="urn:ietf:rfc:3986"/>
        <ns3:value value="urn:uuid:187e0c12-8dd2-67e2-99b2-bf273c878281"/>
    </ns3:identifier>
    <ns3:status value="final"/>
    <ns3:code>
        <ns3:coding>
            <ns3:system value="http://loinc.org"/>
            <ns3:code value="55284-4"/>
            <ns3:display value="Blood pressure systolic &amp;amp; diastolic"/>
        </ns3:coding>
    </ns3:code>
    <ns3:subject>
        <ns3:reference value="Patient/example"/>
    </ns3:subject>
    •
    •
    •
            <ns3:code value="bp-s"/>
            <ns3:display value="Systolic Blood pressure"/>
        </ns3:coding>
    </ns3:code>
    <ns3:valueQuantity>
        <ns3:value value="107"/>
        <ns3:unit value="mm[Hg]"/>
    </ns3:valueQuantity>
</ns3:component>
<ns3:component>
    <ns3:code>
        <ns3:coding>
            <ns3:system value="http://loinc.org"/>
            <ns3:code value="8462-4"/>
            <ns3:display value="Diastolic blood pressure"/>
        </ns3:coding>
    </ns3:code>
    <ns3:valueQuantity>
        <ns3:value value="60"/>
        <ns3:unit value="mm[Hg]"/>
    </ns3:valueQuantity>
</ns3:component>
</ns3:observation>
```

Fig. 2. An example of medical observation codified according to FHIR.

The reasoning model submitted to the service contains three kind of knowledge: i) the linguistic model describing the domain knowledge of interest; ii) the rule model and the whole set of parameters needed to configure the inference scheme; iii) the definition of an arbitrary binding between the linguistic variables and the FHIR structure of the data of interest.

The linguistic model contains the structural knowledge charactering the linguistic variables, and the linguistic terms associated to them. The proposed service has been designed in order to support fuzzy linguistic variables and terms defined by composing several kinds of fuzzy memberships, such as *Trapezoidal, Triangular, Singleton*, and *Piecewise*.

The rule model contains both the procedural knowledge modelling the reasoning logic to reproduce, and the whole set of parameters needed to configure the inference scheme to apply. In detail, the proposed service has been designed in order to support procedural knowledge composed of groups of fuzzy if-then rules, each one containing all rules having the same linguistic variable in their consequent parts. Moreover, each group of rules can be characterized by peculiar logical parameters, i.e. the methods used for the fuzzy *implication*, *aggregation* and *defuzzification* processes, and the

fuzzy *And* and *Or* operators to connect the different conditions in antecedent parts of the rules belonging to a group. Such a way, clinical guidelines, typically composed of isolated recommendations linked to the same final action, can be modeled by a dedicated group of rules characterized by a peculiar inference configuration.

Finally, given a linguistic model, the Fuzzy-FHIR binding describes, on the one hand, how a FHIR input dataset can be interpreted for initializing some linguistic variables, and, on the other hand, how the inferred values of linguistic variables can be used for generating FHIR-compliant recommendations. The proposed service allows specifying the Fuzzy-FHIR binding on the basis of the XML-based representation of FHIR resources of interest. In detail, the Fuzzy-FHIR binding between a linguistic variable *v* and a reference FHIR resource *r* describes the hierarchy that a XML-based representation of *r* must contain to be properly bound to *v*. In case when such a hierarchy appears inside a FHIR input dataset, it is applied to initialize *v* whenever *v* is used as an input variable. Otherwise, in case when *v* is used as an output variable, the required hierarchy is applied, at the end of the inference process, to create XML-based representations of *r*.

All the reasoning model has been implemented according to the XML-based representation given by the FdsL language [14]. This latter is an intuitive XML-based language specifically designed for supporting the representation of medical knowledge in the form of groups of if-then rules, where each group contains all rules having the same linguistic variable in their consequent parts and can be configured by a peculiar inference scheme [14]. As an example, Fig. 3 reports the FdsL representation of a linguistic variable, named *DetectedIssueAction,* whose parameters *name*, *range*, *unit*, *lowestx*, *highestx* describe, respectively, the name of the linguistic concept, the range of admitted values and how they are measured, and the limits of the universe of discourse. Moreover, that variable is also described as composed of three fuzzy sets, named *11*, *12*, and *13*, whose membership function is as *Singleton.*

```
<fuzzyvariable name="DetectedIssueAction" range="real" unit="x" lowerx="0.0" higherx="1.0">
    <fhir_resource_binding reference="ns3:DetectedIssue" target_node="ns3:code">
        <required_attribute name="xmlns:ns3" value="http://hl7.org/fhir"/>
        <required_hierarchy>
            <node name="ns3:mitigation" parent_node="ns3:DetectedIssue"/>
            <node name="ns3:author" parent_node="ns3:mitigation"/>
            <node name="ns3:reference" value="Practitioner/example" parent_node="ns3:author"/>
            <node name="ns3:display" value="Dr. Adam Careful" parent_node="ns3:author"/>
            <node name="ns3:action" parent_node="ns3:mitigation"/>
            <node name="ns3:coding" parent_node="ns3:action"/>
            <node name="ns3:system" value="http://hl7.org/fhir/v3/ActCode" parent_node="ns3:coding"/>
            <node name="ns3:code" parent_node="ns3:coding"/> </required_hierarchy>
    </fhir_resource_binding>
    <fuzzyset name="11" type="Singleton">
        <point x="0.0" y="1.0"/> </fuzzyset>
    <fuzzyset name="12" type="Singleton">
       <point x="0.5" y="1.0"/> </fuzzyset>
    <fuzzyset name="13" type="Singleton">
        <point x="1.0" y="1.0"/> </fuzzyset>
</fuzzyvariable>
```

Fig. 3. FdsL fragment encoding an example of linguistic variable and its Fuzzy-FHIR binding.

In order to also encode the Fuzzy-FHIR binding, the FdsL language, has been extended by introducing a new tag element, named <*fhir_resource_binding*>, which optionally links a linguistic variable to a specific FHIR resource. In detail, given an input (output) linguistic variable *v*, this tag uses its *reference* attribute to indicate the name of the XML node representing the FHIR resource connected to *v*, and its

target_node attribute to indicate the specific child node of the *reference* node which contains (will contain) the value to associate (associated) to *v*. Moreover, this tag also admits a nested one, named <*required_hierarchy*>, to specify the requested hierarchy of the *reference* node, and multiple nested tags, named <*required_attribute*>, to describe names and values of the *reference* node's attributes. With respect to the <*required_hierarchy*> tag, it admits an arbitrary number of child tags, named <*node*>, each one representing a child node of *reference* node. The <*node*> tag uses the *name* attribute to specify the name of the child XML node, the *parent_node* attribute to indicate its parent in the hierarchy, and the *value* attribute to specify a value that the node has to assume. It is worth noting that the value of the *target_node* must be the name of a child node existing in the hierarchy.

With respect to the example reported in Fig. 3, the <*fhir_resource_binding*> tag establishes the binding between the variable *DetectedIssueAction* and the FHIR resource *DetectedIssue*, whose representation is a XML node named *ns3:DetectedIssue*. Moreover, the <*required_attribute*> tag specifies that the *ns3:DetectedIssue* node must have an attribute named *xmlns:ns3*, whose value must be *http://hl7.org/fhir*. Lastly, the <*required_hierarchy*> tag describes the requested hierarchy for the *ns3:DetectedIssue* node and, eventually, the values assumed by its child nodes. Among the specified child nodes, one of them must be indicated by the *target_node* attribute, in this case the node *ns3:code*.

In case when the *DetectedIssueAction* variable is used as an input variable, the binding is used to inspect an input dataset and verify if the required hierarchy exists in it. If it exists, the *DetectedIssueAction* variable is initialized by using the value assumed in the dataset by the node *ns3:code*. On the other hand, in case when the *DetectedIssueAction* variable is used as an output variable, the binding is used to generate a corresponding nodes hierarchy and attributes values, and by assigning to the *ns3:code* node the defuzzified value assumed by *DetectedIssueAction* at the end of the inference process.

2.2 The service operations

The service exposes two different functionalities, namely *configureReasoningModel* and *reasoningEvaluation,* in the form of operations.

The *configureReasoningModel* operation verifies the consistency of a reasoning model. In case of inconsistency, an error message is returned to the client application. Alternatively, the reasoning model is stored and a unique identifier is returned to the client application for a later use. On the other hand, the *reasoningEvaluation* operation receives inference requests on the basis of a reasoning model identifier and the input dataset to evaluate and, thus, it builds a fuzzy inference system on the basis of them. In particular, the input dataset is evaluated to initialize the input linguistic variables. To this aim, the Fuzzy-FHIR binding eventually specified for the input linguistic variables is used. Successively, the inference process is started and the rule model is repeatedly compared with current fuzzy values of linguistic variables in order to infer new knowledge or update the existing one with the final aim of generating one or more recommendations. Lastly, at the end of the inference process, the service evaluates the Fuzzy-FHIR binding eventually specified for the output

linguistic variables in order to generate a FHIR-compliant response on the basis of their inferred fuzzy values and returns it to the client.

4 A proof of concept application

As a proof of concept, the proposed solution has been implemented as a WSDL-based SOAP web service, and tested for modelling and reasoning on a set of clinical guidelines pertaining the evaluation of the blood pressure for a monitored patient. In detail, two simple guidelines have been formulated, each of them made of two clinical recommendations, as reported in the following:

r1a – IF DiastolicPressure is high THEN DetectedIssueCategory is FOOD;
r1b – IF SystolicPressure is high THEN DetectedIssueCategory is DRG;
r2a – IF DetectedIssueCategory is FOOD THEN DetectedIssueAction is 11;
r2a – IF DetectedIssueCategory is DRG THEN DetectedIssueAction is 12.

Each guideline has been modeled as a group of two rules, where the first rule group enables to identify a category associated to a detected issue when suspicious values of blood pressure are recognized, whereas the second group aims at determining the action to take, or the observation to make, in order to reduce/eliminate the risk associated to the detected issue.

Afterwards, the fuzzy linguistic model has been codified in FdsL, as reported in the Figure 4, based on the ranges of the considered health data, taking into account their representation in FHIR. In detail, both the systolic and diastolic blood pressures have been modeled via a set of trapezoidal membership functions, whereas category and action linked to the detected issue have been modeled via a set of singleton membership functions.

Fig. 4. FdsL fragment of the fuzzy linguistic model for the considered case of study.

Moreover, the Fuzzy-FHIR binding between the modeled linguistic variables and some FHIR resources has been established through the *<fhir_resource_binding>* tag. As an example, it is reported the binding established between the *DiastolicPressure* variable and a FHIR *Observation* resource, as the one reported in Fig. 2, containing the value of diastolic blood pressure. Fig. 4 also reports the binding established between the output *DetectedIssueCategory* variable and the FHIR *DetectedIssue* resource. Finally, the rule model and the whole set of parameters needed to configure the inference scheme have been codified in FdsL as reported in the Figure 5.

Lastly, in order to test the service, as a first step, the reasoning model just described has been given as input to the *configureReasoningModel* operation.

```
<?xml version="1.0"?>
<fuzzyinferencesystem name="test">
    <rulesBlock name="DetectedIssue1" outputVar="DetectedIssueCategory" andMethod="min"
            orMethod="max" implicationMethod="min" aggregationMethod="max" defuzzificationMethod="centerOfGravitySingleton">
        <rule name="r1" weight="1.0">
            <IF> <RuleClause variable="DiastolicPressure" connector="is" set="high"/> </IF>
            <THEN> <RuleClause variable="DetectedIssueCategory" connector="is" set="FOOD"/> </THEN> </rule>
        <rule name="r2" weight="1.0">
            <IF> <RuleClause variable="SystolicPressure" connector="is" set="high"/> </IF>
            <THEN> <RuleClause variable="DetectedIssueCategory" connector="is" set="DRG"/> </THEN> </rule>
    </rulesBlock>
    <rulesBlock name="DetectedIssue2" outputVar="DetectedIssueAction" andMethod="min"
            orMethod="max" implicationMethod="min" aggregationMethod="max" defuzzificationMethod="centerOfGravitySingleton">
        <rule name="r1" weight="1.0">
            <IF> <RuleClause variable="DetectedIssueCategory" connector="is" set="FOOD"/> </IF>
            <THEN> <RuleClause variable="DetectedIssueAction" connector="is" set="11"/> </THEN> </rule>
        <rule name="r2" weight="1.0">
            <IF> <RuleClause variable="DetectedIssueCategory" connector="is" set="DRG"/> </IF>
            <THEN> <RuleClause variable="DetectedIssueAction" connector="is" set="12"/> </THEN> </rule>
    </rulesBlock>
</fuzzyinferencesystem>
```

Fig. 5. FdsL fragment of the encoded rule model for the considered case of study.

Successively, the returned identifier and the FHIR input dataset reported in Figure 2 have been used as input for the *reasoningEvaluation* operation. Finally, the FHIR-compliant response generated by this operation is shown in Figure 6.

```
<?xml version="1.0" encoding="UTF-8"?>
<ns3:DetectedIssue xmlns:ns3="http://hl7.org/fhir">
    <ns3:category>
        <ns3:coding>
            <ns3:code value="DRG"/>
            <ns3:system value="http://hl7.org/fhir/v3/ActCode"/>
        </ns3:coding>
    </ns3:category>
    <ns3:mitigation>
        <ns3:action>
            <ns3:coding>
                <ns3:code value="12"/>
                <ns3:system value="http://hl7.org/fhir/v3/ActCode"/>
            </ns3:coding>
        </ns3:action>
        <ns3:author>
            <ns3:display value="Dr. Adam Careful"/>
            <ns3:reference value="Practitioner/example"/>
        </ns3:author>
    </ns3:mitigation>
</ns3:DetectedIssue>
```

Fig. 6. The FHIR-compliant response generated by the service for the considered case of study.

5 Conclusions

In this paper, a standards-based decision support service has been presented, aimed at facilitating the development of healthcare applications enabling: i) the representation of input and output health data by using the emerging standard FHIR; ii) the encoding of uncertain and vague knowledge underpinning clinical guidelines by using Fuzzy Logic. Moreover, thanks to the adoption of the service-oriented paradigm, this solution exhibits important software engineering features such as reuse, ease of maintenance, and cross-platform capabilities.

The novelty of the proposed service relies on its capability to simply establish a bridge among a fuzzy representation of medical knowledge and health data coming from existing medical information systems. Moreover, the logic about the service behavior and the supported health input and output data are decoupled from the current service implementation, giving the possibility of offering personalized decision support depending on the specific application requirements.

As a proof of concept, a WSDL-based SOAP implementation of the service has been tested on a set of clinical guidelines pertaining the evaluation of blood pressure for a monitored patient. Next step of the research activities will regard the design and development of a visual framework offering facilities for the simple configuration and testing of the service, as well as for the inspection of its inferred results.

Acknowledgments. This work has been partially supported by the Italian project "eHealthnet" funded by the Italian Ministry of Education, University, and Research.

References

1. D. M. Malvey, and D. J. Slovensky, "mHealth: Transforming Healthcare". Springer, 2014.
2. A. Minutolo, M. Esposito, G. De Pietro, "Development and customization of individualized mobile healthcare applications", in Proceedings of the 3rd IEEE International Conference on Cognitive Infocommunications (CogInfoCom), pp. 321 – 326, 2012.
3. F. Amato, F. Moscato, "A model driven approach to data privacy verification in E-Health systems". Transactions on Data Privacy, vol. 8(3), pp. 273-296, 2015.
4. F. Amato, M. Barbareschi, V. Casola, A. Mazzeo, "An fpga-based smart classifier for decision support systems". In Intelligent Distributed Computing VII, pp. 289-299, 2014.
5. S. Rodriguez-Loya, K. Kawamoto, "Newer Architectures for Clinical Decision Support". In Clinical Decision Support Systems, pp. 87-97, Springer International Publishing, 2016.
6. G. Jiang et al., "A Standards-based Semantic Metadata Repository to Support EHR-driven Phenotype Authoring and Execution". In IOS Press, 2015.
7. M. Khalilia, M. Choi, A. Henderson, S. Iyengar, M. Braunstein, J. Sun, "Clinical Predictive Modeling Development and Deployment through FHIR Web Services". In AMIA Annual Symposium Proceedings, p. 717, American Medical Informatics Association, 2015.
8. G. C. Lamprinakos et al., "An integrated remote monitoring platform towards telehealth and telecare services interoperability". Information Sciences, vol. 308, pp. 23-37, 2015.
9. Y. F. Zhang, et al., "Integrating HL7 RIM and ontology for unified knowledge and data representation in clinical decision support systems". Computer methods and programs in biomedicine, vol. 123, pp. 94-108, 2016.
10. R.H. Dolin, L. Alschuler, S. Boyer, C. Beebe, F.M. Behlen, P.V. Biron, A.S. Shvo, "HL7 clinical document architecture, release 2, Journal of American Medical Informatics Association, vol. 13, pp. 30–39, 2006.
11. A. Minutolo, M. Esposito, G. De Pietro, "A fuzzy framework for encoding uncertainty in clinical decision-making". Knowledge-Based Systems, vol. 98, pp. 95-116, 2016.
12. J. Warren,G. Beliakov, B. Zwaag, "Fuzzy logic in clinical practice decision support system". Proceedings of the 33rd Hawaii Inter. Conference on System Sciences, 2000.
13. S. Alayón, R. Robertson, S.K. Warfield, J. Ruiz-Alzola, "A fuzzy system for helping medical diagnosis of malformations of cortical development". J. B. Inf. 40, 221–235, 2007.
14. A. Minutolo, M. Esposito, G. De Pietro, "A Fuzzy Decision Support Language for building Mobile DSSs for Healthcare Applications", in Proceedings of Wireless Mobile Communication and Healthcare, LNICST, vol. 61, pp. 263-270, Springer, 2013.
15. E. Muir, "The Rise and Fall of HL7". A blog post of Eliot Muir, funder of Interfaceware, Inc., http://www.interfaceware.com/blog/the-rise-and-fall-of-hl7/, 2011.
16. D. Bender, K. Sartipi, "HL7 FHIR: An Agile and RESTful approach to healthcare information exchange". In Computer-Based Medical Systems (CBMS), IEEE 26th International Symposium, pp. 326–331, 2013.

Electric Mobility in Green and Smart Cities

Adrian-Gabriel Morosan[1], Florin Pop[2], and Aurel-Florin Arion[3]

Computer Science Department[1,2] Engineering Graphic and Design Department[3],
University *Politehnica* of Bucharest, ROMANIA
morosan.ag@gmail.com,florin.pop@cs.pub.ro,aurel@rnvgroup.com

Abstract. Because the electric mobility has its focus on eco-friendly means of transport, a distributed platform designed for a smart city environment that can manage the electrical charging stations is vital. Even though there is only one de facto protocol that is widely implemented, there are many vendors who have chosen different interfaces, protocols and implementations. The purpose of this paper is to present a flexible architecture of a distributed system that can face the addition of new features, interfaces, protocols with focus on REST over WebSocket, a good alternative in terms of efficiency, traffic and costs. The proposed platform is distributed, being divided into several components, each responsible for dedicated parts of the communication and business logic. Our solution is an interactive networked environment for green and smart cities offering flexible support to electrical vehicles management, with a high impact in decarbonizing transport and in keeping the environment clean.

Keywords: Electric Mobility, Smart City, Interactive Networked Environment, OCPP, OSCP, OCHP

1 Introduction

In the actual context where pollution is one of the most challenging problems, electrical vehicles could play a crucial role in decarbonizing transport and in keeping the environment clean. Due to increasing need of electrical vehicles, the smart grid will be harder to monitor and manage. For this purpose, the usage of open protocols in communication between entities will address this issue.

A study commissioned by German National Academy of Science and Engineering has predicted an increasing trend in using electrical vehicles till 2020. This research was conducted based on the trend that is already increasing (from 2015 till now) and based on other different trends related to it: electricity price, battery price and fuel price. As expected, the number of electrical vehicles by 2020 is proportional to higher fuel price and inversely to battery and electricity prices.

Figure 1 shows an increasing trend in purchasing electrical vehicles in Germany. As we can see, the vast majority of them is represented by private cars and only a small part consists of fleet or company cars. Furthermore, the number of cars per user group is proportional to the number of charging points per user

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_17

group. For example, most of the electrical vehicles are owned by private sector, the private wall boxes are the preferred ones [1].

The main contributions of this paper are as follows:

- design and implementation of a distributed architecture of interactive networked environment for green and smart cities offering flexible support to electrical vehicles management;
- an analysis of open protocols for electric mobility;
- design of Business Logic Controller and WebSocket Interface Controller:
- demonstrate the efficiency of REST over WekSocket in comparison with REST over HTTP, in terms of traffic.

The paper is structured in 7 sections. Section 2 presents a critical analysis of open protocols for electric mobility. Section 3 presents the architecture and implementation of Open Charge Point Protocol with a comparative view to similar projects. The experimental methodology is presented in Section 4, which is followed by results and performance analysis in Section 5. In Section 6 we present our future work. The paper ends with conclusions (Section 7).

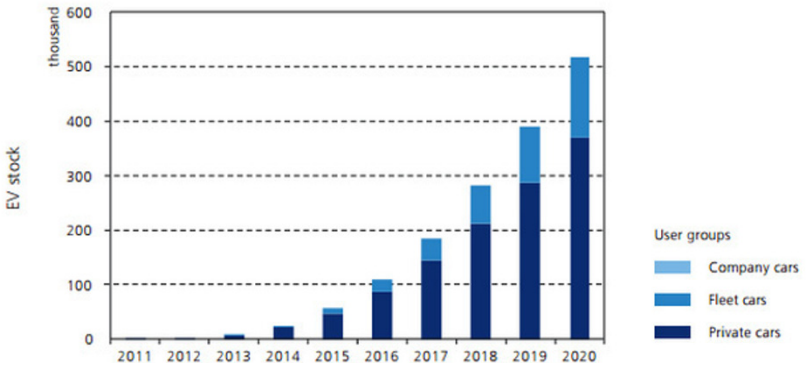


Fig. 1. The increasing need of an alternative to fueled vehicles in Europe [1].

2 Open protocols for electric mobility

Because the new models of smart grids include electrical vehicles, there were some standards in order to establish communication between sub-entities of the entire system. The domain of e-mobility contains specifically developed protocols, used in the interaction between electric vehicles (EVs), charging stations, central servers, power grids, energy providers. The most important protocol with a specific focus on one or many channels of interaction are:

- OCPP - Open Charge Point Protocol;

– OSCP - Open Smart Charging Protocol;
– OCHP - Open Clearing House Protocol;
– ISO/IEC 15118 Road vehicles Vehicle to grid communication interface [2].

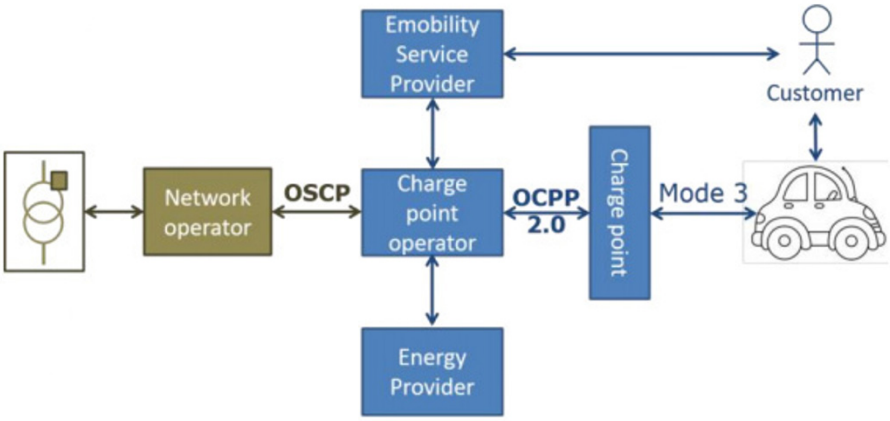


Fig. 2. OCPP and OSCP in EV smart grids [2].

2.1 OCPP - Open Charge Point Protocol

The purpose of this protocol is to standardize the communication between a charging station and a central server. The project was initiated as an open standard. Firstly, the standard used SOAP over HTTP for data exchange, from newer versions REST over HTTP is also available. This protocol explicitly requires both sides to initiate communication. For example, a communication initiated by EVSE - Electric Vehicle Supply Equipment (charge point/charging station) is to request authentication of EVs/customers for charging and one initiated by the central system could be to request a firmware update. The current versions are freely available and used all over Europe, but not only [2].

2.2 OSCP - Open Smart Charging Protocol

The Open Smart Charging Protocol has been implemented at Dutch DSO Exenis and officially released in May 2015. In this case, in comparison with OCPP, the communication is between different entities: DSO (Distribution System Operator) and CSP (Charge Service Provider). DSO manages the loading of the local grid. For this DSO periodically sends a capacity forecast with the information regarding the available capacity for charging EVs. In our context, CSP means the party that pays for the electricity. As presented in figure 2 both OCPP and

OSCP plays different roles and are suitable to work together. In the complete solution, OCPP could be replaced by another protocol with similar purpose [3].

OSCP consists of two main messages: UpdateCableCapacityForecast - the message DSO is sending out to inform about the forecast of cable capacity and the backup capacity to CSO (Charge Service Operator) and RequestAdjusted-Capacity - enables CSO to request an extra capacity when is necessary, being DSOs decision to adhere or not to this request. In the protocol, the frequency of this message is not mandated, however, the default value is 15 minutes.

Besides the main messages, there are three addition messages: GetCapacity-Forecast - this message enables CSO to request a new forecast, UpdateAggre-gatedUsage - message sent by CSO to inform about the total usage per CSO and Heartbeat - used to assure the CSO that forecasting algorithm is still running and forecasts are updated and periodically sent [3].

2.3 OCHP - Open Clearing House Protocol

The purpose of OCHP is to enable charging of electric vehicles across multiple station networks. In this way, multiple actors in the field of electric mobility are connected, in a simple way. OCHP is an open source protocol and free to implement or to participate in the development process of it. From the first version till now (version 1.3), it offers a uniform interface solution, SOAP based.

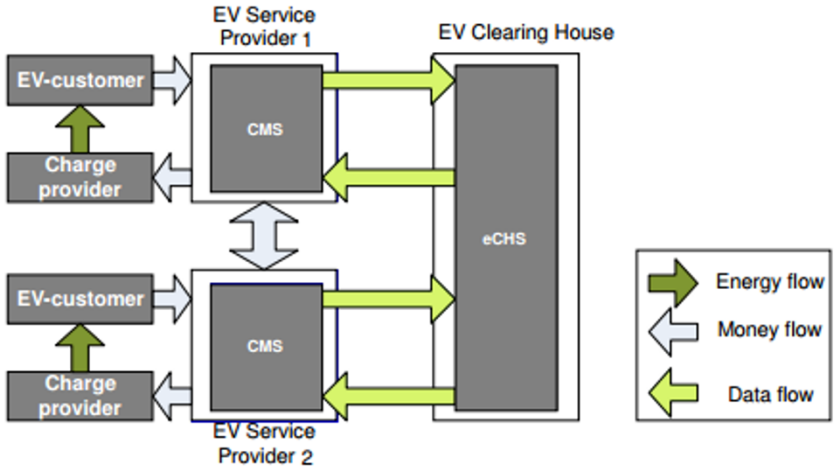


Fig. 3. OCHP in EV smart grids [4].

In order to emphasize the basic idea of the protocol, two different EV Service Providers were selected, although there can be more providers connected to a clearing house. Basically, the protocol enables customers to charge their EVs to a different provider charging station, with whom does not have any kind of

contract. In other words, with the addition of EV Clearing House, a many-to-many EV Service Provider bilateral relationships is converted into a one-to-many between EV Clearing House and EV Service Provider. The CMS abbreviation is referring to Chargepoint Management System and eCHS is referring to European Clearing House System. In comparison with the previous two protocols, this one has a macro view of the system, not being useful for single providers, but create a connection between them [4].

2.4 The evolution of OCPP

The earlier versions of OCPP (1.2/1.5) were exclusively based on SOAP/XML data interchange. The main disadvantage of this approach is the significant overhead in the communication layer. This data interchange could be advantageous in communication without data transmission limit. But regarding the fact that the vast majority of stations are using GPRS for data connection, this overhead could make a difference in terms of costs. In this case, the large data volumes should be as limited as possible. A protocol based on JSON requires less data transmission and it is more suitable in this case [5].

The next version tried to resolve this issue with large data transmissions by using JSON instead of SOAP/XML. The JSON version of the previous protocol has the J extension, like OCPP2.0J. For SOAP/XML is usually used S extension or even no extension at all.

Another important aspect of J version is that is also supporting WebSockets using JSON. WebSockets technology provides a bidirectional communication with a remote host and it is also supported by browsers, because it starts like as an HTTP Connection but it is converted to WebSocket using a switch, known as WebSocket handshake.

Although the problem with data transmission is resolved by adding the alternatives of what type of data is used, there are other different problems unresolved by this path, like device management, pricing and smart charging. Any of these features are successfully added to version 2.0 [5].

2.5 Advantages and disadvantages

Over the time, OCPP requirements and cases were developed, including many situations that should be standardized, for example in the case of no Internet connection. This case with authorization lists of recent users and many others similar cases were treated in latest version of the protocol [7]. Another advantage of this protocol is that from a version to another there are substantial additions, but with no impact on usage of the previous versions. In this case, a central system can check the version implemented on stations and act correspondingly and vice versa [10].

The main focus of the previous version was only on SOAP-over-HTPP. With version 1.5 the REST is also available, but not as mature as the first one. The main disadvantage of the SOAP-over-HTPP is the huge message overhead. The

drawbacks are not visible only on responsiveness of the system but rather on expensive maintenance costs from providers [6].

Another issue regarding OCPP protocol is related to its maturity. The requirements are most of the time vague and interpretable in terms of handling errors and security. Another aspect is related to testing, which is not provided in any way by the standard. Indeed, there are various implementations of test cases, but none of them are indicated by alliance [6].

3 The implementation of OCPP

3.1 Similar projects

One of the open source implementations of OCPP protocol is the project named Motown - MObility Transition Open source Webservices iNterface. In the implementation of this solution, two main architectural approaches are used: hexagonal architecture and Command Query Responsibility Segregation (CQRS) pattern.

The first one, hexagonal architecture, also known as *ports and adapters* means that there is a core application without any dependencies on the user interface, database, or other inputs and outputs and several add-ons. This is very important regarding the fact that there are more and more new features of OCPP protocol from one version to another and keeping the system flexible is very important. Besides new features, there are many *flavours* added by each vendor in order to provide the best experience to the users, which leads to a harder work in the implementation of a central server which communicates with various types of charging stations.

The second decision, CQRS is a relatively simple architectural pattern which separates the updating the information to the retrieving the information. CQRS is commonly used in complex domains and it fits very well with the previous hexagonal approach. The update of the information is done using commands to the domain. As a consequence to the commands, an event that describes the required changes to that domain is triggered [8].

3.2 The architecture

The components of the systems are:

- Database - in which the data is stored
- Business Logic Controller (BLC)- processes the requests from the WebSocket Interface Controller and updates the database
- WebSocket Interface Controller (WIC) - a bridge between the charging stations and administrator which also calls interface of BLC
- Charging station
- Administrator - control the charging station through WIC

The main components of the system are Business Logic Controller and WebSocket Interface Controller.

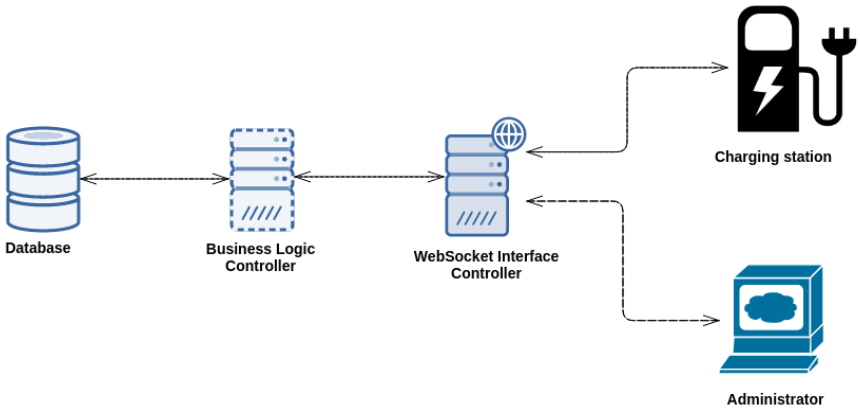


Fig. 4. The architecture of the proposed solution.

3.2.1 Business Logic Controller - BLC The purpose of this component is to update and retrieve information in the database based on the requests from the WIC. This is a separate module from the WebSocket Interface Controller because in this way the system remains flexible, even though other interface controllers (like SOAP over HTTP) are added. In the same case of adding new interface controllers, we need to make sure that only one module has access to the database, for consistency. This module provides a common interface for all other controllers.

3.2.2 WebSocket Interface Controller - WIC WebSocket Interface Controller (WIC) is the middleware component which receives requests from charging points and administrator and calls the BLC interface. The communication between the administrator and charging points is never done directly and only via WIC. Every time a charging point is making a request to the server, the administrator also receives a copy of that message. WIC manages the open connections, having two sets of connections: one for charging stations and one for administrators. This module also validates the messages, against JSON schemas. Open Charge Alliance provides two schemas for every pair of request/response, in this way the interface is the same regardless of implementation. A similar approach is done in SOAP case, where WSDL files are used.

4 Experimental methodology

The purpose of this measurements is to demonstrate the efficiency of REST over WekSocket in comparison with REST over HTTP, in terms of traffic. For this purpose, there were chosen only the most frequent requests initiated by the charging stations for 24 hours of a weekday. Let be:

$$M = \{Authorize, Heartbeat, StarTransaction, MeterValues, StopTransaction\} \quad (1)$$

The rest of the requests are not frequent and are determined by errors, reset or update firmware. For a better understanding of testing, hardware and software specifications are also presented in this section.

4.1 REST over WebSocket

WebSocket is a new feature of HTML5 and it is defined as a technology that enables full-duplex communication with a remote host. Besides the full-duplex communication, it is added just a minimal overhead in comparison with Ajax polling and Comet solutions that are used to support a full-duplex communication by maintaining two HTTP connections. These two aspects make WebSocket a feasible solution for building real-time and scalable communication systems.

WebSocket protocol starts with a handshake between the client and the server by which HTTP protocol is updated. However, it uses the same ports as HTTP (80 and 443). The implementation of this protocol is facile on both sides. On the front-end, it is supported by almost every browsers and on the back-end on almost all application servers. Even web servers, like Apache Tomcat, supports it (from version 8) [9].

In most of the cases, the charging stations do not have Internet connection through WI-FI connection, but GPRS. The traffic is vital because is directly related to the cost of the maintenance. On the other hand, about half of the requests are initiated by the central server, which means that we need an efficient full-duplex communication between charging stations and central systems, WebSocket being the best solution in this case.

The format of the messages is also important. REST provides a simple and highly efficient structure of messages. On the other hand, the SOAP messages are bigger and the traffic is also increased. Regarding the constraints of network traffic, REST is a far better option than SOAP. In conclusion, REST over WebSocket is the best combination for OCPP implementation in terms of efficiency, traffic and costs.

4.2 Assumptions

The tests simulated a normal week day, on public charging stations. We assumed that the period interval that charging station is used is: 08:00 AM - 10:00 PM (14 hours). Another assumption is that the type of the charging station is a fast charge, meaning that a normal car stays at charging no more than 20-30 minutes. In order to measure the traffic in the network, we need to know the frequency of the requests. As we previously assumed, there are 20-30 minutes for each charging, meaning that there are 28 chargings in a day. The numbers of chargings is the same with the numbers called methods (pairs of request/response): Authorize, StartTransaction and StopTransaction.

MeterValues request is sent by the charging station in order to let the central server (CS) know about the progress of the charging. If this kind of request was not implemented, the server can lose the data for an entire charging if the Internet connection is lost or other technical problems occur. The frequency of this request is set in the configuration of each charging station, but in most of the cases is 300 seconds from the beginning of a transaction till the end of it. If we assume that the average time for a charging is 25 minutes (in the rest of 5 minutes card authorization and physical connection/disconnection of the electric car from the station are done), that means that there are 4 MeterValues requests sent for each charging. A similar purpose has Heartbeat request, by which the charging station announces the CS that the connection is up and running. The default value for this request is 300 seconds.

There are other 2 values that can determine the efficiency of an approach over another. Firstly, it is the Keep-Alive messages sent every 45 seconds in the WebSocket transmission in order to check if the TCP connection is up and running. Secondly, the connection timeout for TCP in HTTP approach. Regarding the fact that the most frequent requests (Heartbeat and MeterValues) are sent every 300 seconds and the TCP connection timeout is about 30 seconds, we can assume that the are no requests that share the same TCP connection, meaning that every HTTP request in the REST over HTTP approach will start with three-way handshake ($3WH_{TCP}$) and will end with four-way handshake($4WH_{TCP}$).

4.3 Hardware and software specifications

The two approaches were simulated using a Wi-Fi connection 802.11b/g/n (2.4 GHz), with the latency of 50ms. The operating systems for CS and simulated CP are Ubuntu 15.04 Linux Mint 17.1, respectively. The application server used is GlassFish Server Open Source Edition 4.0. The source code that runs on the server side is written in Java and the client in Javascript.

5 Experimental results and performance analysis

Any request in the REST over HTTP approach is preceded by a TCP three-way handshake - 214 bytes and, if the there are no other requests on the connection, it is succeeded by a four-way handshake - 264 bytes.

In addition to these, there are the HTTP requests/responses: request generated by the CP (CP_req) and response from the CP (CS_resp). As we can see in the table 1 the response is split into 2 different messages with its own ACK - 66 bytes. In the table 1, a Heartbeat request is sent to the CS using REST over HTTP. The traffic of a single OCPP request over HTTP is:

$$Traffic_{HTTP}(t) = 3WH_{TCP} + CP_req(t) + CS_resp(t) + 3 \times ACK + 4WH_{TCP} \times No(t) \quad (2)$$

Table 1. Table of times and lengths in REST over HTTP.

No.	Source	Destination	Protocol	Length (bytes)	Info
1	CP	CS	HTTP	582	POST
2	CS	CP	TCP	66	[ACK]
3	CS	CP	TCP	354	[TCP segment of a reassembled PDU]
4	CP	CS	TCP	66	[ACK]
5	CS	CP	HTTP	71	HTTP/1.1 200 OK
6	CP	CS	TCP	66	[ACK]

Table 2. HTTP requests/responses containing OCPP message.

Request/Response	CP_req (bytes)	CS_resp (bytes)
Heartbeat	582	425
Authorize	606	481
StartTransaction	714	499
MeterValues	1314	499
StopTransaction	1060	480

In the table 2 are shown the lengths of each HTTP requests/responses containg OCPP message. As a result, the total traffic generated by the REST over HTTP approach is:

$$Total_{HTTP} = \sum_{t \in M} Traffic_{HTTP}(t) \approx 902KB. \tag{3}$$

In comparison with HTTP approach, REST over WebSocket maintains only one TCP connection open, all the requests being sent through the same connection, assuming that no errors, that could end the connection, occur. In this approach, the traffic generated by each method is composed of three messages: request sent by the CP, response from the CS and the ACK from the CP. The network traffic generated by each pair of request/response is composed of:

$$Traffic_{WS}(t) = CP_req(t) + CS_resp(t) + 3 \times ACK \tag{4}$$

Table 3. WebSocket requests/responses containing OCPP message.

Request/Response	CP_req (bytes)	CS_respons (bytes)
Heartbeat	128	147
Authorize	154	180
StartTransaction	236	220
MeterValues	597	111
StopTransaction	701	201

The connection is opened only once, the messages sent in order to establish it are: TCP three-way handshake (214 bytes), an HTTP request (689 bytes) and response (465 bytes) for upgrading protocol from HTTP to WebSocket and the Connection established WebSocket message (126) sent by CS. Each of these messages is accompanied by an ACK (66 bytes). In a similar way, the WebSocket connection is closed. There is WebSocket request sent by CS in order to close the WebSocket connection and finally, a four-way handshake (264 bytes). In addition to this traffic, keep-alive messages are added. As a result, the traffic generated by REST over WebSocket is:

$$Total_{WS} = 3 \times WH_{TCP} + Open_{WS} + \sum_{t \in M} Traffic_{WS}(t) + Close_{WS} + 4 \times WH_{TCP} + KeepAlive_{TCP} \approx 253KB \quad (5)$$

where $KeepAlive_{TCP}$ is the traffic generated by Keep-Alive TCP messages.

As a result from the measurements of the two approaches, we can say that REST over WebSocket is 3.5 times more efficient in terms of traffic than REST over HTTP. The main reason is the fact that there is an important overhead in opening and closing a new TCP connection. If the frequency of the requests would be higher, the TCP connection could be reused, but using default values, this is very unlikely. On the other hand, with an increased numbers of requests, the frequency of Keep-alive messages used in WebSocket communication would be drastically decreased, because there are used when no other messages are exchanged in more than 45 seconds.

If the requests would be initiated by both sides, the difference would be even bigger, because the HTTP approach requires 2 TCP connections with the same traffic on each of them and only one TCP connection in WebSocket approach.

6 Future work

The main goal of the future work is to integrate the SOAP over HTTP interface to the existent system. Even though, as we demonstrated above, REST over WebSocket is a better alternative, there are many vendors who have chosen implement only SOAP over HTTP version on their charging stations. Because the goal is a centralized system for charging vehicles in smart grids, the integration of those stations is vital.In addition, security and load-balancing will be treated. First one is important because the data has an economic impact and the second one is also important due to the increasing numbers of charging stations and, as a result, in the number of requests.

7 Conclusion

The smart charging technology is a trending topic in the research field, raising important problems in the near future. With the increasing share of electric

vehicles and renewable energy sources, the existence of a flexible grid in this area becomes necessary. In this paper was proposed an architecture which can provide efficiency, flexibility to new add-ons, even though they are in terms of new features, interfaces, protocols. However, due to the flexibility of the system, it can be easily adjusted for other scenarios. Because OCPP requires a full-duplex communication, WebSocket subprotocol has been chosen and because the size of the messages is also important JSON format was the best solution. The software solution was divided into several modules, each responsible for dedicated parts of the communication and business logic.

Acknowledgment

This work was supported by a grant of the Romanian National Authority for Scientific Research and Innovation, CNCS - UEFISCDI, project number PN-II-RU-TE-2014-4-2731 - *DataWay*: Real-time Data Processing Platform for Smart Cities: Making sense of Big Data.

References

1. Bogdan Ovidiu Varga. "Electric vehicles, primary energy sources and CO_2 emissions: Romanian case study." Energy 49 (2013): 61-70.
2. Fiona Williams, Matthias Sund, Alexander von Jagwitz, "Electric mobility functional ICT Architecture Description", 12 May 2013, FI.ICT-2011-285135 FINSENY
3. Open Charge Alliance, "Open Smart Charging Protocol 1.0 - Interface description between DSO and Central System", 4 September 2015
4. Peter Buycks, Arjan Wargers, Raymund Rutten, "Open Clearing House Protocol (OCHP)", 21 May 2012
5. Open Charge Alliance, "Open Charge Point Protocol - Interface description between Charge Point and Central System", 3 November 2014
6. Johannes Schmutzler, Claus Amtrup Andersen, and Christian Wietfeld. "Evaluation of OCPP and IEC 61850 for smart charging electric vehicles." In Electric Vehicle Symposium and Exhibition (EVS27), 2013 World, pp. 1-12. IEEE, 2013.
7. Marc Mutin, Christian Gitte, and Hartmut Schmeck. "Smart Grid-Ready Communication Protocols And Services For A Customer-Friendly Electromobility Experience." In GI-Jahrestagung, pp. 1470-1484. 2013.
8. Motown official page, `https://github.com/motown-io/motown/wiki/developers-guide`, Accessed on May 2016.
9. Qigang Liu, Xiangyang Sun, "Research of Web Real-Time Communication Based on Web Socket", 2012.
10. Open Charge Alliance, "Background Open Charge Alliance", June 2015

SR-KNN: An Real-time Approach of Processing *k*-NN Queries over Moving Objects

Ziqiang Yu, Yuehui Chen, Kun Ma*

Abstract Central to many location-based service applications is the task of processing *k*-nearest neighbor (*k*-NN) queries over moving objects. Many existing approaches adapt different index structures and design various search algorithms to deal with this problem. In these works, tree-based indexes and grid index are mainly utilized to maintain a large volume of moving objects and improve the performance of search algorithms. In fact, tree-based indexes and grid index have their own flaws for supporting processing *k*-NN queries over an ocean of moving objects. A tree-based index (such as R-tree) needs to constantly maintain the relationship between nodes with objects continuously moving, which usually causes a high maintenance cost. Grid index is although widely used to support *k*-NN queries over moving objects, but the approaches based on grid index almost require an uncertain number of iterative calculations, which makes the performance of these approaches be not predictable.

To address this problem, we present a dynamic Strip-Rectangle Index (SRI), which can reach a good balance of the maintenance cost and the performance of supporting *k*-NN queries over moving objects. SRI supplies two different index granularities that makes it better adapt to handle different data distributions than existing index structures. Based on SRI, we propose a search algorithm called SR-KNN that can rapidly calculate a final region space with a filter-and-refine strategy to enhance the efficiency of process *k*-NN queries, rather than iteratively enlarging the search space like the approaches based on grid index. Finally, we conduct experiments to fully evaluate the performance of our proposal.

Ziqiang Yu
The university of Ji'nan, Shandong province, China, 250022, e-mail: ise_yuzq@ujn.edu.cn

Yuehui Chen
The university of Ji'nan, Shandong province, China, 250022, e-mail: yhchen@ujn.edu.cn

Kui Ma
The university of Ji'nan, Shandong province, China, 250022, e-mail: ise_mak@ujn.edu.cn

* Corresponding author

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_18

1 Introduction

Processing k nearest neighbor (k-NN) queries over moving objects is a fundamental operation in many location-based service applications. For example, a location-based social networking service may help a user find k other users that are closest to him/her. Some taxi-hailing applications such as UBER need monitoring the nearby taxies and users for a user or a taxi with submitting requests to applications respectively. In location-based advertising, a store may want to broadcast promotion messages only to the potential customers that are currently closest to the store. Such needs can be formulated as k-NN queries, where each user, taxi, or customer can be regarded as a moving object.

Consider a set of N_p moving objects in a two dimensional region of interest. An object o can be represented by a quadruple $\{id_o, t, (o_x, o_y), (o'_x, o'_y)\}$, where id_o is the identifier of the object, and t is the current time; (o_x, o_y) and (o'_x, o'_y) represent the current and previous positions of o respectively. The old location (o'_x, o'_y) can help us remove the obsolete positions of moving objects. In this study, we adopt the snapshot semantics because we make no assumption on the motion of objects, which conforms to the situation of moving objects in reality. The snapshot semantics refer to the answer of query q at time t is only valid for the past snapshot of the objects and q at time $t - \Delta t$, where Δt is the delay due to query processing. Since this study aims to processing k-NN queries over moving objects in real-time, we thus need to reduce the value of Δt as small as possible. In order to achieve this purpose, we focus on main-memory-based solutions.

The problem of k-NN query processing over moving objects has attracted considerable attentions in recent years. The existing works about this problem can be broadly classified into tree-based approaches and grid-based approaches. Tree-based approaches refer to the works [1] that adopt tree index structures (such as R-tree, B^+-tree) to process the k-NN queries on spatial-temporal data. Here, the R-tree [2] is a data structure used for indexing multi-dimensional information such as geographical coordinates, rectangles or polygons and it has been adopted to answer k-NN queries in some works. The general idea of these works is first searching the nearest neighbor for a given query, and then determining k nearest neighbors of this query. R-tree, as a spatial data index, well adapts to index the static spatial data, but it is not suitable for the maintenance of continuous moving objects. This is because the nodes of R-tree has to be split or merged frequently with objects constantly moving, and even the R-tree requires to be organized repeatedly. Therefore, indexing a large scale of moving objects with the R-tree will cause a large maintenance cost .

Grid index is a typical spatial index that partitions the whole search region (2-D surface in this study) of interest into equal-sized cells, and indexes objects and/or queries (in the case of continuous query answering) in each cell respectively. Extensive existing works propose various search algorithms based on the grid index to handle k-NN queries over moving objects. Most existing grid-based approaches to k-NN search [3, 4, 5] iteratively enlarge the search region to identify the k-NNs. For example, given a new query q, the YPK-CNN algorithm [3] initially locates a rectangle R_0 centered at the cell covering q; it then iteratively enlarges R_0 until it

encloses at least k objects. Let p' be the farthest object to query q in R_0. The circle C_r centered at q with radius $\| q - p' \|$ is guaranteed to contain the k-NNs of q, where $\| \cdot \|$ is the Euclidean norm. The algorithms then computes the k-NNs using the objects in the cells intersected with C_r. The other existing grid-based approaches are based on similar ideas. These approaches have a common defect that they need to iteratively search k objects and the number of iterations is unpredictable. In some cases, these approaches probably cause extensive search iterations especially on the data in non-uniformly distribution, which will degrade their performance.

To address this challenge, we propose a dynamic Strip-Rectangle index called SRI to support the processing of k-NN queries over moving objects. SRI is a two level index structure . The first level index of SRI is the strips that partition the whole region of interest. All strips can cover the whole search region and no overlap exists between any two different ones. Further, each strip is divided into smaller rectangles that form the second level index. SRI can dynamically adjust the sizes of strips and rectangles based on the distribution of moving objects to make each strip and rectangle guarantee covering at least ξ objects. This characteristic makes SRI better adapt to support the processing of k-NN queries and handle various distributions of spatial data. Based on SRI, we design an algorithm called SR-KNN to handle k-NN queries over moving objects without iterations occurred in grid-based approaches. For a given query q, SR-KNN adopts a filter-and-refine strategy to rapidly calculate a small search region that covers k neighbors of q, and then obtain k-NNs from this search region. Our contributions can be summarized as follows.

- We propose SRI, a strip and rectangle combined index structure that can well support the processing of k-NN queries over a large scale of moving objects in different distributions.
- Based on SRI, we design the SR-KNN algorithm that can improve the efficiency of processing spatial k-NN queries by avoiding the unpredictable iterative calculations, which solves the major flaw of existing grid-based algorithms.
- Extensive experiments are conducted to sufficiently evaluate the performance of our proposal.

2 Related work

The problem of k-NN query processing over moving objects has attracted considerable attentions in recent years. In this section, we present a brief overview of the literature.

The R-tree has been adopted extensively (e.g., [1, 6, 7, 8]) to answer nearest neighbor queries. Ni et al. [9], Roussopoulos et al. [10], and Chaudhuri et al. [11] use the TPR-tree to index moving objects and propose filter-and-refine algorithms to find the k-NNs. Gedik et al. [12] describe a motion-adaptive indexing scheme based on the R-tree index to decrease the cost of update in processing k-NN queries. Yu et al. [13] first partition the spatial data and define a reference point in each partition,

and then index the distance of each object to the reference point employing the B^+-tree structure to support k-NN queries.

Grid index is widely used to process spatial queries [3, 14, 15, 16, 17, 18, 19]. Zheng et al. propose a grid-partition index for NN search in a wireless broadcast environment [14]. The Broadcast Grid Index (BGI) method proposed by [15] is suitable for both snapshot and continuous queries in a wireless broadcast environment. Šidlauskas et al. [18] propose PGrid, a main-memory index consisting of a grid index and a hash table to concurrently deal with updates and range queries. Wang et al. [19] present a dual-index, which utilizes an on-disk R-tree to store the network connectivities and an in-memory grid structure to maintain moving object position updates.

3 The SRI structure

In building SRI, the region of interest R in an Euclidean space (normalized to the $[0, 1)$ square) is first partitioned into non-overlapping strips. In this study, we make the partition be done along the x axe, thus the whole region is divided into multiple vertical strips. For each strip, SRI further divides it into smaller rectangles without overlap and this partition is done along the y axe. An example of SRI structure is shown in Fig. 1.

A strip S_i ($1 \leq i \leq N_v$, where N_v is the number of strips) in SRI takes the form of $\{id_i, lb_i, ub_i, p_i, \Lambda_i\}$, where id_i is the unique identifier of S_i, lb_i and ub_i are the lower and upper boundaries of the strip respectively, p_i is the number of moving objects in this strip, and Λ_i is a list of identifiers of rectangles contained in this strip. Similarly, a rectangle R_j covered by S_i can be represented as $\{rid_j, b_j, u_j, \Gamma_j\}$, where rid_j is the identifier of the rectangle, b_i and u_i are the lower and upper boundaries of R_j respectively, and Γ_j is a list of objects in R_j. Since the strips and rectangles are both non-overlapping and every object must fall in one strip and one rectangle, we can deduce $\sum_{i=1}^{n} p_i = Np$ and no object belongs to two different strips, where N_p is the total number of moving objects. For any strip S_k with m rectangles, we can infer $\sum_{j=1}^{m} |\Gamma_j| = p_k$, and $\nexists o_t \in \Gamma_s \cap \Gamma_t$ ($s \neq t$). Fig. 1 describes attributes of index elements in SRI.

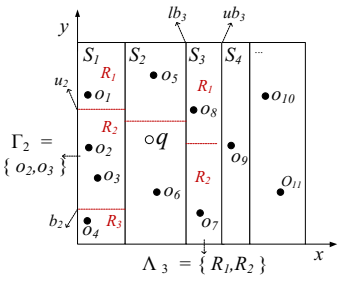


Fig. 1 The structure of SRI

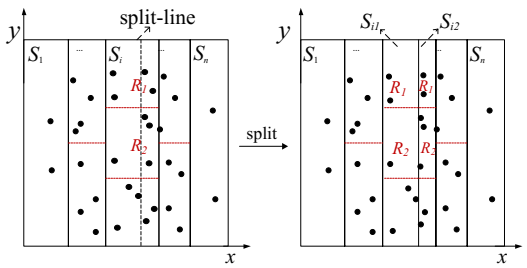


Fig. 2 The split of S_i

In SRI, we require every strip to contain at least ξ and at most θ objects, i.e., $\xi \leq p_i \leq \theta$ for all strips S_i. The strips are split or merged as needed to ensure this condition is met when object locations are updated. We call ξ and θ the *minimum occupancy* and *maximum capacity* of a strip respectively. Typically $\xi << \theta$. Be similar with the strip, every rectangle R_j is also required to contain at least ξ and at most β objects ($\xi < \Gamma_j < \beta$), where ξ and β are the *minimum occupancy* and *maximum capacity* of a single rectangle. Each rectangle also needs to be spit or merged with its adjacent rectangle within the same strip to ensure the volume of its moving objects belongs to $[\xi, \beta]$. In this study, a strip will be divided into multiple rectangles to form the index with a smaller granularity, thus the value of β is specified smaller than that of θ.

So far, the structure of SRI is more clear. Strips form the first level index and they are sorted in ascending order according to their boundaries. The rectangles of each strip construct the second level index and they are also maintained in order based on their boundaries within every strip. SRI has two level indexes, but it only needs to store one piece of all moving objects. This is because every strip does not store the locations of its moving objects but just record their quantity, which not only reduces the memory and maintenance costs but also is critical for designing k-NN algorithms. The benefits of SRI will be discussed later.

3.1 Insertion

When an object o_i sends the message $\{id_o, t, (o_x, o_y), (o'_x, o'_y)\}$ to the server, we need to insert the new position (o_x, o_y) of o_i and delete its old location (o'_x, o'_y).

Object o_i is inserted into SRI with two steps: (1) determining the strip S_i that o_i falls into its boundaries ($lb_i \leq o_x < ub_i$) and modifying the value of p_i ; (2) searching the rectangle R_j that satisfies $b_j \leq o_y < u_j$ from strip S_i and then inserting o_i into the rectangle R_j. The insertion is done by appending its id_o into the object list Γ_j. Initially, there is only one strip covering the whole region of interest and the strip itself is a rectangle.

When an object o_i is inserted into a rectangle R_j in the strip S_i, which probably causes two types of splits: rectangle spit and strip split. After o_i being inserted into rectangle R_j, it will be split if the volume of objects in it exceeds the maximum capacity, i.e., $|\Gamma_j| > \beta$. A split method that can adapt to the data distribution is to split R_j and generate two new ones that hold approximately the same number of objects. In this method, we first find an object o such that o_y is the median of the y coordinates of all objects in rectangle R_j, which implies that there are approximately $|\Gamma_i|/2$ objects whose y coordinates are less than or equal to o_y. This can be accomplished in $O(|\Gamma_i|)$ time using the binning algorithm. Next, we set the line $y = o_y$ as the split-line, according to which R_j can be split into two new strips R_{i1} and R_{i2}. Once R_j is split, the attributes of new strips can be determined as follows. The lower boundary of R_{i1} is the same as that of S_i, and its upper boundary is the split-line. For R_{i2}, it uses the split-line as the lower boundary, and the upper boundary of R_i as its upper boundary. The *id* of R_i is inherited by S_{i1}, and a new *id* is assigned to R_{i2}.

As to strip S_i, it will also be split if the number of moving objects covered by itself is greater than θ as an object o being inserted. The split method about strip is similar with that about the rectangle and Fig. 2 gives an example of strip split. When S_i is split into S_{i1} and S_{i2}, the attributes of new strips also can be easily deduced just like determining the attributes of new rectangles above. Meanwhile, every rectangle R_k in S_i also need to be split into R_{k1} and R_{k2} by the split-line. Then the objects in R_k will be assigned into R_{k1} and R_{k2} based on the split-line. That is, $\Gamma_{k1} \cup \Gamma_{k2} = \Gamma_k$ and $\Gamma_{k1} \cap \Gamma_{k2} = \emptyset$. Since R_{k1} and R_{k2} belong to two different strips, so both of them can use the identifier and the lower and upper boundaries of R_k.

3.2 Deletion

If object o disappears or moves out of a rectangle, it has to be deleted from the rectangle that currently holds it. To delete an object o, we need to determine which rectangle current holds it, which can be done using its previous position (o'_x, o'_y). After deleting an object, if the rectangle R_j in strip S_i has less than ξ objects (i.e., R_j has an underflow), it will be merged with an adjacent rectangle in S_i. Let this adjacent strip be R_h. R_j will be deleted from Strip S_i, and the merged strip will inherit the *id* of the R_h, and its lower and upper boundaries are set to be the lower and upper boundaries of R_j and R_h, respectively. The object lists Γ_j and Γ_h are merged.

When object o moves from strip S_i to another strip or disappears, the number of moving objects in strip S_i needs to minus 1, i.e., $p_i = p_i - 1$. At this time, if p_i is smaller than the *minimum occupancy* ξ, then S_i also needs to be merged with the adjacent strip S_j with fewer moving objects. Since $p_i < \xi$, S_i contains only one rectangle. In this case, we assign the objects in S_i into the corresponding rectangles in strip S_j, and then update the boundaries of S_j to make it cover the space of S_i. The boundaries of every rectangle in the new strip S_j remain the same.

4 The SR-KNN algorithm

The SR-KNN algorithm follows a filter-and-refine paradigm. For a given k-NN query q, the algorithm first prunes the search space by identifying *candidate strips* that are guaranteed to contain at least k neighbors of q. From candidate strips, it then identifies *candidate rectangles* that also covers at least k neighbors of q, which can further narrow down the search region. Next, it examines the objects contained in these candidate rectangles and identify the k-th nearest neighbor found so far. Using the position of this neighbor as a reference point, it calculates the final region that guarantees covering k-NNs of q and obtain the final result from it. We present the pseudocode of SR-KNN in Algorithm **??**. Now we present its details.

4.1 Calculating candidate strips

For a given query q, SR-KNN can directly identify the set of strips that are guaranteed to contain k neighbors of q, which we call the candidate strips.

Step 1: Calculating the number of candidate strips. Assume that the number of candidate strips is c. The idea is that from each strip we select χ ($1 \leq \chi \leq \xi$) objects that have the shortest Euclidean distances to q, such that $\chi * c \geq k$, where χ can be specified by users. This way, we have found at least k neighbors for q. Of course, these objects may not be the final k-NNs, but they can help us prune the search space and serve as the starting points for computing the final k-NNs. Hence, the number of candidate strips c is set to be $\lceil k/\chi \rceil$.

Step 2: Identifying the set of candidate strips. In this step, we identify the strips that are considered to be "closest" to q based on their boundaries. We use d_i^l and d_i^u to denote the distances from q to the lower and upper boundaries of S_i respectively. For the example shown in Fig. 3, the line l_i is perpendicular to lb_2 of S_2 and the distance from q to lb_2 is d_2^l. The distance between S_i and q is defined to be $dist(S_i, q) = \max\{d_i^l, d_i^u\}$.

If query q is located in S_i, then S_i is automatically a candidate strip and inserted into $\mathscr{C}^V$. Next, we decide whether its neighboring strips are candidate strips. Starting from the immediately adjacent strips, we expand the scope of search, adding to $\mathscr{C}^V$ the strip j that has the next least $dist(S_j, q)$. This procedure terminates when $|\mathscr{C}^V| = c$ or all strips have been processed. Fig. 3 gives an example, in which S_3 is determined to be a candidate strip first. Then by comparing d_2^l with d_4^u, we decide S_4 to be the next candidate strip. Next, we find S_2 also a candidate strip.

4.2 Calculating candidate rectangles

For a query q (x_q, y_q), candidate rectangles refer to the rectangles that are "closest" to q and at least cover k neighbors of q. We set the candidate rectangles to q must be covered by its candidate strips.

Step 1: Determining the number of candidate rectangles. In this step, we also suppose χ ($1 < \chi < \xi$) objects are chosen from each rectangle, then the number of candidate rectangles is $\lceil k/\chi \rceil$.

Step 2: Identifying the set of candidate rectangles. In SRI, there exists a center in every rectangle that has equal distance to the lower and upper boundaries. We adopt the distance between the center of a rectangle to q as the metric to identify the rectangles that are "closest" to q. Since the centers of all rectangles in a strip have the same x coordinate, for any two rectangles in the same strip, we can immediately infer which rectangle is closer to q only based on their y coordinates. Hence, we can rapidly identify the closest rectangle to q within every strip without computing distances between their centers and q, which will reduce extensive calculation costs.

To identify candidate rectangles, we use two sets $\mathscr{R}_c$ and $\mathscr{R}_t$ to separately store candidate rectangles and intermediate results. First, we find the closest rectangle to q in each candidate strip and put these rectangles into $\mathscr{R}_t$. Next, we choose the

closest rectangle R_f to q from $\mathscr{R}_t$ and add R_f into $\mathscr{R}_c$. Third, we put a rectangle R_s into $\mathscr{R}_t$, where R_s and R_f belong to the same strip S_i and R_s is closer to q than other rectangles in S_i except for R_f. We execute the second and third steps repeatedly until $\mathscr{R}_c$ contains $\lceil k/\chi \rceil$ rectangles or $\mathscr{R}_t$ is empty. Finally, $\lceil k/\chi \rceil$ candidate rectangles can be contained by $\mathscr{R}_c$. Fig. 3 shows an instance of identifying candidate rectangles and blue points are centers of rectangles.

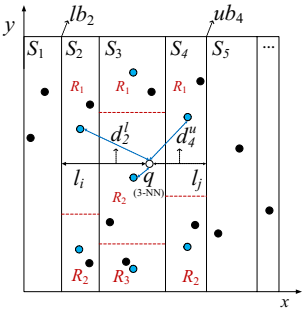


Fig. 3 Identifying candidate strips and rectangles

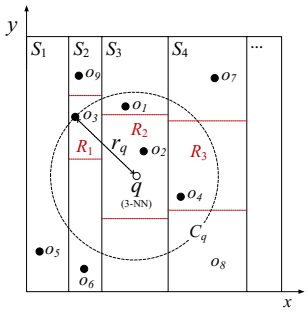


Fig. 4 Procedure of processing query q

4.3 Determining the final search region

After candidate rectangles being determined, we form the set of *supporting objects* Υ by selecting from each candidate rectangle χ objects that are closest to q. We then identify the supporting object $o \in \Upsilon$ that is the k-th closest to q. Let the distance between o and q be r_q. The circle with (q_x, q_y) as the center and r_q as the radius is thus guaranteed to cover k-NNs of q. Next, we identify the set of rectangles $\mathscr{F}$ that intersect with this circle, and search for the final k-NNs within the objects in $\mathscr{F}$.

Fig. 4 shows an example, where the query q is a 3-NN query and $\chi = 1$. We first identify the candidate strips $\{S_2, S_3, S_4\}$ as well as the candidate rectangles $\{R_1, R_2, R_3\}$, and then find three closest supporting objects (o_3, o_2, o_4) from candidate rectangles. Next, we set the radius r_q to be the distance between q and o_3 and the circle C_q is guaranteed to contain the 3-NNs of q. After scanning all objects that are located within C_q (by examining all rectangles intersecting C_q), we find that the 3-NNs are o_1, o_2, and o_4.

4.3.1 Advantages of SR-KNN

- **Powerful pruning strategy**: According to SRI index, SR-KNN can quickly narrow down the search space that covers the final results of queries by two pruning steps. It first identifies candidate strips to locate a much smaller search region that covers k neighbors of the query and further prunes the search region by calculating candidate rectangles, which significantly enhance the search performance.

- **Low costs of calculating candidate rectangles**: In one strip, SR-KNN can rapidly infer which rectangle is closest to q without computing the distances from the center of each rectangle to q, which effectively reduces expenses of calculating candidate rectangles and improves the whole efficiency of SR-KNN algorithm.
- **Avoiding multiple iterations**: Be different with grid-based algorithms, SR-KNN utilizes a cascading pruning strategy to narrow down the search space instead of iteratively enlarging the search space. Regardless of uniform or non-uniform data distribution, SR-KNN always can rapidly locate a small final search region with only three steps, which makes it be superior to grid-based algorithms.

5 Experiments

The experiments are conducted on a server with a 2.4GHz Intel processor and 8GB of RAM. We use the German road network to simulate three different datasets for our experiments. In these datasets, all objects appear on the roads only. In the first dataset (UD), the objects follow a uniform distribution. In the second dataset (GD1), 70% of the objects follow the Gaussian distribution, and the other objects are uniformly distributed. The third dataset (GD2) also has 70% of the objects following the Gaussian distribution, but they are more concentrated. In all four datasets, the whole area is normalized to a unit square, and the objects move along the road network, with the velocity uniformly distributed in [0, 0.002] unless otherwise specified.

5.1 Performance of index construction and maintenance

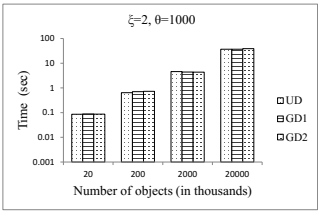


Fig. 5 Computation time for building SRI

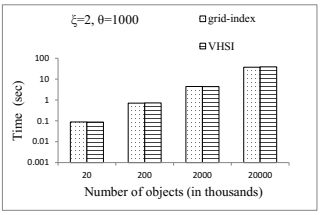


Fig. 6 Comparison of SRI and grid-index w.r.t building time

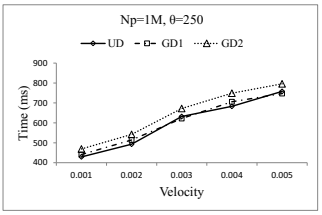


Fig. 7 Maintenance cost w.r.t velocity

Time of building SRI. We first test the time of building SRI from scratch to index different numbers of objects. Fig. 5 shows the time of building SRI as we vary the number of objects. In our study, the size of each object is approximately 50B, so we at most handle 1GB of data in this experiment. The time it takes to build the index increases almost linearly with the increasing number of objects.

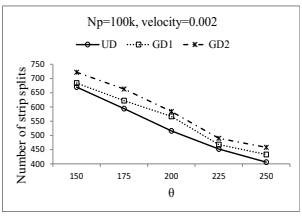


Fig. 8 Number of strip splits w.r.t. θ

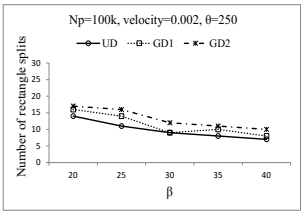


Fig. 9 Number of rectangle splits w.r.t. θ

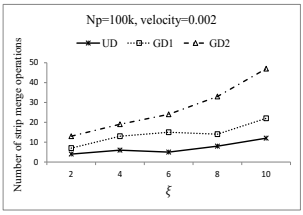


Fig. 10 Number of strip merge operations w.r.t. ξ

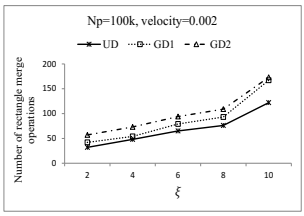


Fig. 11 Number of rectangle merge operations w.r.t. ξ

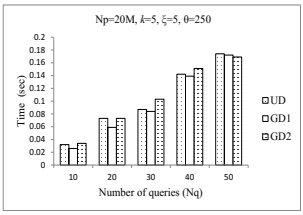


Fig. 12 Performance of RS-KNN w.r.t. number of queries

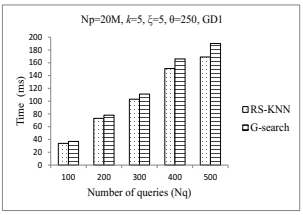


Fig. 13 Comparison of RS-KNN and G-search based on GD1

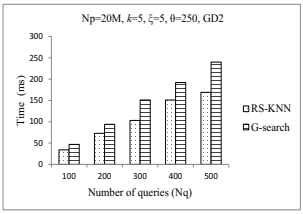


Fig. 14 Comparison of RS-KNN and G-search based on GD2

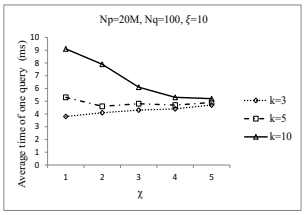


Fig. 15 Performance of RS-KNN w.r.t. ξ

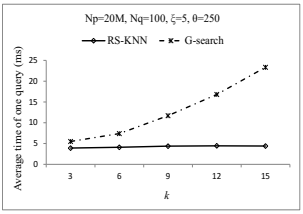


Fig. 16 Query evaluation time w.r.t. k

Comparison of building times. We then t test the time of building SRI and grid-index to index different numbers of objects based on the GD2 dataset. Fig. 6 demonstrates that SRI and grid-index need almost equal building time to handle the moving objects, which certifies the good performance of SRI with respect to indexing objects.

Effect of the velocity of objects. Fig. 7 demonstrates the effect of the velocity of objects on the computation time for maintaining SRI based on three datasets. In this set of experiments, we first build the SRI for 1M objects, and then 100K object are chosen to move continuously with varying velocities. As expected, the faster the objects move, the more split and merge operations happen, leading to an increase in maintenance time.

Effect of θ on SRI. Fig. 8 shows the effect of the maximum capacity of strip, θ, on the frequency of strip split. The number of moving objects indexed is 100K. As can be observed from Fig. 8, the strip split frequency is approximately reversely

proportional to the value of θ; a greater θ value would result in a reduction in the number of splits. Of course, θ cannot be overly large, because that will increase the time for processing queries.

Effect of β on SRI. Fig. 9 demonstrates the effect of the maximum capacity of rectangle on the frequency of rectangle split. This group of experiments also index 100k moving objects and test the average number of rectangle splits in every strip. The results show that the rectangle split frequency is also approximately reversely proportional to the value of θ; a greater β value would reduce the number of splits. Similarly, the large value of β will increase the time of processing queries, so β cannot be set overly large.

Effect of ξ on RSI. Fig. 10 and Fig. 11 show the influence of the minimum occupancy, ξ, on the frequency of strip and rectangle merge operations. A larger value of ξ means that underflow will occur more often and thus cause more strip and rectangle merge operations. Additionally, the number of rectangle merge operations is greater than that of strip merge operations.

5.2 Performance of query processing

We now perform experiments to evaluate the performance of SR-KNN, and make a comparison of SR-KNN and G-search algorithms.

Processing time. We feed a batch of queries into our system in one shot, and measure the time between the first query entering the system and the *k*-NN results of all queries having been obtained. As can be observed from Fig. 12, SR-KNN achieves similar performance for different distributions. This is because every strip and rectangle in SRI separately contains at most θ and β objects and typically each query only involves a few strips and rectangles. Therefore, the data distribution only has a slight impact on the query processing time.

In Fig. 13 and Fig. 14 , we make SR-KNN and G-search algorithms process the same batch of queries and measure their processing times with varying number of queries based on the GD1 and GD2 datasets. The results show that SR-KNN outperforms G-search on both datasets but more significantly on GD2, the reason is that G-search needs more iterative calculations to process *k*-NN queries on the non-uniformly dataset.

Effect of χ on SR-KNN. In SR-KNN, we select χ objects from each candidate strip to form the set of supporting objects. Fig. 15 shows the influence of χ on the cost of processing queries. In this set of experiments, we feed 100 queries into the system and record the average processing time of a query. The results show that χ has little effect on the processing time when *k* takes smaller values (3 and 5). But *k* increases, the influence becomes more obvious.

Effect of *k*. Finally, we study the influence of *k* on the two algorithms. As shown in Fig. 16, the processing time of SR-KNN almost remains unchanged as *k* increases, the reason being that we can adjust the value of χ accordingly to accommodate the increase in processing time. When *k* increases, G-search needs more iterations to compute the results, thus its processing time increases more rapidly than RS-KNN.

6 Conclusion

The problem of processing k-NN queries over moving objects is fundamental in many applications. In this study, we propose SRI, a novel index can better support the processing of spatial k-NN queries than other indexes. Based on SRI, we design the SR-KNN algorithm that answers k-NN queries over moving objects without numerous iterative calculations occurred in grid-based approaches and achieves a good performance.

References

1. K. L. Cheung and A. W.-C. Fu, "Enhanced nearest neighbour search on the r-tree," *ACM SIGMOD Record*, vol. 27, no. 3, pp. 16–21, 1998.
2. A. Guttman, "R-trees: a dynamic index structure for spatial searching," in *SIGMOD*, 1984, pp. 47–57.
3. X. Yu, K. Pu, and N. Koudas, "Monitoring k-nearest neighbor queries over moving objects," in *ICDE*, 2005, pp. 631–642.
4. K. Mouratidis, D. Papadias, and M. Hadjieleftheriou, "Conceptual partitioning: An efficient method for continuous nearest neighbor monitoring," in *SIGMOD*, 2005, pp. 634–645.
5. M. Cheema, "CircularTrip and arctrip: Effective grid access methods for continuous spatial queries," in *DASFAA*, 2007, pp. 863–869.
6. Y. Tao, D. Papadias, and Q. Shen, "Continuous nearest neighbor search," in *VLDB*, 2002, pp. 287–298.
7. K. Mouratidis and D. Papadias, "Continuous nearest neighbor queries over sliding windows," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 6, pp. 789–803, 2007.
8. M. S. H. A.-K. Sultan Alamri, David Taniar, "Tracking moving objects using topographical indexing," *Concurrency and Computation: Practice and Experience*, 27(8): 1951-1965, 2015.
9. K. Raptopoulou, A. Papadopoulos, and Y. Manolopoulos, "Fast nearest-neighbor query processing in moving-object databases," *GeoInformatica*, vol. 7, no. 2, pp. 113–137, 2003.
10. T. Seidl and H. Kriegel, "Optimal multi-step k-nearest neighbor search," in *SIGMOD*, 1998, pp. 154–165.
11. S. Chaudhuri and L. Gravano, "Evaluating top-k selection queries," in *VLDB*, 1999, pp. 399–410.
12. B. Gedik, K. Wu, P. Yu, and L. Liu, "Processing moving queries over moving objects using motion-adaptive indexes," *Knowledge and Data Engineering*, vol. 18, no. 5, pp. 651–668, 2006.
13. C. Yu, B. Ooi, K. Tan, and H. Jagadish, "Indexing the distance: An efficient method to knn processing," in *VLDB*, 2001, pp. 421–430.
14. B. Zheng, J. Xu, W.-C. Lee, and L. Lee, "Grid-partition index: a hybrid method for nearest-neighbor queries in wireless location-based services," *The VLDB Journal*, vol. 15, no. 1, pp. 21–39, 2006.
15. K. Mouratidis, S. Bakiras, and D. Papadias, "Continuous monitoring of spatial queries in wireless broadcast environments," *IEEE Transactions on Mobile Computing*, vol. 8, no. 10, pp. 1297–1311, 2009.
16. M. F. Mokbel, X. Xiong, and W. G. Aref, "SINA: scalable incremental processing of continuous queries in spatio-temporal databases," in *SIGMOD*, 2004, pp. 623–634.
17. X. Xiong, M. F. Mokbel, and W. G. Aref, "SEA-CNN: Scalable processing of continuous k-nearest neighbor queries in spatio-temporal databases," in *ICDE*, 2005, pp. 643–654.
18. D. Šidlauskas, S. Šaltenis, and C. S. Jensen, "Parallel main-memory indexing for moving-object query and update workloads," in *SIGMOD*, 2012, pp. 37–48.
19. H. Wang and R. Zimmermann, "Snapshot location-based query processing on moving objects in road networks," in *SIGSPATIAL GIS*, 2008, pp. 50:1–50:4.

Intrusion Detection for WSN Based on Kernel Fisher Discriminant and SVM

Zhipeng Hu[1], Jing Zhang [2], Xu An Wang [3]

[1]Basic electronic technology teaching and research section, Officer's College of CAPF ,Chengdu, Sichuan Province,China
2504300689@qq.com
[2]Basic computer teaching and research section, Officer's College of CAPF,Chengdu, Sichuan Province,China
634080084@qq.com
[3]Engineering University of CAPF, Xi'an, Shaanxi Province, China
wangxazjd@163.com

Abstract. As the energy and computing ability are limited in wireless sensor networks, so almost all of the traditional network intrusion detection schemes cannot be applied. That WSN's intrusion detection based on Kernel Fisher Discriminant and SVM is brought forward. According to the principle that the classifiers' sensitivity is different when different types of data is processed, the data is assigned to Kernel Fisher Discriminant and SVM. So that Data can be processed by the corresponding optimal classifier, and detection efficiency can be raised. Theoretical analysis and simulation results show that the proposed schemes not only can detect intrusions effectively, but also lower energy consumption than others.

1 Introduction

The wireless sensor network named WSN[1] which has caused the network vulnerable to outside attacks and threated the security of network information and normal use because of the openness of its deployment area and the broadcast nature of wireless communication. Intrusion detection technology is a deep defense technology. It is judged a variety of malicious attacks in the network accordingly by predicting the network data traffic, simulating the host running state, detecting the amount of a variety of network features and then gives its response.

Existing the WSN intrusion detection algorithms[2] had been some lack of uniform evaluation criteria, based on the traditional network intrusion detection algorithm performance evaluation criteria proposed by Porras and Debar, for the characteristics if the node resource-constrained, joined the low-power characteristics and existence continuity, this paper presents a distributed sensor network intrusion detection algorithm based on kernel Fisher discriminant analysis and SVM. A lot of experiments found that an algorithm had a particularly sensitive to certain types of data. Using this feature, we processed the data collected with suitable algorithms, and

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_19

then determined whether there is the network intrusion action. During the test we encountered multiple classification, we solve the problem with Binary tree method. Experimental results showed that the accuracy and timeliness of the new algorithm we proposed has improved greatly, and the energy consumption reduced. Simultaneously, we had been studied the lifetime of the algorithm.

2 Traditional network intrusion detection system performance evaluation criteria

Evaluation Criteria of the traditional network intrusion detection system is based on three evaluation factors given by Porras and other, later Debar and other added two performance evaluation measure[3-4]again:

①Accuracy: The accuracy refers to the IDS can correctly identify the variety of invasion from kinds of invasion acts.

②Performance: The performance refers to the processing speed IDS processed source data.

③ Completeness: The completeness refers to the ability that IDS can detect all the attacks.

④Fault Tolerance: The Fault Tolerance refers to that IDS itself must be able to withstand attacks against itself, especially Denial-of-Service attacks.

⑤Timeliness: The timeliness refers to that IDS system must analyzed as quickly as possible if IDS exist the action of the system intrusion, and reported the results out. And allow the system to do some responds before the attack did not cause more harm invasion.

Compared with traditional network, the WSN have great differences in the energy and metering capability, so we evaluate the WSN intrusion detection system, not only to meet several properties of the above, it must also have the two characteristics such as the low energy consumption and the survival continuity.

① low energy consumption: Because Wireless sensor nodes were restricted by the energy in the network, the intrusion detection system must had low energy consumption characteristics, so that the detection node consumed energy as small as possible in order to ensure to detect the survival period of the node.

② Survival continuity: The division of a variety of the WSN nodes is different, some nodes need work continuously, so their energy consumption is high. So that kind of node can failure due to the dissipation of energy, thus will affect the normal operation of the entire network. In this case you must use some strategy to complement these failed nodes and help them fulfill its original function.

3 The WSN Intrusion Detection Algorithm

3.1 The WSN Model

If the WSN is randomly distributed in a certain area[5-6] which include the sensor nodes simply named node, this network can be divided into several clusters, each cluster has a cluster head node and some ordinary nodes, aggregate nodes and the detection nodes, the cluster head node is responsible for coordinating to work the modes in the cluster, and simultaneously maintaining communication with the sink node; the Ordinary physical nodes collected ambient signal, but also store and forward data from other nodes like the routers; the detection node is responsible for detecting intrusions; the aggregate node is responsible for Summing the detection information and determining whether it has intrusion action. The network topology shown in Fig 1:

3.2 Intrusion

Though calculating various network features changes collected with some algorithm, the detection node can determine whether there is a network intrusion[7] as the following table 1(Table lists only partial information).

Table 1: network intrusion

Network characteristic data	Invasion mode
data	tampering attacks
packet transmission frequency	energy attack
packet loss rate	select forward attack
packet reception rate	forgery attack
packet transmission distance	hello attack

The detection nodes would compare the network characteristics collected with the known attack patterns of the known invasion patterns data signatures, and then find the known network intrusions.

3.3 The Multi-classification method based on binary tree

The classifier general designs for the problem of dichotomous. There are many types of attacks in reality, the response should be different against the different types of invasion attacks. So the detection system required to process the Multi-classification problems, the usual practice is to change multi-class problem into a plurality of binary classification[8]. This paper used a binary tree method, each classifier could only classify the samples of a class. So in dealing with k-class classification problems such as the following: First, sort k-class data, respectively train classifier with the samples C lass 1 samples and Class 2 ..... Class k, so we get first detection source; then train classifier with Class 2 samples and Class 3 ..... Class k samples, so we get second detection source, and so on, until separated all k-class

samples. For example 4-class classification, its binary tree structure[7] shown in Fig 2.

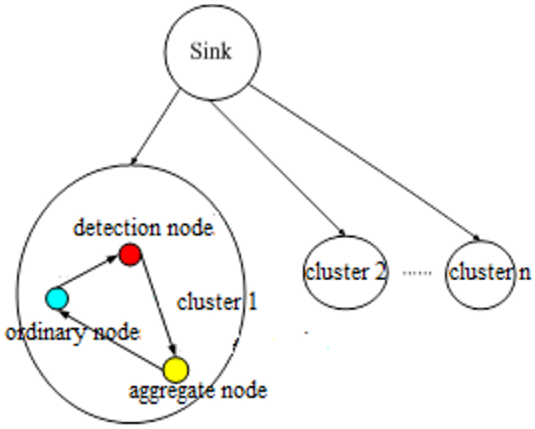


Fig 1: network topology

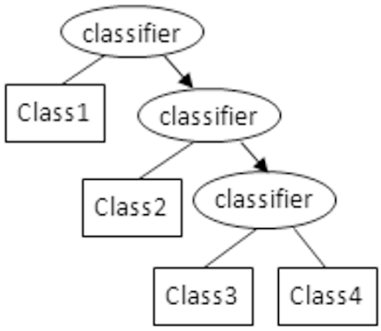


Fig2: the binary tree structure of 4-class classification

3.4 Kernel Fisher discriminant analysis theory

There are several network features on the network node X_i, assuming that the number of monitored properties were d, these features can be composed of d-dimensional vector. There are n detection nodes in the cluster, according to the size of the clusters, you can decide how much modes a cluster include, if the detection node collected n_1 samples of the normal behaviors ω_1 class, remarked $X_1 = \{X \mid X_{11}, \cdots\cdots, X_{n1}\}$; the detection node collected n_2 samples of the intrusion behaviors ω_2 class, remarked $X_1 = \{X \mid X_{12}, \cdots\cdots, X_{n2}\}$. If the detection algorithm can correctly distinguish the normal behavior ω_1 class and the intrusion behavior ω_2 class, so the detection algorithm can achieve the purpose of intrusion detection.

The kernel Fisher linear discriminant[9-10] is currently more advanced classification algorithms, the algorithm has the characteristics of the small amount of computation and the high classification accuracy, so it was widely used in the pattern recognition, the artificial intelligence and other direction. The core of the algorithm is to find the optimal projection vector as Figure 3. The waiting tested sample is projected on this optimized vector, its projection is most likely to be separated.

To find the ways of the most optimal projection vector is to calculate the maximize generalized Rayleigh entropy as formula (3-1).

$$J(\omega) = (\omega^T S_b \omega) / (\omega^T S_\omega \omega) \qquad (3\text{-}1)$$

In the formula (3-1), $S_b = (m_1 - m_2)(m_1 - m_2)^T$, S_b is the scatter matrix between the class and the class in the samples, $S_\omega = \sum_{i=1,2} \sum_{x \in X_i} (x - m_i)(x - m_i)^T$, S_ω is the scatter matrix inter-class in the Samples, m_i is the average value which calculation method as formula (3-2).

$$m_i = (1/n_i) \sum_{x \in X_i} x, i = 1,2 \tag{3-2}$$

By solving the eigenvector of the matrix $N^{-1}M$,we can gain the optimal solution, x projection on w is as formula (3-3).

$$(\omega \cdot \phi(x)) = \sum_{i=1}^{n} \alpha_i k(x_i, x) \tag{3-3}$$

In practice, N may be non-positive definite matrix, in order to preventing this situation from happening, we usually to add μ unit matrixes to N, that is the matrix N_μ instead of the matrix N as following formula (3-4) .

$$N_\mu = N + \mu I \tag{3-4}$$

In the formula (3-4), I is the identity matrix.

3.5 SVM

The sample x is k-dimensional vector[11], in a certain region the category of 1 sample is determined by the formula $(x_1, y_1), \cdots (x_l, y_l) \in R \times \{\pm 1\}$, If the Sample is hyperplane, the category of 1 sample is determined by the formula (3-14)

$$w \cdot x + b = 0 \tag{3-5}$$

The samples can be divided into two categories, in the formula (3-5), the dot represents the dot product of the vector. The best hyperplane is that which is the farthest away from the two kinds of sample. Obviously, in the formula (3-5) the result still can satisfy the equation after multiplying between w,b and the coefficient. Without loss of generality, to all samples xi , the minimum of the formula $|w \cdot x + b|$ is 1, so the minimum distance between the samples and the hyperplane is the formula $|(w \cdot x_i + b)| / \|w\| = 1 / \|w\|$ .The best hyperplane should satisfy the constraint as formula(3-6).

$$y_i [w \cdot x_i + b] \geq i, i = 1, \ldots, l. \tag{3-6}$$

The data of the training algorithm of the linear support vector displayed with the form of the dot product $(x_i \cdot x_j)$. Now the input space is mapped into a feature space

with the nonlinear mapping, and remarked $\Psi : R_k \to H$, if there is a kernel function K as formula (3-7).

$$K(x_i, x_j) = (\Psi(x_i) \cdot \Psi(x_j)) \tag{3-7}$$

We can do some calculation in the feature space without knowing the specific mapping Ψ .

Instead of the dot product form of a linear SVM with the kernel function, the dual planning of the formula (3-7) can be changed into the formula (3-8).

$$\begin{aligned} &\max W(a) = \sum_{i=1}^{l} \alpha - \frac{1}{2} \sum a_i y_i a_j y_j K(x_i \cdot x_j), \\ &s.t. \sum_{i=1}^{l} \alpha_i y_i = 0, \\ &\alpha_i \in [0, C], i = 1, \ldots, l, \end{aligned} \tag{3-8}$$

The nonlinear decision function of SVM is formula (3-9).

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{l} \alpha_i y_i K(x_i \cdot x) + b \right). \tag{3-9}$$

4 Simulation

The experiment was finished with the platform of MATLAB2012A and NS2[10],the simulated data is the WSN dataset provided by the Naval Research Lab[11-12] in which contained the normal network dataset{NS1，NS2} and the attack dataset {AS1，AS2，AS3，AS4}, we simulated and implemented four types of the attack scenarios such as the Passive Sinkhole Attacks (PSA), Periodic Route Error Attacks (PREA), Active Sinkhole Attacks (ASA)，Denial of Service (DoS) with the attack dataset {AS1，AS2，AS3，AS4} in which also included a large number of normal network data. This four type of attacks are common the WSN intrusions, in this paper, we do some experiments with the above-mentioned five types of data.

4.1 The accuracy of the assessment

First, we randomly selected the training samples and the testing samples with each 3000, in these data the proportion of the normal data is 60%, the proportion of the DDoS sample attack points is 15% , the proportion of the Passive Sinkhole Attacks sample is 10%,the Periodic Route Error Attacks sample is 10%, the Active Sinkhole Attacks sample is 5%. After training a classifier with the training samples data, we do the performance test with the test samples.

Then, we do the simulation with the algorithm described above, and we compared its test results with BP, the SVM methods and the Kernel Fisher Discriminant algorithms. The detection rate and the false detection rate after dealing as shown in Table 2.

Table 2: The detection rate and the false alarm rate

Attack Type	Kernel Fisher Discriminant		BP Network	
	Detection rate	False detection rate	Detection rate	False detection rate
PSA	83.12	5.21	83.24	12.41
Dos	77.63	3.27	91.06	9.22
PREA	62.73	33.35	53.33	48.66
ASA	90.30	4.54	90.59	8.67

Attack Type	SVM		Distributed Sensor	
	Detection rate	False detection rate	Detection rate	False detection rate
PSA	92.33	7.59	94.63	4.11
Dos	97.0	5.63	93.54	2.01
PREA	69.77	40.22	71.12	29.84
ASA	94.55	5.88	97.55	3.98

The comparison of the experimental results could see that the detection rate of this scheme was significantly higher than BP and SVM and kernel Fisher discriminant analysis. The results of the experiment showed that the algorithm could effectively detect the PSA attack, the ASA attack and the DDOS attacks, and the false detection rate is very low. However, due to the smaller attack samples of the PREA types and the incomplete classifier training, so the detection rate is not high.

4.2 Timely assessment

It was not enough to evaluate the performance of intrusion detection algorithm that we only did a comparison on the two aspects of the detection rate and the false detection rate, we should consider the requirement of the timeliness. In the detection process it costed mainly in the process of the data calculation and the data transmission.

According to the data frame format of the IEEE802.15.4 MAC layer, we can divided the need of the packets into two categories[12]: The first category is that the detection node sends a detection result to the aggregate node; the second category is that the aggregate node sends eventual result to the network, its packet length l is 8 bytes including 2-byte frame control field, 1-byte frame order number, 2-byte address message, 1-byte valid data the MAC layer(the test results), 2 bytes of FCS.

We assumed that there were m pieces of the test samples, when the samples are detected the amount of data transmission is ml (l is the length of a packet of packet).We calculated Data transmission time as the formula $t_{tele} = \frac{ml}{R_b} + \frac{D}{c}$, R_b is

the data transmission rate of the physical layer (The data transmission rate of the IEEE802.15.4 physical layer is 250kbit / s), c is the speed of light, D is the transmission distance(D Generally is considered to be 10 meters).

The calculation of time of the various detection algorithms showed as the following Table 3.

Table 3: the detection time of each classification

Algorithm species	Test time (s)
distributed	31.44
SVM	37.30
BP	41.01
The kernel Fisher discriminant analysis	36.86
the WSN's intrusion detection based on Kernel Fisher Discriminant and SVM	29.96

From Table 3, we can see that the calculation time of new scheme in this paper is less than the other kinds of classifiers, so the real-time is higher than the other.

4.3 Energy Assessment

To a node, the increased communication energy consumption is much greater than the calculation consumption, so this paper only discussed communication energy.

Chndrakasan[13] proposed an energy consumption model: the sensor node transmits 1k-bit data to consume the energy as formula(4-1).

$$E_{send} = k \times E_{static} + k \times \varepsilon_{amp} \times d^2 \tag{4-1}$$

In the formula (4-1), ε_{amp} is the signal amplification factor of the amplifier, E_{static} is the energy of the transmit and receive circuits, d is the distance the signal sent.The sensor node calculated to consume the energy as formula (4-2).

$$E_{count} = P \times \Delta t \tag{4-2}$$

In the formula (4-2), P is the average power of the processor, Δt is the computation time.

We respectively made the simulation experiments with SVM, BP, kernel Fisher discriminant analysis and the WSN's intrusion detection algorithm based on the Kernel Fisher Discriminant and SVM of this paper, their energy consumption shown in Fig 4.

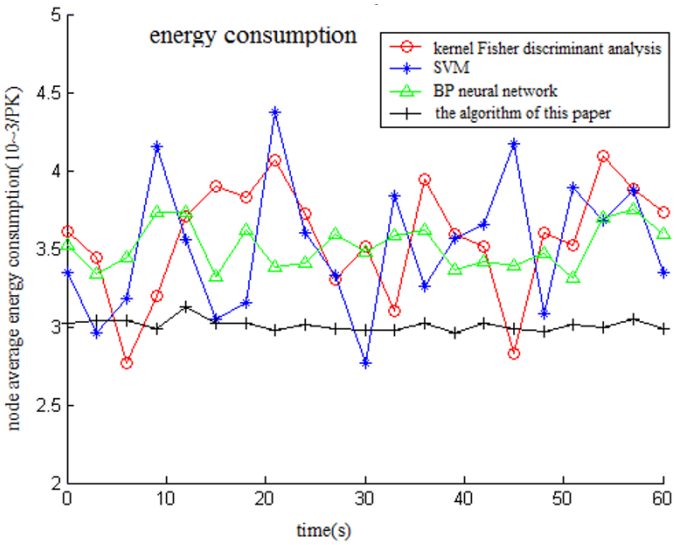


Fig 4:the energy consumption of four algorithms

From the Fig 4 we can see that the energy consumption of the algorithm of this paper always remains at a low level, why? Because the algorithm select the most appropriate way to solve the most appropriate problem, and reduced the computational burden of the detection node.

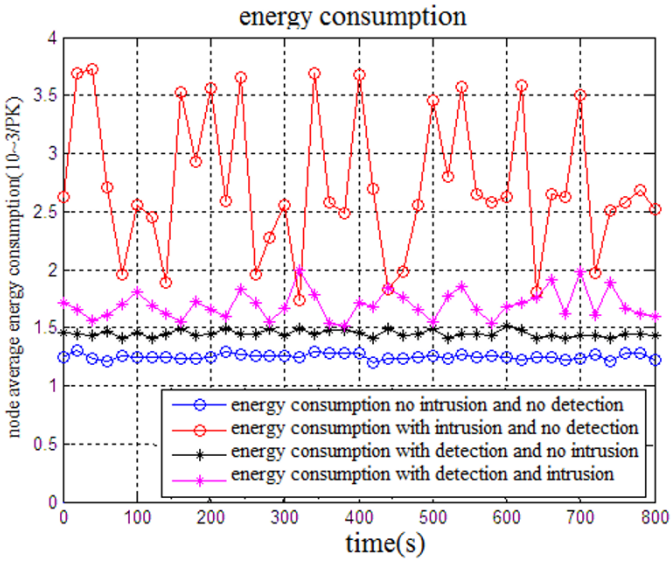


Fig 5: the energy consumption's comparison with and no intrusion(detection)

The comparison of the energy consumption of the node in the network before and after adding the intrusion detection system is shown in Fig 5.From the Fig 5, In the case of having intrusion action, after adding the network detection system, the network nodes greatly reduced the energy consumption of nodes, and improved the network lifetime.

4.4 Lifetime assessment

According to the network model and the consumption model, we respectively simulated the energy consumption of the detection nodes, the aggregate nodes and the ordinary nodes, the energy consumption of those three categories of nodes as shown in Fig 6.

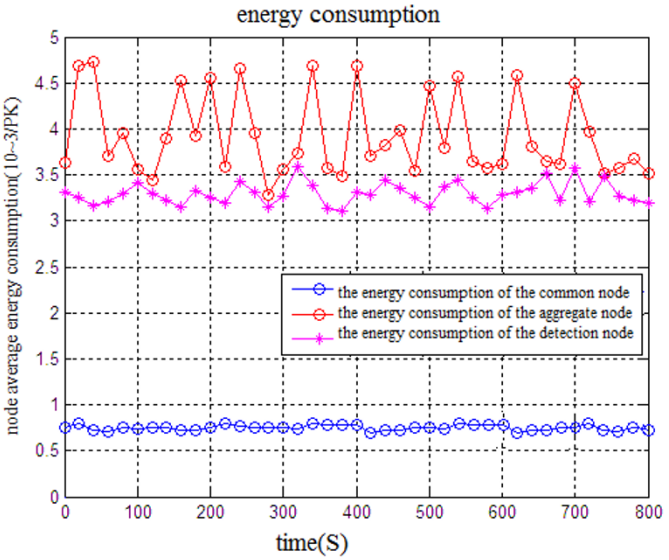


Fig 6: the energy consumption of those three categories of nodes

From Fig 6, we could see that the energy consumption of the detection node, the aggregation node were more than six times of the energy consumption of the normal node. The initial energy of these modes were the same, so the detection nodes and aggregation nodes must firstly deplete their energy than the common nodes. The way to solve this problem generally was to supplement the energy of the detection nodes and aggregation nodes with the redundant nodes'.

We assumed that the initial energy of the node was E, network had n redundant nodes(These nodes were usually in a dormant state, their energy consumption was negligible), k1 detection nodes, and k2 aggregate nodes, k3 ordinary nodes, n1, n2 were the number of redundant nodes to complement the detection nodes and the aggregation nodes, n3 was the number of the redundant nodes set aside to deal with unexpected situations(n3 was generally constant) , $\bar{t}$ was per unit time , then the lifetime of the detection node shown in Fig (4-3).

$$\frac{(n_1+k_1)E-n_1(E_{sendin}+E_{revin})}{(E_{count}+E_{send1})k_1/\bar{t}}=t_1 \tag{4-3}$$

E_{send1} is the consumed energy when the detection node send the detection result to the aggregate Node, E_{sendin} and E_{revin} are the energy consumption that the

network sent the detection algorithm to the new added detection node when the network added the failure detection node, E_{count} is the calculated energy consumption.

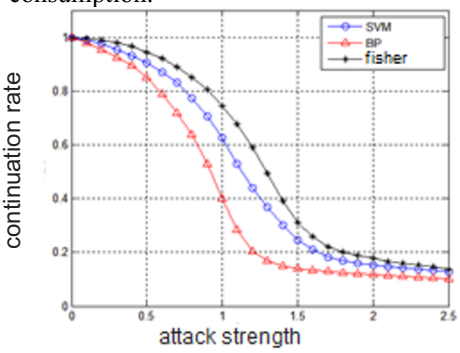


Fig 7: the influence of attack strength α on a continuation of the rate λ

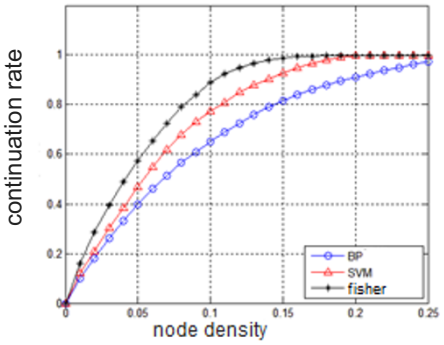


Fig 8: the influence of the node density ρ on the continuation rate λ

We simulated BP, SVM, the kernel Fisher discriminant analysis three detection algorithm, and Compared the change of their continuation rates λ, we found that the main factor affecting the network survivability and continuity was the attack strength α and the node density ρ. When the node density ρ was a constant, the simulation figure of the attack strength α on the influence of the continuation rate λ was shown as Fig 7. From the Fig 7, when the node density ρ was unchanged, the network continuation rate λ of the wireless sensor reduced with the Increase of the attack strength α.

When the attack strength α is a constant, the simulation figure of the influence of the node density ρ on the continuation rate λ was shown as Fig 8.From the Fig 8, when the attack strength α was unchanged, increasing the node density ρ could increase the continuation rate λ.When $\lambda = 1$, if you continued to increase the density of nodes ρ, it would result in increased network load due to node density was too large, even might occurred the phenomenon of inter-node signals interference. From the Fig 8, we also saw that the different detection algorithms also had some impact on the continuation rate λ, the main reason is that the energy consumption of the internal communications of the different detection algorithms was different, so the energy consumption calculation also had an impact on this.

From the above analysis we can draw: In order to achieve optimal full extension network ,when we deployed the WSN, according to the using detection algorithm, in ensuring the continuation rate $\lambda = 1$, according attack power network α, we should dynamically adjust x.

5 conclusion

This paper designed a distributed the WSN intrusion detection algorithm which could make the classification to process the most appropriate itself data. The theoretical analysis and simulation results showed that this algorithm had a higher accuracy, timeliness and the lower energy consumption. But whatever the detection node used any algorithm, its energy consumption always had been 6 times of the common node, and the detection node died faster than the ordinary node. Therefore finally this paper discussed deeply the lifetime, by using the redundant nodes in the network to supplement the detection nodes and the aggregate nodes to solve the problem of the detecting nodes' excessive death. From the theoretical analysis and simulation experiments we knew that the lifetime depended on not only the algorithm itself but also the node density and the attack intensity.

References

1. Song Lijun,Li Nayuan,Wang Aixin. An Improved Security Protocol for Wireless Sensor Network Routing[J].Chinese Journl of Senors and Actuators, 2009, 10: 1471-1475
2. Yang Libing,Mu Dejun,Cai Xiaoyan.Study on intrusion detection for wireless sensor network[J]. Application Research of Computers, 2008, 11: 3204-3209
3. Visual analytics for intrusion detection in spam emails Jinson Zhang; Mao Lin Huang; Doan Hoang DOI: http://dx.doi.org/10.1504/IJGUC.2013.056254 187-196
4. A study on network security monitoring for the hybrid classification-based intrusion prevention systems Oscar Rodas; Marco Antonio To DOI: 10.1504/IJSSC.2015.069240
5. Hu Zhipeng,Wei Lixian,Shen JunWei,Yang Xiaoyuan. An Intrusion Detection Algorithm for WSN Based on Kernel Fisher Discriminant[J]. Chinese Journal of Sensors and Actuators. 2012.7: 1189-1193
6. Use of wireless sensor networks for distributed event detection in disaster management applications Majid Bahrepour; Nirvana Meratnia; Mannes Poel; Zahra Taghikhaki; Paul J.M. Havinga DOI: 10.1504/IJSSC.2012.045569
7. Zhu Qi,Song Rushun,Yao Yongxian. SVM-based cooperation intrusion detection system for WSN[J]. Application Research of Computers ,2010, 27(4):1489-1492
8. Decentralised malicious node detection in WSN Alaa Atassi; Naoum Sayegh; Imad H. Elhajj; Ali Chehab; Ayman Kayssi DOI: 10.1504/IJSSC.2014.060685
9. An effective attack detection approach in wireless mesh networks Felipe Barbosa Abreu; Anderson Morais; Ana Cavalli; Bachar Wehbi; Edgardo Montes de Oca; Wissam Mallouli DOI: 10.1504/IJSSC.2015.069204100-114
10. The Network Simulator-NS2[EB/OL]. http://www.isi.edu/nsnam/ns, 2006-09-17
11. Downard I. Simulating Sensor Networks in NS2. Technical Report[R]. NRL/FR/5522-04 - 10073, Naval Research Laboratory, Washingt on, D. C. ,U. S. A. , May 2004.
12. Yan K Q, Wang S C, Liu C W. A Hybrid Intrusion Detection System of Cluster-Based Wireless Sensor Networks[C]. Proceedings of the International Multi Conference of Engineers and Computer Scientists, 2009I: 956-963
13. Heinzelman W B, Chndrakasan A P, Balakrishnan H. An Application-Specific Protocol Architecture for Wireless Micro sensor Networks[C]. IEEE Transaction on Wireless Communications, 2002, 1(4): 660-670.

Automatic Verification of Security of OpenID Connect Protocol with ProVerif

Jintian Lu[1], Jinli Zhang[2], Jing Li[3], Zhongyu Wan[4], Bo Meng[5]
Correspondent author: Bo Meng
[1-3,5] School of Computer, South-Central University for Nationalities,
MinYuan Road #708, HongShan Section,430074,Wuhan, Hubei, China;
[4]School of Netcenter,Jianghan University SanJiaoHu Road #8,CaiDian Section,430056
Wuhan,Hubei,China
[1]Ljt45@hotmail.com;[2]jinlysa@163.com;[3]stevenlee710@sina.com;
[4]wanzhongyu@jhun.edu.cn;[5]mengscuec@gmail.com

Abstract. Owning to the widely deployment of OpenID Connect protocol in the important applications, in order to provide a strong confidence in its security for the people, in this study, we firstly review OpenID Connect protocol. And then, we use the formal language: Applied PI calculus to model OpenID Connect protocol and provide a security analysis with the automatic tool ProVerif. Finally, we find it does not have the secrecy and have some authentications. We present some approaches to address the security problems in OpenID Connect protocol.

1 Introduction

In order to prevent identity-oriented attacks and simplify the identity management systems[1], some of identity management had been proposed, such as OpenID Connect[2],OAuth2.0[3],CardSpace and OpenID[4]. OpenID Connect is a replacement for the OpenID scheme. It provides federated identity management and authentication[5] by adding authentication capabilities on the top of OAuth2.0 protocol. Owning to the widely deployment of OpenID Connect by Google[6], Microsoft[7] and PayPal[8] .In order to provide a strong confidence in the security[9], Wanpeng Li et al.[10] reveal serious vulnerabilities of a number of types, all of which allow an attacker to log in to an RP website as a victim user in Google's implementation of OpenID Connect.

Not like the OAuth2.0,it already has been analysis using formal methods[11],very little research has been conducted on the security of OpenID Connect. So, it's important to analyze the security of OpenID Connect. We use the formal language : Applied PI calculus to model the OpenID Connect security protocol and provide a security analysis with the automatic tool ProVerif[12].

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_20

2 OpenID Connect Protocol

OpenID Connect was published in February 2014,since it is based on OAuth. In OpenID Connect, it has three parties: End_User, Relay Part(RP)[13] and OpenID Provider(OP).The End_User is a owner of resource which stored in OP.RP is a application providing a certain service which requires authentication which is delegated to the corresponding OP. OP is capable of Authenticating the End-User and providing Claims to a Relying Party about the Authentication event and the End-User. In OpenID Connect, there are three authentication flows: Authorization Code Flow, Implicit Flow and Hybrid Flow[13]. In this study, we choose the Hybrid Flow to implement because it has better representativeness in implementation of OpenID Connect.

OpenID Connect protocol mainly consist of three parties: End_User(E_U),OpenID Provider(OP) and Relay Part(RP).It enables RP to verify the identity of the E_U based on the authentication performed by an Authorization Server which contained in OP, as well as to obtain basic profile information of E_U.The core OpenID Connect functionality is authentication builds on top of OAuth2.0 and the use of Claims to communicate information about the E_U,it describes the security[14] and privacy.

The authentication of OpenID Connect Protocol mainly contains six messages exchange among E_U,OP and RP when use the Hybrid Flow.

(1) authenticationRe:=client_id||response_type||scope||redirect_uri||state

To access the protect resource of E_U,RP generated message(1) authenticationRe and sends it to the Authorization Server in OP for authentication. Parameters in this request mainly contains client_id which is a client identify valid at the Authorization Server; response_type value determines the authorization processing flow to be used, when using the Hybrid Flow, this value is code id_token token, but in our implementation, it's value is code id_token because we delete the token which is optional; scope is a OpenID scope value, if it doesn't present, the behavior is entirely unspecified; redirect_uri exactly match one of the Redirection URI values the client pre-registered at OP; state is a Opaque value used to maintain state between the request and the callback.

(2) authenticationE_U:=ask_authentication

When Authorization Server received the message(1),it would validate all of the parameters according to the OAuth2.0 specification, and it must verify all of parameters are present and their usage confirmed to this specification. Then, Authorization Server will generate the message(2) authenticationE_U which contains ask_authentication to authenticate the E_U and sends it to End_User.

(3) authorization :=username||userpassword

When E_U received message(2),it would validate whether the parameter is ask_authentication, if true, the E_U will create the message(3) authorization contains username and userpassword and sends it to Authorization Server of OP which indicate that End_User authorizes to OP. The username is the name which registered name previously at OP, userpassword is the secret of the E_U when registered previously at OP.

(4) authenticationResp:=code||id_token

By the time Authorization Server received message (3), it will get username and userpassword which sent from E_U and get the information of grant. Therefore, it will generate message (4) authenticationResp which contains code and id_token and sends it to RP. The code is the authorization code which used to exchange the access_token at Token Endpoint.

(5) tokenrequest:=grant_type||code||redirect_uri||client_secret||client_id

When RP received message (4),it will be valid if the code is correct and not has been used. Then it generates message(5) tokenrequest which contains grant_type that value is code indicate that the RP wants to use the code to receive the access_token from Token Endpoint of OP, code which is match with the code of message(4), redirect_uri is matched with message(1), client_secret,the secret shared by the RP and OP, client_id is the same client_id of message(1) indicate that the message(5) is sent by the same RP with message(1).Finally, it sends message(5) to OP.

(6) token_response :=access_token||id_token||token_type||expiress_in
signedMessage

After Token_Point of OP received message (5), firstly it must verify the value of grant_type and code is authorization code, if verified, Token Endpoint generates message (6) token_response which contains access_token is a evidence that the user has been authenticated, id_token is same with message (4),token_type which value is Bear, expiress_in is the longest live time of access_token, signedMessgae is the digital signature of parameters above use the private key of OP. Finally, it sends message (6) to RP. When RP receives message (6), firstly, it verifies the signature use the public key of OP, if verification is successful, the protocol is finished. The detail procedure is shown as Fig.1.

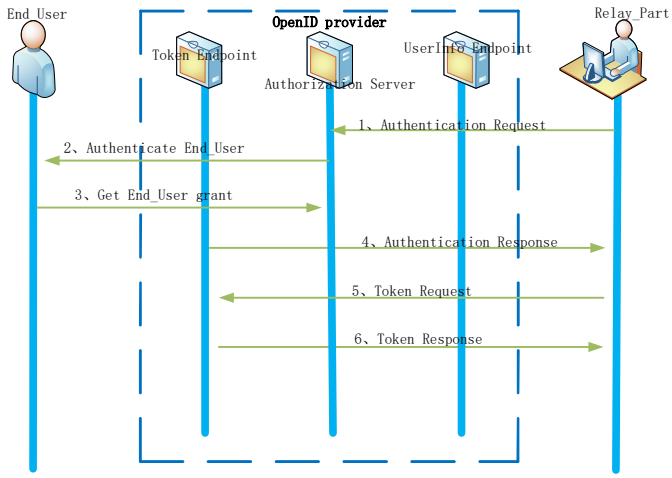


Fig.1. The Procedure of OpenID Connect Protocol

3 Formalizing the OpenID Connect protocol using Applied PI Calculus

The Applied PI calculus was made by Abadi et.al [15], in 2001, it is a formal language which used to modeled correspondence between concurrent process. Applied calculus added functions and equation primitives on the basis of structure of correspondence and concurrence. Message can not only contains names, but also the values consist of names and functions. PI calculus is convenient to describing the type of standard data, it is convenient to formal modeling of security protocols too. For example, it denoted values of any types with variables, denoted values of atom with names and denoted the cryptographic primitives[16] such as encryption ,decryption, signature[17] and XOR with functions, so it can be used to modeled and analysis the complicated security protocols.

3.1 The Function and Equational Theory

The function and equational theory are introduced in this part. We use the Applied PI calculus to model OpenID Connect Protocol.Fig.2.describes the function and equational theory in OpenID Connect protocol.

We use fun sign(x,PR) to sign the message x with the private key PR and verification algorithm versign(x,PU) to verify the digital signature x with the public key PU ,and the fun decsign(x,PU) recover message from the digital signature x with the public key PU .The function fun PR(b) accepts private value b as input and produces private key as output. The function fun PU(b) accepts public key value b as input and produces public key as output.

```
fun sign(x,PR).
fun PR(b).
fun PU(b).
fun versign(x,PU).
fun decsign(x,PU).
equation versign(sign(x,PR),PU)=x.
equation decsign(sign(x,PR),PU)=x
```

Fig.2. The Functions and the Equational Theory

3.2 Processes

The complete processes of OpenID Connect mainly contains four processes: Main process, End_User process, OpenID Provider process and Relay Part process. The main process OpenID consist of End_User process, OpenID Provider process and Relay Part process, shown as Fig.3.

```
OpenID=
!processOP|!processRP|!processEU
```

Fig.3. Main Process

Process End_User was modeled use the Applied PI calculus as Fig.4.At first, it receives the message2 through free channel c from Process OP by the statement in(c,m2) ,then it will check whether the parameter ask_authentication is equal with message2 and it generates own username username ,userpassword userpassword and the construct secretX .Finally, it generates message authorization contains username and userpassword and sends it to process OpenID Provider through free channel c .

```
processEU=              (*****Process E_U********)
in(c,m2);          (*******E_U recieves message2 from OP******)
if ask_authentication=m2 then
let secretX=userpassword in
let authorization = (username,userpassword) in
out(c,authorization).          (****E_U sends message3 to OP****)
```

Fig.4. End_User Process

OpenID Provider process was modeled use Applied PI calculus as Fig.5.Firstly,it receives mesasge1 from Relay Part Process and extract parameters client_id_op,response_type_op,scope_op,redirect_uri_op,state_op from message m1 ,after that if response_type_op is equal with code_id_token and scope_op is equal with code ,it will create message authenticationE_U contains ask_authentication and sends it to End_User process through free channel c .Next, it receives message3 from End_User process and exact username_op and userpassword_op from message m3 ,if username_op and userpassword_op is matched with username and userpassword stored in OP, it will create parameter code_op and id_token_op and sign them with private key keyop2 and then sends generated message authorizationResp contains parameters code_op,id_token_op,signedM4 and sends this message to Relay Part through free channel c .Later, it receives message5 from RP and exact grant_type , code_op , redirect_uri_op , client_secret_op and client_id_op ,from message5 and check the value of grant_type and code_op is authorization code, then it creates parameters access_token_op , id_token_op , token_type_op , expires_in_op , and sign them use digital signature with private key keyop1 of OP, it generate massage token_response contains parameters access_token_op , id_token_op ,

token_type_op , expires_in_op , signedMessage and sends this message to Relay Part through free channel c .So far, OpenID process ended.

```
processOP=                    (*****Process OP**********)
in (c,m1);              (*****OP recieves message1 from RP******)
let (client_id_op,response_type_op,scope_op,redirect_uri_op,state_op)=m1 in
if scope_op=scope then
if response_type_op=code_id_token then
new ask_authentication;
let authenticationE_U=ask_authentication in
out(c,authenticationE_U);       (*******OP sends message2 to End_User*******)

in(c,m3);               (********OP recieves message3 from End_User*******)
let (username_op,userpassword_op)=m3 in
if username_op=username then
if userpassword_op=userpassword then
new code_op;new id_token_op;
let signedM4=sign((code_op,id_token_op),PR(keyop2)) in
let authorizationResp=(code_op,id_token_op,signedM4) in
out(c,authorizationResp);  (*******OP sneds message4 to RP********)

in(c,m5);               (*******OP recieves message5 from RP********)
let(grant_type_op,code_op,redirect_uri_op,client_secret_op,client_id_op)=m5 in
if grant_type_op=code then
if code_op=code then
new access_token_op;new id_token_op;new token_type_op;new expires_in_op;
let signedMessage=sign((access_token_op,id_token_op,token_type_op,expires_in_op),PR(keyop1)) in
let token_response=(access_token_op,id_token_op,token_type_op,expires_in_op,signedMessage)in
out(c,token_response).        (**OP sends m6 which was signed to RP**)
```

Fig.5. Process OpenID Provider

The Relay part process was modeled by Applied PI calculus as Fig.6, firstly it generates message authenticationRe which contains lient_id , response_type , code , redirect_uri and state ,it sends this message to OpenID Provider process. Next, it receives message4 m4 sent from OP through free channel c and exacts code_rp , id_token_rp and signedM4 from message 4 m4 , after that, it will verify the signed message signedM4 signed use digital signature with public key keyop2 ,of OpenID Provider. Then, it generates message tokenrequest which contains declared parameters grant_type_rp , code_rp , redirect_uri_rp , client_id and client_secret and sends this message to OpenID Provider through free channel c .Finally, relay part process receives message m6 sent from OpenID Provider and exact access_token_rp , id_token_rp , token_type_rp , expiress_in_rp and signedMessage1 .It verifies the signed message signedMessage1 with public key keyop1 of OP, if verification is successful, it creates a parameter finished and sends it through free channel, so far, protocol ends.

```
let processRP=                    (********Process RP ***********)
new client_id;new response_type;new scope;new redirect_uri;new state;
let authenticationRe=(client_id,response_type,scope,redirect_uri,state) in
out(c,authenticationRe);    (**RP sends message1 to OP**)

in(c,m4);                (******RP recieve message4 from OP******)
let (code_rp,id_token_rp,signedM4)=m4 in
if versign((signedM4),PU(keyop2))=(code_rp,id_token_rp) then

new grant_type_rp;new code_rp;new redirect_uri_rp;
new client_secret_rp;new client_id_rp;
let tokenrequest=(grant_type_rp,code_rp,redirect_uri_rp,client_secret_rp,client_id_rp) in
out(c,tokenrequest);     (**RP sends m5 to OP**)

in(c,m6);
let (access_token_rp,id_token_rp,token_type_rp,expires_in_rp,signedMessage1)=m6 in
if versign(signedMessage1,PU(keyop1))=(access_token_rp,id_token_rp,token_type_rp,expires_in_rp) then
new finished;
out(c,finished).        (***************Finished*****************)
```

Fig.6. Relay Part Process

4 Automatic Verification of Secrecy and Authentications with ProVerif

We use the statement query attacker:secretX. in ProVerif to verify the secrecy of userpassword which is username registered previously at OP. ProVerif uses the non-injective agreement to model the authentication, shown as Table 1. So we use query ev:e1==>ev:e2 to model the authentication. It's true when event e1 executed and then event e2 had been executed before the event e1.

Table 1. The Authentications

Non-injective agreement	Authentications
ev:endauthusera_s(x)==>ev:beginaauthusera_s(x)	Authorization Server authenticates End_User
ev:endautha_suser(x)==>ev:beginaautha_suser(x)	End_User authenticates authorization server
ev:endauthRPE_p(x)==>ev:beginaauthRpE_p(x)	Token Endpoint authenticates RP
ev:endauthE_pRP(x)==>ev:beginaauthE_pRP(x)	RP authenticates Token Endpoint

As we know, the input formats consist of Horn clauses and the extension of Applied PI calculus, but they have the same output system. In this paper, we use the Applied PI calculus as the input. The model using Applied PI calculus must to

translated into the syntax of ProVerif and the inputs of ProVerif in extension of PI calculus. The Fig.7-Fig.10 are the inputs for OpenID Connect protocol.

```
query attacker:secretX.

query ev:endauthusera_s(x)==>ev:beginaauthusera_s(x).
(**authorization server authenticates End_User**)
query ev:endautha_suser(x)==>ev:beginaautha_suser(x).
(**End_User authenticates authorization server**)
query ev:endauthRPE_p(x)==>ev:beginaauthRpE_p(x).
(**Token Endpoint authenticates RP**)
query ev:endauthE_pRP(x)==>ev:beginaauthE_pRP(x).
 (**RP authenticates Token Endpoint**)
```

Fig.7. Query Secrecy and Authentications in ProVerif

```
            .......
if ask_authentication=m2 then
event endautha_suser(m2);
let secretX=userpassword in
let authorization = (username,userpassword) in
event beginaauthusera_s(authorization);
out(c,authorization).
```

Fig.8. The End_User Process in ProVerif

```
        ......
event endauthE_pRP(signedM4);
        ......
let tokenrequest=(grant_type_rp,code_rp,redirect_uri_rp,client_secret_rp,client_id_rp) in
event beginaauthRpE_p(tokenrequest);
out(c,tokenrequest);
        ......
```

Fig.9. The Relay Part Process in ProVerif

```
                ......
let authenticationE_U=ask_authentication in
event beginaautha_suser(authenticationE_U);
out(c,authenticationE_U);
                ......
if userpassword_op=userpassword then event endauthusera_s(m3);
                ......
let authorizationResp=(code_op,id_token_op,signedM) in
event beginaauthE_pRP(signedM);
out(c,authorizationResp);
                ......
event endauthRPE_p(m5);
                ......
if ask_authentication=m2 then event endautha_suser(m2);
let secretX=userpassword in
let authorization = (username,userpassword) in
event beginaauthusera_s(authorization);
out(c,authorization).
```

Fig.10. The OpenID Provider Process in ProVerif

We run these input of OpenID Connect use the ProVerif as shown as Fig.7-Fig.10. The result shown as Fig.11-Fig.17.Fig.11 shows the result of query attacker:secretX .We can find that the secretX does not have secrecy from the result because the secretX was sent by a plaintext way, it is easy to attacker to monitor the free channel c and get secretX .Therefor, secretX doesn't have secrecy. To solve this problem we can use some security mechanisms[18], for example digital signature.

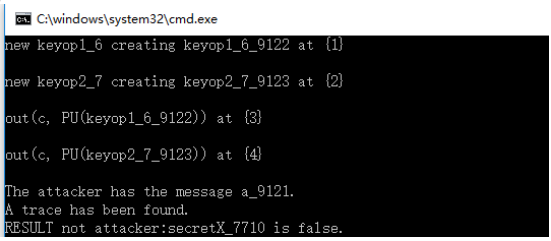


Fig.11. The Result of Secrecy

Fig.12 shows the result of query ev:endautha_suser(x)==>ev:beginaautha_suser(x), we can know End_User can not authenticate Authorization Server because it's a plaintext way that Authorization Server sends the parameter ask_authentication ,it can be gotten by attacker easily, so we can use the encryption or digital signature to solve the problem.

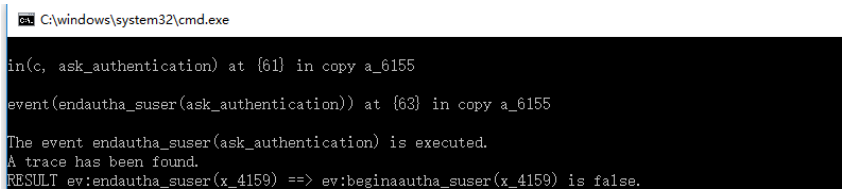


Fig.12. The Result of End_User Authenticates Authorization Server

Fig.13 shows the result of query ev:endauthusera_s(x)==>ev:beginaauthusera_s(x) , we can find Authorization Server can not authenticates End_User, because the protocol does not use the security approach when End_User sends username and userpassword to Authorization Server. For the security of protocol, we can use digital signature to enhance the security.

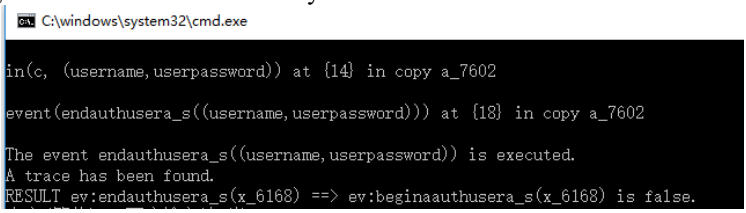


Fig.13. The Result of Authorization Server Authenticates End_User

Fig.14 shows the result of query ev:endauthRPE_p(x)==>ev:beginaauthRpE_p(x) , we can find the result is false that indicates the Token Endpoint can not authenticate the Relay Part from Fig.14,because the RP does not use any security approaches when it sends the message tokenrequest to OP. We can use the digital signature to enhance the security.

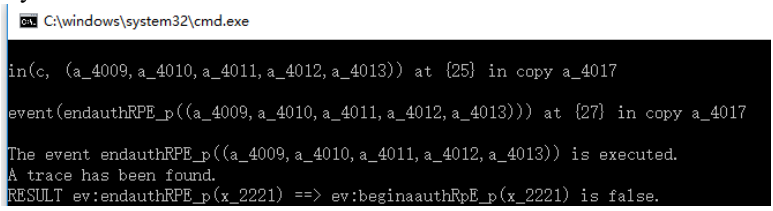


Fig.14. The Result of Token Endpoint Authenticates Relay Part

Fig.15 shows the result of query ev:endauthE_pRP(x)==>ev:beginaauthE_pRP(x) , the result is true indicates that RP can authenticate Token Endpoint because it use the digital signature which is a way of security when Token Endpoint sends the message authorizationResp to RP .

Fig.15. The Result of Relay Part Authenticates Token Endpoint

5 Conclusion

In this study, we use the formal language to model the authentications of communicating parties of OpenID Connect protocol, the result shows us the no-security and have some authentications between the communicating parties. In the future work, we will mainly to solve some no-authentication.

References

1. Ronak R.Patel,Bhavesh O.Enhance OpenID Protocol in identity Management.International Journal of Application or Innovation in Engineering &Management.Vol.2,No.4,Apr. 2013: 248-252
2. Nat S., John B., Michael J. ,Breno D M. ,Mortimore C.. OpenID Connect Core 1.0, 2014.http://OpenID.net/specs/OpenID-connect-core-1_0.html.
3. Dick Hardt. The OAuth 2.0 authorization framework. October 2012.http://tools.ietf.org/html/rfc6749.
4. David R. ,Brad F. . OpenID Authentication 2.0 - Final, 2007. http://OpenID.net/specs/OpenID-authentication-2_0.html.
5. Alireza P S., Joao P S. Authentication, authorization and auditing for ubiquitous computing: a survey and vision. International Journal of Space-Based and Situated Computing.Vol.1 No.1,2011:59-67
6. Google OpenID Connect 1.0, 2015. https://developers.google.com/accounts/docs/OpenIDConnect.
7. Microsoft OpenID Connect, 2014. https://msdn.microsoft.com/en-us/library/azure/dn6455
8. PayPal OpenID Connect 1.0, 2014. https://developer.paypal.com/docs/integration/direct/identity/log-in-with-paypal
9. Blanchet B. Automatic proof of strong secrecy for security protocols//Proceeding of the 2004 IEEE Symposium on Security and privacy,California,2004:86-100
10. Wanpeng L. and Chris J M. Analysing the Security of Google's implementation of OpenID Connect. Information Security Group, Royal Holloway, University of LondonTW20 0EX.Aug.2015:1-27
11. Guangye S., Mohamed M. FASER (Formal and Automatic Security Enforcement by Rewriting) by BPA algebra with test. International Journal of Grid and Utility Computing.Vol4,No.2-3,2013:204-211
12. Blanchet B. An efficient cryptographic protocol verifier based on prolog rules//Proceeding of the 14th IEEE Computer Security Foundations Workshop, Cape Breton,2011:82-96

13. OpenID Connect Core 1.0 incorporating errata set 1. http://OpenID.net/specs/OpenID-connect-core-1_0.html#toc
14. Terry B., Brian L., Thomas S., John S. ,Gautam K.. An example of the use of Public Health Grid (PHGrid) technology during the 2009 H1N1 influenza pandemic. International Journal of Grid and Utility Computing.Vol.4,No.2-3,(2013):148-155
15. Abadi M.,Fournet C.Mobile values,new names, and secure communication//Proceeding of the 28th ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages,London,2001:104-115
16. Alessandro B. Gerardo. Federico T. Secure and efficient design of software block cipher implementations on microcontrollers. International Journal of Grid and Utility Computing.vol.4,No.2-3,2013:110-118
17. N. Bharadiya B. ,Soumyadev M., R.C. Hansdah. An authentication protocol for vehicular ad hoc networks with heterogeneous anonymity requirements. International Journal of Space-Based and Situated Computing.Vol.4,No.1,2014:1-14
18. David M. ,Herve G. Security in wireless sensor networks: a survey of attacks and countermeasures. International Journal of Space-Based and Situated Computing.Vol.1,No.2-3,2011:151-162

Jintian Lu is now a postgraduate at school of computer, South-Center University for Nationalities, China. His current research interests include security protocols and formal methods.

Jinli Zhang is a postgraduate at school of computer, South-Center University for Nationalities, China. Her current research interests include security protocols and formal methods.

Jing Li is now a postgraduate at school of computer, South-Center University for Nationalities, China. His current research interests is network security

Zhongyu Wan is a postgraduate at school of computer, South-Center University for Nationalities, China. Her current research interests include security protocols and formal methods.

Bo Meng was born in 1974 in China. He received his M.S. degree in computer science and technology in 2000 and his Ph.D. degree in traffic information engineering and control from Wuhan University of Technology at Wuhan, China in 2003. From 2004 to 2006, he worked at Wuhan University as a postdoctoral researcher in information security. From 2014 to 2015, he worked at University of South Carolina as a Visiting Scholar. Currently, he is a full Professor at the school of computer, South-Center University for Nationalities, China. He has authored/coauthored over 50 papers in International/National journals and conferences. In addition, he has also published a book “secure remote voting protocol” in the science press in China. His current research interests include security protocols and formal methods.

Low Power Computing and Communication System for Critical Environments

Luca Pilosu, Lorenzo Mossucca, Alberto Scionti, Giorgio Giordanengo, Flavio Renga, Pietro Ruiu, Olivier Terzo,Simone Ciccia, Giuseppe Vecchi

Abstract The necessity of managing acquisition instruments installed in remote areas (e.g., polar regions), far away from the main buildings of the permanent observatories, provides the perfect test-case for exploiting the use of low power computing and communication systems. Such systems are powered by renewable energy sources and coupled with reconfigurable antennas that allow radio-communication capabilities with low energy requirements. The antenna reconfiguration is performed via Software Defined Radio (SDR) functionalities by implementing a phase controller for the array antenna in a flexible low power General Purpose Platform (GPP), with a single Front-End (FE). The high software flexibility approach of the system represents a promising technique for the newer communication standards and could be also applied to Wireless Sensor Networks (WSNs).
This paper presents the prototype that is devoted to ionospheric analysis and that will be tested in Antarctica, in the Italian base called Mario Zucchelli Station, during summer campaigns. The system, developed to guarantee its functionality in critical environmental conditions, is composed of three main blocks: Communication, Computing and Power supply. Furthermore, the computing and communication system has been designed to take into account the harsh environmental conditions of the deployment site.

1 Introduction

Low power computing and communication technologies have recently gained a key role for helping the scientific community to improve the capability of processing and exchanging data, while keeping low the impact on power consumption. Low power computing inherits most of the challenges faced in managing large computing infrastructures (i.e., optimizing the resource allocation and their utilization, scalability and ease of use), putting the stress on a coordinated operation of several low power units aiming at obtaining the highest processing capabilities. From this perspective, embedded computer domain (which comprises low power computing) is moving from traditional low performance platforms to more powerful, highly parallel, general purpose systems. The availability of small computer boards equipped with full chipsets and I/O interfaces represents an inexpensive opportunity to explore parallel

Luca Pilosu, Lorenzo Mossucca, Alberto Scionti, Giorgio Giordanengo, Flavio Renga, Pietro Ruiu, and Olivier Terzo
Istituto Superiore Mario Boella (ISMB), Torino, Italy, e-mail: {pilosu,mossucca,scionti,giordanengo,renga,ruiu,terzo}@ismb.it

Simone Ciccia, Giuseppe Vecchi
Department of Electronics and Telecommunication (DET), Politecnico di Torino, Torino, Italy, e-mail: {simone.ciccia,giuseppe.vecchi}@polito.it

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_21

computer systems in many critical application scenarios. Heterogeneity of hardware modules largely contributes to the adoption of such platforms, since heterogeneity in the form of hardware accelerators (e.g., GPUs embedded in the same System-on-Chip along with CPUs) allow to take advantage from specialization without defeating flexibility and adaptability of the whole system. Finally, the relative low price of such platforms further extends their attractiveness. On the other hand, low power communication aims at obtaining excellent data transmission features while remaining compliant to the standard protocols (e.g., IEEE 802.11) and keeping the power consumption at acceptable limits. Concurrently to the evolution of wireless standards 802.11 b/g/n, reconfigurable antennas have gained a lot of attention for optimizing Signal to Noise Ratio (SNR), and reducing interferences by directing the beam towards terminals and/or access points, with the purpose of increasing link capacity, extending battery life and reducing terminals costs [1, 2, 3, 4]. Recently, base-band beamforming and Multi-Input Multi-Output (MIMO) are the most widely used technologies; however, in low power communication this does not represent the optimal choice, since MIMO systems rely on multi front-end receivers and power-hungry Digital Signal Processors (DSPs) [5, 6, 7, 8].

This work would like to demonstrate that high quality communication link can be achieved by means of terminals equipped with reconfigurable antenna - single Front-End and a software defined radio controller to steer the main beam in the direction of the base station, thus ensuring that the global energy consumption of the link is still lower when compared to the traditional approach implemented with a general on-chip IEEE802.11 wireless module. These technological improvements, i.e., low power processing and wireless communication, have been integrated together for the purpose of a case study in a critical environment: the Antarctic region. Antarctica is a valuable resource for scientific research in several domains: today the continent hosts stable research stations from 28 countries. It is also a unique scenario for stressing various technologies, thanks to its extreme environmental and isolation conditions. The scientific application addressed by the system described in this paper is based on the acquisition of GNSS satellite data for the study of radio propagation in the ionosphere [9, 10]. The prototype architecture, which is depicted in Fig.1, is split into two parts that will be installed in the Italian Mario Zucchelli Station (i.e., the permanent observatory): the *base station* will be placed inside the main building of the observatory, meaning the module does not have particular constraints regarding power supply and temperature; the *measurement station* is located 500 meters far away from the base station, and it needs to take into account some crucial environmental aspects, such as climatic constraints, power supply and management, distances and Line of Sight (LoS) conditions for wireless communications.

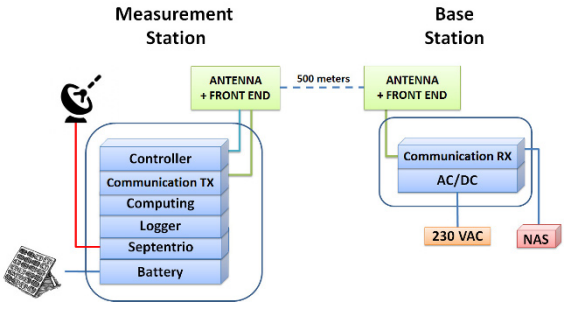


Fig. 1 Main blocks of the prototype architecture

GNSS data are currently acquired in the observatory but, so far, this study has consisted of three mandatory phases:

1. raw data acquisition on a fixed measurement station, located into a container outside the base station;
2. processing of the received data on a dedicated PC;
3. data storage on a networked hard disk, exploiting a wired link between the measurement and the base stations.

This configuration implies that two main constraints emerge: *(i)* the GNSS receiver can only be used in the measurement station that has power supply and it is reached by the network infrastructure; *(ii)* expensive hardware must be dedicated to a single task. The goal of the activity presented here is to extend the existing architecture, by implementing a ground-disposable system powered by renewable sources, and equipped with low power processing and wireless communication capabilities, in such a way it can overcome the limitations of the current fixed installations.

The rest of the paper is organized as follows: Section 2 provides a description of the scientific and technological background. In Section 3 the design is described, highlighting the main constraints and the motivation of the chosen approach and architecture. Section 4 presents the experimental results of the preliminary tests on the building blocks of the prototype. Section 5 draws conclusions and gives some ideas for future works.

2 Background and Motivation

The continuous progresses in the miniaturization of electronic chips allow to create powerful computing platforms which can easily fit the space of a very small board. Such types of computing platforms, also known as Single Board Computers (SBCs), integrate all the components of a traditional desktop-like computer (e.g., CPU, main memory, I/O interconnections, storage, etc.) in a compact boards. Raspberry Pi [11] and Arduino [12] are two popular projects proposing SBCs with the size of a credit card, with the latter more oriented to the creation of systems able to interact with the outside environment, thanks to a large set of sensors and actuators available. Although initially designed to cover applications in the embedded domain (this is mainly due to the adoption of low power embedded processor or micro-controllers), nowadays, SBCs sport high-performance Systems-on-Chip (SoCs) coupling a powerful CPU and a GPU within the same silicon die. Such platforms are able to run also complex parallel applications. These improved features make them more attractive for implementing advanced cyber-physical applications. Furthermore, inexpensive hybrid solutions mixing general purpose computational capabilities and reconfigurable logic (i.e., Field Programmable Gate Arrays – FPGAs) [13] are becoming popular. It is worth to note that, albeit the performance exhibited by such computing modules are continuously growing [14], their power (and mostly energy) consumption remains within an acceptable range, so that it is possible to effectively power them through a battery or using renewable energy sources.

The growing demand for processing and storage capabilities lead in the past computer designers to implement parallel platforms in order to crunch as much as possible data. Historically, computer clusters have been used to process large amount of data recurring to standard commercial off-the-shelf computers instead of more expensive dedicated ones, like those available in a supercomputer system. Today, the large availability of high-performance SBCs allow the creation of clusters hosted in a standard server chassis. HPE Moonshot [15] is an example of a commercial

high-density server system, which is able to pack up to 180 server nodes in a single enclosure. Similarly, several research projects are investigating on the design and use of such systems both in the high-performance and cyber-physical contexts. For instance, the FIPS project [16] developed a heterogeneous datacenter-in-a-box equipped with up to 72 ARM-based low power processors, and mixing also standard X86 processors and FPGA acceleration modules. The 1U server box is mainly designed to run complex simulations coming from the life-science domain. Conversely, the AXIOM project [17] proposes a modular design based on the UDOO boards [18] for accelerating video-processing applications that are common in the embedded domain.

The usage of small computing clusters made of inexpensive boards is becoming attractive for the scientific community [19, 20, 21], since they can provide autonomous computing and wireless communication capabilities, which are of great interest to collect data from several sensors or to act on the environment through actuators. In many applications, such systems can be powered by a battery pack or by external renewable sources. In these contexts, the adoption of a multi-board solution becomes almost mandatory, since tasks regarding monitoring the power source or controlling the wireless interface require dedicated computing resources. In particular, adaptable wireless communication systems based on smart antennas result to be high computationally demanding, thus requiring fast computation capabilities. This becomes even more true when wireless communications are managed through more flexible software-defined functionalities, since all the low level communication tasks are run as dedicated software components. However, to our best knowledge there is not yet a well defined architecture that provides the computational capabilities of a small cluster, which can be powered through a battery pack, and that is able to communicate through a wireless channel through a reconfigurable smart antenna.

Given this premise, our solution represents a first attempt to pack cluster computing and reconfigurable wireless communication capabilities into a complete autonomous system, which is able to operate in critical environmental conditions, as those of the Antarctic region.

2.1 Environmental constraints

Extreme climatic conditions are the main characteristic of the polar environment, and their impact on electronic equipment is crucial. For an effective selection of the appropriate devices and precautions, data logged from a weather station close to the site where the prototype will be installed have been downloaded and analyzed [22], both for taking into account the main constraints and for evaluating the available resources in the area. Temperature, shown in Fig.2 (a), is the first aspect that must be considered, and in this area it spans from a minimum of $-35\,^{\circ}$C to a maximum of $+8\,^{\circ}$C. This means that all the electronic equipment must operate in extended temperature range, being the standard range from $0\,^{\circ}$C to $+40\,^{\circ}$C. Wind, as can be seen in Fig.2 (b), is always very strong in that area, being a very promising resource for wind power, but there is also a tangible risk of prejudicing the equipment's functionality. In this specific case, it was not feasible to develop a wind-power supply, because of the presence of a magnetic observatory nearby that could possibly suffer from interference issues. Also solar energy is an important source of supply for Antarctica, and it is commonly adopted because, despite having a dark period of some months in winter (see Fig.2 (c)), the solar radiation values are quite high during summer time. Therefore, this is the renewable power source that has been chosen for this specific application, specifically addressing a summer campaign.

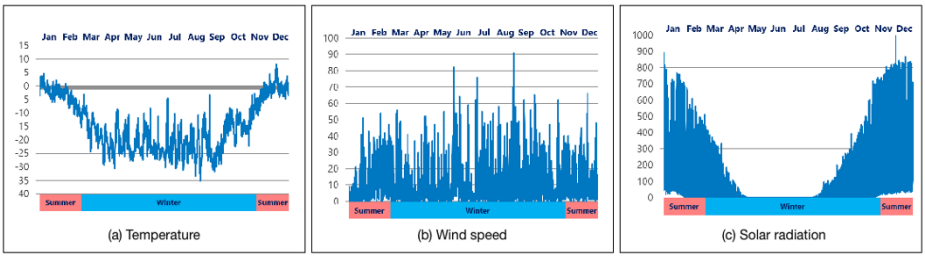


Fig. 2 Environmental conditions: (a) Hourly temperature over one year [°C], (b) Hourly wind speed over one year [knots], and (c) Hourly solar radiation over one year [W/m^2].

3 Design and architecture

As depicted in Fig.1, the prototype is composed by the following subsystems that will be described in next paragraphs:

- base station side: Reconfigurable Antenna, Front-End, Communication Receiver, Antenna controller, Power supply (provided by 230 VAC), Network Attached Storage (NAS);
- measurement station side: Reconfigurable Antenna, Front-End, Communication Transmitter, Antenna controller, Power supply (provided by Photovoltaic panels), Computing board, Logger board, Septentrio PolaRxS receiver and GNSS Antenna.

Power efficiency is the main concern of this ground disposable prototype so that, in the following paragraphs, the main logical components of the prototype architecture are described and the specific issues for the design of each building block are detailed. The final prototype implementation of the measurement station is shown in Fig.3 (a).

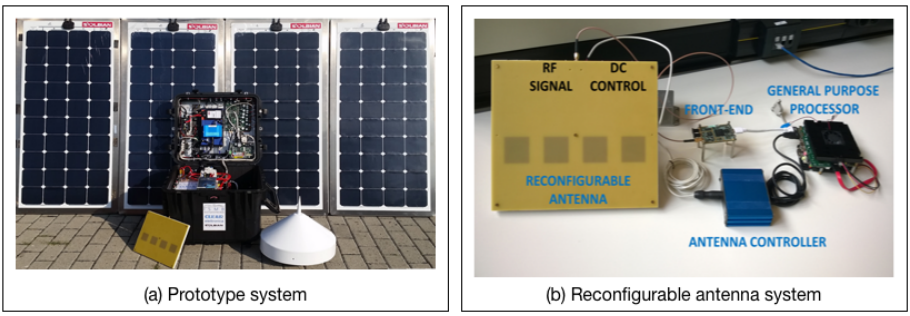


Fig. 3 Prototype: (a) integrated system of the measurement station, and (b) the reconfigurable antenna system.

3.1 Power management

A fine power management is of crucial importance for the measurement station, which is intended as a completely autonomous device (in particular from the energy provisioning point of view), powered by renewable sources (i.e. solar energy). Solar

power availability depends from many factors: weather conditions, latitude of the installation, yearly seasons and time of the day. Therefore, care should be taken in order to optimize the overall power efficiency of the entire system. In this paragraph, the architecture and behavior of Power Management Unit (PMU) are described in details. The PMU has been designed and developed by *CLEAR Srl* In Fig.4 the

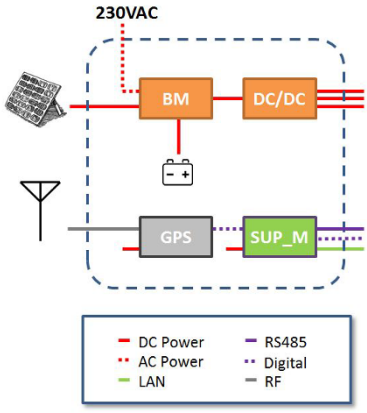


Fig. 4 Power supply unit.

power management architecture is depicted. The Battery Manager (BM) is in charge of extracting all the available electrical energy coming out from the photovoltaic panels and storing it into the battery stack. The BM implements also a Maximum Power Point Tracking (MPPT) algorithm in order to maximize the power extracted as the external conditions change (e.g., panel temperature, solar radiation, presence of shades on a subset of photovoltaic cells). The DC/DC converter module is in charge of extracting the power available from the battery, generating the voltage needed to power the systems. The GPS module in Fig.4 provides the absolute timing information to every component of the system architecture. An input power supply 230*VAC* is also included in order to be able to test the device in laboratory, without the need to connect photovoltaic modules.

The supervisor SUP_M is the module in charge of further optimizing the power consumption of the whole system. Such module feeds the electrical power to all the other components of the system (e.g., GNSS receiver, computing board, antenna) using energy saving policies. In fact, according to a series of preset operating profiles, the modules that are not used are safely switched off using the procedure in Algorithm 1.

Various operating profiles (duty cycles) can be set on each power line, in order to implement specific energy saving behaviors (e.g., the TX module of the measurement station can be powered on only when the acquisition and pre-processing of new GNSS data is completed). These profiles can be changed and stored on the fly using remote commands. In Table 1 the main electrical components used in PMU are reported.

Algorithm 1 Procedure to safely power off the slave modules

```
1: procedure SWITCH OFF PROCEDURE
2:     ShutdownCompleted ← 0                          ▷ Initial value for shutdown status flag
3:     SendCommand (SHUTDOWN)                         ▷ Send the SHUTDOWN command
4:     ResetTimeoutTimer(t)
5:     while ShutdownCompleted = 0 do
6:         GetCommand (c)                             ▷ Check for command answer
7:         GetPowerConsumption (p)                    ▷ Monitor the slave power consumption
8:         GetEvent (e)                               ▷ Check for timeout event
9:         if ((c == CLEAR_TO_SHUTDOWN) or
10:            (p < MIN_POWER_THRESHOLD) or
11:            (e == EVT_TIMEOUT)) then
12:            ShutdownCompleted ← 1                  ▷ Set shutdown status flag to true
13:        end if
14:    end while
15:    PowerOff()                                     ▷ Safely power off the slave module
16: end procedure
```

Table 1 Electrical components used in PMU

Component	Manufacturer	Description
AD8221	*Analog Devices*	*Precision Instrumentation Amplifier*
L76	*Quectel*	*GNSS Module*
STM32F407	*STM*	*ARM Cortex-M4 32b MCU+FPU, 210DMIPS, up to 1MB Flash/192+4KB RAM, USB, OTG HS/FS, Ethernet, 17 TIMs, 3 ADCs, 15 comm. interfaces & camera*
TEN40	*Traco Power*	*DC/DC converter, TEN 40N Series, 40 Watt*
THL20WI	*Traco Power*	*DC/DC converter, THL 20WI Series, 20 Watt*
THN30	*Traco Power*	*DC/DC converter, THN 30 Series, 30 Watt*
TMR9	*Traco Power*	*DC/DC converter, TMR 9 Series, 9 Watt*
IK0505SA	*XP POWER*	*DC/DC converter, IK Series, 0.25 Watt*

3.2 Communications

The ability to reconfigure an antenna may allow saving energy in autonomous systems as, for instance, the nodes of a deployable wireless sensor network [23]. The fundamental concept relies on saving energy radiated in unwanted directions. This is commonly achieved by means of directional antennas, but the real advantage of reconfigurable antennas is their capability to focus their energy toward the intended direction dynamically, without requiring a mechanical movement. Generally, this is accomplished by varying the voltage of internal components of the antenna (e.g., the beamformer). Since less transmission power will be required, this leads to a more energy-efficient development [24]. An overview of the communication system is reported in Fig.3 (b), which presents the basic components employed on both base and measurement stations. The realized antenna is a 4x1 array connected to a Radio-Frequency (RF) beamformer that, by means of voltage controllable phase, is able to steer the main beam. These four signals are summed within the RF beamformer and the resulting signal is downconverted to baseband and discretized by a single FE. The digital signal is then processed by a GPP which executes the software implementation of the employed communication standard and the algorithm controlling the antenna, that scans the space to search for the LoS. Both software are fully optimized to be fitted in low power boards. The controller is a PC-to-Antenna beamformer interface which essentially converts the digital input in to a voltage with the purpose of controlling the position of the main beam.

The link has been optimized for low power point-to-point communication. It derives from the idea of placing the measurement station in an arbitrary position without knowing a-priori the base station location. When the measurement station needs to transmit the collected data, by means of the reconfigurable antenna controlled via

SDR, it aligns the main beam autonomously with the one of base station. Furthermore, thanks to the modular structure, a power supply controller switches off the transmitter module of the measurement station when it has finished to send data. This allows a fine-grain control on the power consumption of the communication module which requires peak power when transmitting and zero consumption in the idle time. A different situation applies to the base station: in fact the availability of AC power supply allows this side to be always on. The Base station is responsible to transmit a carrier signal to be found by the measurement station and waiting to receive the data transmitted from the latter.

3.3 Computing

The computing module is devoted to the pre-processing of raw data collected from a Septentrio PolaRxS receiver, and it is activated by the logger module whenever a given amount of raw data are received, then it is powered off immediately after the elaboration in order to optimize power consumption. This module allows to execute several analyses based on the type of data retrieved. One of these analyses is surely related to the ionosphere which is the single largest contributor to the GNSS (Global Navigation Satellite System) error budget, and ionospheric scintillation in particular is one of its most harmful effects. The Ground Based Scintillation Climatology (GBSC), developed by INGV [25], can ingest data from high sampling rate GNSS receivers for scintillation monitoring like the widely used GISTM (GPS for Ionospheric Scintillation and PolaRxS). Each analysis needs a predefined elaboration environment with particular software and libraries, so in order to separate each environment Docker tool [26] has been implemented. Docker is a new high-level tool which automates the deployment of applications inside software containers and provides an additional layer of abstraction and automation of operating system-level virtualization. It is designed to help delivering applications faster by using a lightweight container virtualization platform surrounded by a set of tools and workflows that help developers in deploying and managing applications easier. The idea is to associate each application with a container, where these containers can be loaded on to the main board and then upload and download to different locations. Containers are basically self-contained executable applications. To this end, the developer of the application is in charge of collecting all the required dependency components and libraries, and pack them within the application itself. Docker is built on top of LXC Linux containers API, and its main feature is that it offers an environment (in terms of functionalities and performance) as close as possible to a traditional Virtual Machina (VM), but without the overhead of running a whole separate kernel. In addition, it allows to expose kernel features used to simulate the presence of dedicated hardware resources, as well as a full OS (guest OS) with its own memory management installed with virtual device drivers, and controlled by a hypervisor, which works on the top of the host OS. By adding low overhead over host machines, containers perform better than other virtualization methods based on traditional hypervisors, such as KVM and Xen. Unlike a VM, which runs the full operating system, a container can be even only a single process.

4 Results

Several tests have been done on each building block of this project. First, each part has been tested separately from the others, then they have been tested together to define an internal communication protocol able to manage all the modules composing

the complete prototype. The following sections illustrate the results obtained from the functional tests on: *i*) sizing of the renewable energy components, taking into account the prototype consumption; *ii*) tests in the environmental chamber, bringing the instrumentation to extreme conditions; and *iii*) communication module, in particular antenna reconfigurability and LoS data transmission.

4.1 Sizing of the renewable energy components

In order to correctly carry out the sizing of the PMU, the available data logged from the above mentioned weather station [22] has been considered. As a first step, the local radiance has been measured all over one year (Fig.2 (c)). This information is important to achieve the correct dimensioning of the photovoltaic panels to be installed on the site. The system should be dimensioned in the worst case period, therefore the months of November and March have been selected as the dimensioning temporal points. It is worth noting that, in the energy budget evaluation process, an optimized operating profile has been considered in order to achieve the correct dimensioning in the real operating conditions (i.e., suitable power on-off duty cycle has been considered for the slave components that don't need to be always powered on).

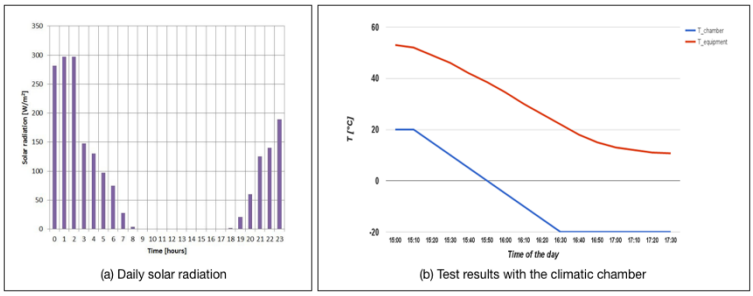


Fig. 5 Daily solar radiation (March 1st) used to dimensioning the solar panels (a); and tests of the equipment in the environmental chamber: temperature comparison between chamber and equipment box [°C] (b).

Considering the above mentioned periods, the typical daily radiance trend is depicted in Fig.5 (a). The overall daily radiation for the March 1st is $1706 W/m^2$ per day, which corresponds to 1.7 equivalent hours[1]. A single photovoltaic module[2] is characterized by $125 Wp$ peak power, therefore it produces:

$$E_{day} = h_{eq} \cdot P_{STC} = 212.5 Wh$$

The mean electrical load power requested by the selected operating profile is about $30 W$, with a peak value of $50 W$. This mean power corresponds to $720 Wh$ of requested energy per day. Consequently, it is possible to calculate the minimum number of photovoltaic modules requested for the system:

$$N_{panels} > \frac{E_{eff}}{E_{load}} = 3.38$$

[1] Considering $1000 W/m^2$ radiation in STC conditions.

[2] SOLBIAN FLEX SP125 model.

Rounding this value to the next larger integer gives the minimum number of required panels. This rounding operation also takes into account the losses due to the DC/DC converters and the efficiency of the energy storage unit (i.e., the battery stack). We finally selected 4 panels of the above mentioned photovoltaic module.

For the sizing of the battery stack, an autonomy of 3 days without renewable sources has been considered: this corresponds to $2160Wh$ of stored energy ($30W$ per $72h$). The LiFeYPO4 battery from GWL Power [3] is characterized by a $320Wh$ capacity, it is therefore possible to calculate the minimum number of battery cells requested for the system, which is:

$$N_{batteries} > P_{load} \cdot \left(\frac{h}{C}\right) = 30W \cdot \frac{72h}{320Wh} = 6.75$$

Considering the losses due to the energy storage unit, we selected 8 cells of the above mentioned battery model.

4.2 Tests in the environmental chamber

The equipment selected for the prototype has been tested in an environmental chamber with a twofold goal:

1. a preliminary validation of the full functionality of the single building blocks of the architecture when brought to low temperatures;
2. an evaluation of the internal temperature of the enclosure where the hardware is placed, when external conditions are varied.

As for the first objective, this was assumed to be intrinsically guaranteed by the specification of the single devices selected. On the other hand, the coexistence of several devices, switched on and running in a single box, could drive to higher temperatures within the box, and possibly to overheating conditions. As can be seen in Fig. 5 (b), the internal temperature of the box retraces the external condition, with a considerable and almost constant gap of $+30°C$. This is a positive remark for the polar application specifically addressed, but should be taken into consideration for possible future extensions in different environmental conditions.

Considering the output of the previous test, where the equipment was switched on for the whole duration, a further assessment has been done: the environmental chamber has been set to the minimum temperature of $-20°C$ for 3 hours, while the devices under test were switched off. After this period of time, the equipment was switched on in order to see if every device was able to start working at low temperature without being warmed by other ones nearby. Also in this case, the test was successful, giving a positive feedback on the real working conditions in the final test-bed.

4.3 Base station detection and data transmission

The settings for the reconfigurability test are illustrated in Fig. 6 (a). Base station is always listening and waiting for data, while transmitting a carrier signal with the purpose of being revealed.

[3] The battery family based on LiFeYPO4 technology has been selected because it is particularly suitable for harsh environmental conditions.

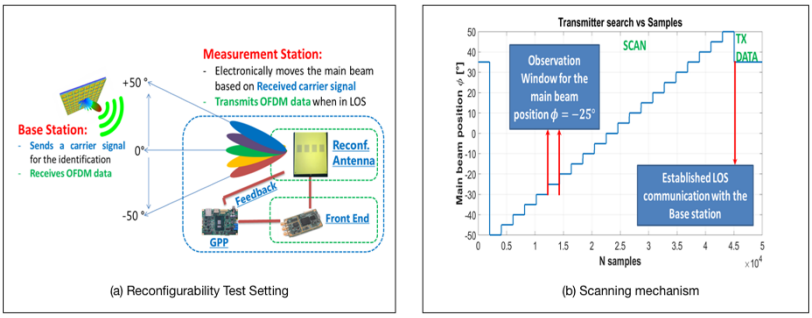


Fig. 6 Reconfigurability Test Settings (a), and the Scan and Lock Algorithm (b).

When measurement station is ready to transmit collected data, its communication module is switched on, and before starting the transmission the antenna controller algorithm computes the received power for each position in which the main beam can be steered (e.g., Fig. 6 (b), executes a scan that ranges from $-50\,°C$ to $+50\,°C$ with a step of $5\,°C$).

After the best position is detected (signal is the strongest), the LoS condition is established and the measurement station sends the pre-processed GNSS data to the base station. The scan and lock mechanism reported in Fig. 6 (b), is repeated every time a transmission is required. Scan time can also be reduced, since it mainly depends on sampling frequency, observation windows for the power estimation and scan step. Thanks to software flexibility, they can be adjusted based on needs.

5 Conclusions and future work

An advanced prototype of a low power computing and communication system, self-sustained by means of renewable energy and advancements in communication, has been designed and realized for being deployed in the Antarctic continent within the scope of the Italian National Research Program for Antarctic Research (PNRA).

The system is meant to support the collection and transmission of scientific data in field, being easily movable and suitable to extreme environmental conditions. All the constituting components of the prototype have been selected to be robust to low temperatures, and tested in an environmental chamber before the on-site campaign; this has allowed to recreate the temperature conditions that characterize the Antarctic summer time, when the system is intended to be operative. Communication between measurement station and base station has been successfully tested outdoors, at various transmission rates, through the approximation of the positions of the two stations in a scenario as close as possible to the field set-up.

Next steps include the field tests of the system in the Antarctic region, which is also a crucial step to collect feedback that will allow a fine tuning of the single building blocks, as well as possible improvements of the whole system. The system's flexibility will also be enhanced, leveraging its modularity to work on different scenarios (e.g., changing the deployment site will drive to different constraints on hardware and renewable sources), preserving the current features already implemented for this experimentation.

6 Acknowledgements

The authors are grateful to the PRNA-Programma Nazionale di Ricerche in Antartide for supporting the project "Upper atmosphere observations and Space Weather" within PNRA D.C.D. 393 del 17/02/2015 PNRA14_00110 - Linea A1 and to all contributors: CLEAR elettronica Srl, SOLBIAN energie alternative Srl, INGV and ENEA.

References

1. Y.J. Guo and Pei-Yuan Qin. Advances in reconfigurable antennas for wireless communications. In *9th European Conference on Antenna and Propagation*, April 2015.
2. S. S. Jeng and C. W. Tsung. Performance evaluation of ieee 802.11g with smart antenna system in the presence of bluetooth interference environment. In *2007 IEEE 65th Vehicular Technology Conference - VTC2007-Spring*, pages 569–573, April 2007.
3. Wenjiang Wang, Sivanand Krishnan, Khoon Seong Lim, Aigang Feng, and Boonpoh Ng. A simple beamforming network for 802.11b/g wlan systems. In *Communication Systems, 2008. ICCS 2008. 11th IEEE Singapore International Conference on*, pages 809–812, Nov 2008.
4. D. C. Chang and C. N. Hu. Smart antennas for advanced communication systems. *Proceedings of the IEEE*, 100(7):2233–2249, July 2012.
5. M. Uthansakul and P. Uthansakul. Experiments with a low-profile beamforming mimo system for wlan applications. *IEEE Antennas and Propagation Magazine*, 53(6):56–69, Dec 2011.
6. A. Hakkarainen, J. Werner, K. R. Dandekar, and M. Valkama. Widely-linear beamforming and rf impairment suppression in massive antenna arrays. *Journal of Communications and Networks*, 15(4):383–397, Aug 2013.
7. K. T. Jo, Y. C. Ko, and Hong-Chuan Yang. Rf beamforming considering rf characteristics in mimo system. In *2010 International Conference on Information and Communication Technology Convergence (ICTC)*, pages 403–408, Nov 2010.
8. A. S. Prasad, S. Vasudevan, R. Selvalakshmi, K. S. Ram, G. Subhashini, S. Sujitha, and B. S. Narayanan. Analysis of adaptive algorithms for digital beamforming in smart antennas. In *Recent Trends in Information Technology (ICRTIT), 2011 International Conference on*, pages 64–68, June 2011.
9. et al. Prikryl P. Gps phase scintillation at high latitudes during geomagnetic storms of 7-17 march 2012 - part 2: Interhemispheric comparison. *Annales Geophysicae*, 2015.
10. et al. Prikryl P. An interhemispheric comparison of gps phase scintillation with auroral emission observed at the south pole and from the dmsp satellite. *Annals of Geophysics*, 2013.
11. Raspberry pi. `https://www.raspberrypi.org`. Accessed: 2016-09-07.
12. Arduino single board computer. `https://www.arduino.cc`. Accessed: 2016-09-07.
13. The parallella board. `https://www.parallella.org`. Accessed: 2016-09-07.
14. D. Richie, J. Ross, S. Park, and D. Shires. Threaded mpi programming model for the epiphany risc array processor. *Journal of Computational Science*, 9:94 – 100, 2015.
15. Hpe moonshot system. `https://www.hpe.com/us/en/servers/moonshot.html`. Accessed: 2016-09-07.
16. Fips project. `https://www.fips-project.eu/wordpress/`.
17. Roberto Giorgi. Scalable embedded systems: Towards the convergence of high-performance and embedded computing. In *EUC-2015*, 2015.
18. E. Palazzetti. *Getting Started with UDOO*. Packt Publishing, 2015.
19. Simon J. Cox, James T. Cox, Richard P. Boardman, Steven J. Johnston, Mark Scott, and Neil S. O'Brien. Iridis-pi: a low-cost, compact demonstration cluster. *Cluster Computing*, 17(2):349–358, 2014.
20. Yiran Zhao, Shen Li, Shaohan Hu, Hongwei Wang, Shuochao Yao, Huajie Shao, and Tarek Abdelzaher. An experimental evaluation of datacenter workloads on low-power embedded micro servers. *Proc. VLDB Endow.*, 9(9):696–707, May 2016.
21. Sheikh Ferdoush and Xinrong Li. Wireless sensor network system design using raspberry pi and arduino for environmental monitoring applications. *Procedia Computer Science*, 34:103 – 110, 2014.
22. Climantartide web site. `http://www.climantartide.it/`. Accessed: 2016-09-07.
23. M. Orefice G. Dassano. Voltage controlled steerable array for wireless sensors networks. In *2nd European Conference on Antennas and Propagation EuCAP*, November 2007.
24. Constantine A. Balanis. *Antenna Theory - Analysis and Design*. Wiley, third edition, 2005.
25. O. Terzo L. Spogli L. Alfonsi V. Romano A. Scionti, P. Ruiu. Demograpе: Managing scientific applications in a cloud federated environment. In *CISIS-2016*, 2016.
26. Docker. `https://www.docker.com/`. Accessed: 2016-09-07.

Risk Management Framework to Avoid SLA Violation in Cloud from a Provider's Perspective

Walayat Hussain[*], Farookh Khadeer Hussain[*], Omar Khadeer Hussain[‡]

[*]School of Software
Decision Support and e-Service Intelligence Lab
Centre for Quantum Computation and Intelligent Systems
University of Technology Sydney, Sydney, New South Wales 2007, Australia

[‡]School of Business, University of New South Wales Canberra,
*walayat.hussain@ uts.edu.au; farookh.hussain@uts.edu.au, ‡o.hussain@adfa.edu.au

Abstract. Managing risk is an important issue for a service provider to avoid SLA violation in any business. The elastic nature of cloud allows consumers to use a number of resources depending on their business needs. Therefore, it is crucial for service providers; particularly SMEs to first form viable SLAs and then manage them. When a provider and a consumer execute an agreed SLA, the next step is monitoring and, if a violation is predicted, appropriate action should be taken to manage that risk. In this paper we propose a Risk Management Framework to avoid SLA violation (RMF-SLA) that assists cloud service providers to manage the risk of service violation. Our framework uses a Fuzzy Inference System (FIS) and considers inputs such as the reliability of a consumer; the attitude towards risk of the provider; and the predicted trajectory of consumer behavior to calculate the amount of risk and the appropriate action to manage it. The framework will help small-to-medium sized service providers manage the risk of service violation in an optimal way.

Keywords: Risk management framework; cloud computing; SLA violation prediction.

1 Introduction

The flexibility, scalability and wide range of computing services using a "pay as you go" model has made cloud computing one of the most promising technologies in today's world. Both large and small enterprises benefit from the scalability of virtualized resources and the opportunities created by large-scale complex parallel processing at no upfront cost. Cloud computing also reduces the cost of support and maintenance, infrastructure, automation and computerization by better managing organizational income and outgoings, and eliminating tiresome activities [1, 2]. As a result of these benefits, many organizations have leveraged a cloud-computing platform, and this has created both new opportunities and new challenges for cloud providers.

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_22

Service level agreement (SLA) management is among one of the important issues a cloud provider cannot ignore [3]. An SLA is a primary legal document that defines a course of action between a consumer and a provider. In order to avoid SLA violation, a service provider needs viable SLA formation and a management model that intelligently executes an SLA with a consumer based on their resource usage history [4]. In our previous work [5] we proposed a profile-based viable SLA violation management model comprising two time phases – a pre-interaction time phase and a post-interaction time phase. The pre-interaction phase comprises an identity manager module and a viable SLA module, responsible for forming a viable SLA and assigning the amount of resources based on their transaction trend. In this paper, we propose the modules of the post-interaction phase that are responsible for predicting the likely resource usage of a consumer, forming the threshold, and possible risk management when the system predicts a likely violation.

Our approach for SLA violation management in the post-interaction start time phase is based on the notion of Risk, which is a ubiquitous factor associated with many of business transactions notable for its negative occurrence due to external or internal vulnerabilities [6]. Risk has the capacity to change the results of an interaction in an adverse way; therefore it is crucial to manage it as it leads to causing negative results. When both parties execute an SLA, our proposed approach which we term as the Risk Management Framework to avoid SLA violation (RMF-SLA) has a prediction module that forecasts the likely resource usage behavior of the consumer over a given future period of time. When the predicted results exceed a pre-defined threshold, the risk management module of RMF-SLA is activated. This module considers the provider's attitude toward risk, the reliability of the consumer, and the predicted trajectory of the consumer's use of resources to determine either to accept or not accept the risk of SLA violation and thereby recommends an appropriate action to manage it.

The rest of the paper is organized as follows. Section 2 discuss about related literature. In Section 3 we describe our proposed risk management based RMF-SLA framework. In Section 4 we describe the applicability of the framework in managing SLA violations and Section 5 concludes the paper.

2 Related Work

Albakri et al. in [7] proposed a security risk assessment model, including both a cloud provider and a cloud consumer, to manage the risk of SLA violation. Involving both the provider and the consumer in risk assessment allows best-case evaluation. Based on ISO27005, the authors divided their framework into six modules. Risk assessment begins by establishing information security risk management. After establishing a context, the risk is assessed, by both the consumer and provider, to determine the potential for loss and the reason. After assessment, the risk is treated and then

becomes the definition of the risk acceptance criteria. The proposed model is developed to work on SaaS model. Zhang et al. [8] proposed a risk management model based on ISO/IEC 27001 and NIST management guidelines, to identify the threats and vulnerabilities related to the cloud computing environment, however, the approach did not consider the notion of risk, proactive risk detection or its early remedy. Majd and Balakrishnan in [9] proposed a trust model based on four components, namely consistency, likeness, trust logical relation and satisfaction for provider selection. They used fuzzy logic to identify these components and applied a TOPSIS method to select the most reliable provider. They compared their approach with existing approaches and demonstrated the effectiveness of their proposed model. Morin et al. [10] identified the risks and challenges linked with cloud computing. They proposed an SLA risk management framework that helps to manage risk in runtime cloud computing environments to enhance governance, risk and compliance. Wang et al. in [11] proposed a quantitative risk assessment model that defined risks, assets, vulnerabilities and made a quantitative calculation to estimate the possibility of risk occurrence. The model quantifies risk based on the existing best practice approach. In other work, attack-defense tree (ADT) method was used by Wang et al. [12] to handle the threat and risk analysis in cloud computing environments. The ADT method considers the attack cost of vulnerabilities and the defense cost for defensive actions. The model calculates the defensive cost based on the attacking cost with a ratio of success. The model uses Bayesian network theory to determine the amount of risk and help defenders take appropriate action, however the suitability of the framework depends on the availability of authentic statistical data and the selection of a proper defense method.

From the above discussion, it is seen that approaches in the literature identify risks, vulnerabilities and the necessary actions required for proper risk management in cloud computing environment. However, that analysis is from different viewpoints of risk. From the viewpoint of a cloud service provider's (particularly a small and medium cloud service provider) perspective various research gaps particularly in SLA management by considering risk exist. These gaps relate to the proper identification and estimation of risks and then developing a decision system that help the service provider to take an appropriate action to mitigate the risk of SLA violation. We propose such an approach in this paper that will assist the cloud service provider in SLA management by considering the notion of risk. Our proposed framework utilizes a Fuzzy Inference System (FIS) by considering three inputs - reliability of consumer, risk attitude of the provider and predicted trajectory to generate an output of an appropriate action that the cloud provider should take to avoid the risk of SLA violation. In next section we describe our risk management framework for SLA management in detail.

3 RMF-SLA Framework to manage SLA Violation in Post-Interaction Time-Phase

In our previous work [5], we proposed a viable SLA model to assist cloud providers in forming and executing SLAs with consumers. The viable SLA framework monitors the SLA both at pre-interaction and post-interaction time phases. When both parties execute their SLAs, then in the post-interaction time phase a provider needs a SLA management framework that accurately predicts likely service violations due to the excess resource usage of its consumers and manages the risk of it accordingly. Our proposed risk management framework RMF-SLA as presented in Figure 1 achieves that by using five modules as follows:

a) Threshold formation module (TFM)
b) Runtime QoS monitoring module (RQoSMM)
c) QoS prediction module (QoSPM)
d) Risk identification module (RIM)
e) Risk management module (RMM)

The description of each module is presented in below sections:

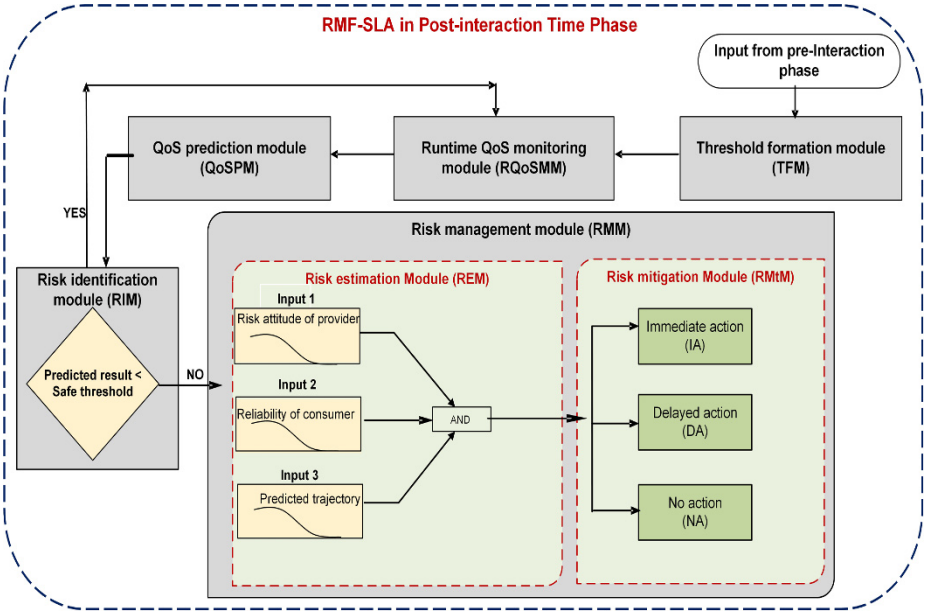


Figure 1: Risk management framework to avoid SLA violation (RMF-SLA)

3.1 Threshold formation module (TFM)

This is the first module in RMF-SLA, which define the threshold value according to which SLA violation management is carried out. For informed SLA management, we propose that a provider defines two thresholds, namely Agreed threshold value (T_a) and Safe threshold value (T_s).

T_a is the threshold value that the provider and consumer agreed to when finalizing and forming their SLA. T_s is provider-defined customized threshold value which is set to alert the provider when it finds the runtime QoS parameters exceeds the T_s value and invoke risk management module to take necessary action to manage the risk of SLA violation. To explain with an example, let us suppose that a consumer and a provider both agreed to a maximum processing time of 90 seconds for the request "Copy Blob". The provider sets a safe threshold by making that threshold stricter with a maximum processing time of 70 seconds. The criterion of 90 seconds is T_a and the criterion of 70 seconds is T_a.

3.2 Runtime QoS monitoring module (RQoSMM)

The second module in RMF-SLA is runtime QoS monitoring module that takes the runtime resource usage behavior of a consumer and passes it to the QoS prediction module to predict the future QoS and for comparison with the formed threshold value.

3.3 QoS prediction module (QoSPM)

This module is responsible for predicting the future resource usage behavior of the consumer. In our previous work [13] we compared different prediction algorithms on real cloud datasets. We considered the different prediction methods, including stochastic and neural network methods, on a time series real cloud dataset from Amazon EC2 and found that the ARIMA method gave most optimal results among all the methods. For an optimal prediction result, the prediction method considers all previous observations including the most recent data from the Runtime QoS monitoring module to predict the future intervals. By considering the value from RQoSMM, QoSPM is able to recalibrate the prediction results for better accuracy.

3.4 Risk identification module (RIM)

The module compares the output of the QoS prediction module with the T_s value. If the value of predicted result exceeds the T_s value, then it invokes the risk management module to assess the level of risk and take appropriate necessary actions to mitigate the risk of SLA violation.

3.5 Risk management module (RMM)

This module estimates the risk of an SLA violation occurring and recommends plans for mitigating them to avoid SLA violation. RMM is composed of two sub modules, namely the risk estimation module and risk mitigation module. The modules are described below:

a) *Risk estimation module (REM):* REM is responsible for estimating the extent of risk of SLA violation occurring and assists the provider in making their decision to either accept or decline the risk. The module uses a fuzzy inference system with three inputs namely the reliability of the consumer, the provider's attitude toward risk, and the predicted trajectory of the SLO being considered over a future period of time to determine the level of risk and appropriate actions to mitigate SLA violation.

b) *Risk mitigation module (RMtM):* Based on the output of the risk estimation module, this module recommends appropriate action to be taken to mitigate the risk. In high-risk cases, a provider needs to take immediate action and/or arrange deficient resources to avoid SLA violation. Additionally, in such case the provider may have to stop taking new requests until the risk is managed. If a risk is identified as a medium-risk, then the provider needs to take action within a certain time period to eliminate the risk, and when the risk is identified as a low-risk, the provider accepts the risk without taking any action. These actions to be taken upon the determined level of risks are provider dependent and can vary.

4 Evaluation of RMF-SLA to avoid the risk of SLA violation

To validate the proposed approach, we collected a real cloud dataset from Amazon EC2 EU [14] by using PRTG service [15]. For the purposes of validation, we consider that the collected dataset represents the usage of a consumer in a SLO. These input values are monitored by RQoSMM and fed into QoSPM for their prediction over a future period of time. The next step is the RIM where the predicted QoS value is compared with the define safe threshold and the appropriate course of action determined as explained below:

a) Scenario 1: Predicted trajectory exceeds the defined safe threshold (T_s) and moving towards the agreed threshold (T_a)

The simulation for the first scenario is run by using the QoS values on 3rd February 2016 from 1:45AM to 12:30PM. The output of the prediction and defined safe thresholds are as presented in Figure 2.

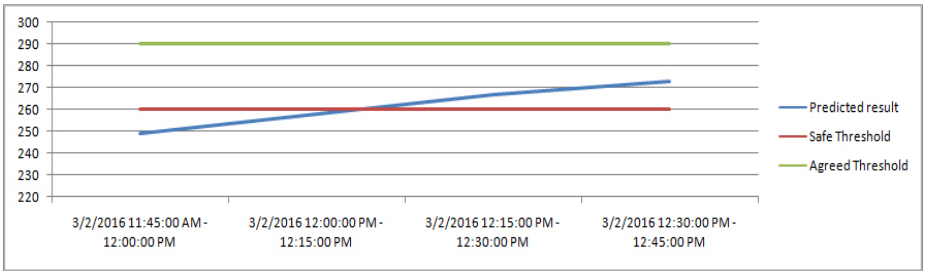


Figure 2: Scenario 1 predicted results, safe and agreed thresholds

The red line represent the defined T_s, the green line represents the agreed T_a and blue line represents the predicted trajectory from QoSPM which has exceeded the T_s and moving towards T_a. Let us suppose that the reliability value of the consumer being monitored is 85 out of 100, and risk attitude of the provider is a value 3 out of the scale of 5. The values are input to the FIS along with the direction of the predicted trajectory (which in this case is towards T_a). According to the defined fuzzy rules the FIS recommends the provider to take an immediate action to mitigate the risk at the earliest possible time as the blue line is moving towards the agreed threshold.

b) Scenario 2: Predicted trajectory exceeds the defined safe threshold (T_s) but moving towards the agreed threshold (T_a)

The simulation for the first scenario is run by using the QoS values on 3rd February 2016 from 12:45AM to 1:30PM. The output of the prediction and defined safe thresholds are as presented in Figure 3. In this case, the reliability of consumer is considered as 20 out of 100 and the risk attitude of a provider as 3 on a scale of 5.

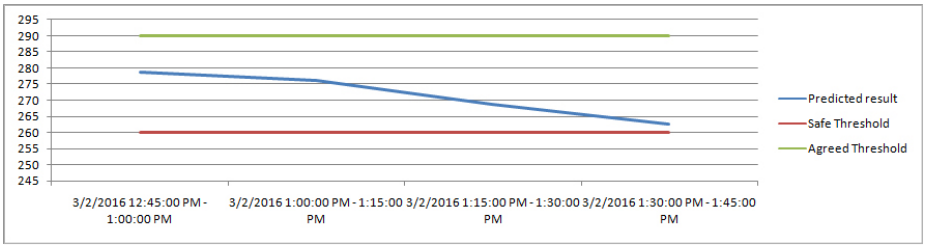


Figure 3: Scenario 2 predicted results, safe and agreed thresholds

From Figure 3, we can observe that although the predicted trajectory exceeds the T_s, however it is moving away from the T_a and coming back towards the T_s value. When above input values are given to FIS then according to the fuzzy rules an output of delayed action is recommended. The delayed action is in spite of the reliability of a user being low but as the predicted value is moving towards the safe threshold away from the agreed one, there is a greater chance of a SLA violation not happening if the situation remains same. However, in this case if the risk attitude of the provider was

risk averse then the output from the system differs and it may recommend for immediate action.

From above scenarios, we can see that depending on the reliability of consumer, risk attitude of the provider and the predicted trajectory RMF-SLA recommends the service provider to take one of the following actions from immediate action, delayed action or no action and avoid SLA violations.

5 Conclusion

The elastic nature of the cloud frees cloud consumers from the scalability issues of computing resources, and SLAs are key agreements that define all business terms and obligations in these business arrangements. Service providers need to manage the risk of potential violations on the defined SLAs. This work-in-progress addresses the issue of avoiding SLA violations for small size and medium size cloud providers through risk management to help service providers better manage their SLAs. The proposed RMF-SLA considers the reliability of the consumer; the provider's attitude toward risk; and the predicted trajectory of SLO usage to decide whether to accept the risk or not and, if the risk is not accepted, what actions are necessary to be taken in order to avoid SLA violation. In our future work we will implement and validate our framework at a provider to manage their SLA assurance real-time.

References

1. Khajeh-Hosseini, A., D. Greenwood, and I. Sommerville. *Cloud migration: a case study of migrating an enterprise IT system to IaaS*. in *Cloud Computing (CLOUD), 2010 IEEE 3rd International Conference on*. 2010. IEEE.
2. Hashem, I.A.T., et al., *The rise of "big data" on cloud computing: Review and open research issues.* Information Systems, 2015. **47**: p. 98-115.
3. Hussain, W., F.K. Hussain, and O.K. Hussain. *Maintaining Trust in Cloud Computing through SLA Monitoring*. in *Neural Information Processing*. 2014. Springer.
4. Hussain, W., F.K. Hussain, and O. Hussain, *Comparative analysis of consumer profile-based methods to predict SLA violation*, in *FUZZ-IEEE*, IEEE, Editor. 2015, IEEE: Istanbul Turkey.
5. Hussain, W., et al., *Profile-based viable Service Level Agreement (SLA) Violation Prediction Model in the Cloud*, in *2015 10th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC)*. 2015, IEEE: Krakow, Poland. p. 268-272.
6. Hussain, O.K., et al., *A methodology to quantify failure for risk-based decision support system in digital business ecosystems.* Data & knowledge engineering, 2007. **63**(3): p. 597-621.

7. Albakri, S.H., et al., *Security risk assessment framework for cloud computing environments.* Security and Communication Networks, 2014. **7**(11): p. 2114-2124.
8. Zhang, X., et al. *Information security risk management framework for the cloud computing environments*. in *Computer and Information Technology (CIT), 2010 IEEE 10th International Conference on*. 2010. IEEE.
9. Majd, E. and V. Balakrishnan, *A trust model for recommender agent systems.* Soft Computing, 2016: p. 1-17.
10. Morin, J.-H., J. Aubert, and B. Gateau. *Towards cloud computing SLA risk management: issues and challenges*. in *System Science (HICSS), 2012 45th Hawaii International Conference on*. 2012. IEEE.
11. Wang, H., F. Liu, and H. Liu, *A method of the cloud computing security management risk assessment*, in *Advances in Computer Science and Engineering*. 2012, Springer. p. 609-618.
12. Wang, P., et al. *Threat risk analysis for cloud security based on Attack-Defense Trees*. in *Computing Technology and Information Management (ICCM), 2012 8th International Conference on*. 2012. IEEE.
13. Hussain, W., F.K. Hussain, and O. Hussain, *QoS Prediction Methods to Avoid SLA Violation in Post-Interaction Time Phase*, in *11th IEEE Conference on Industrial Electronics and Applications (ICIEA 2016)* 2016, IEEE: Hefei, China.
14. CloudClimate. *Watching the Cloud*. Available from: http://www.cloudclimate.com/.
15. Monitor, P.N.; PRTG Network Monitor]. Available from: https://prtg.paessler.com/.

Enhancing Video Streaming Services by Studying Distance Impact on the QoS in Cloud Computing Environments

Amirah Alomari, Heba Kurdi

Computer Science Department
College of Computer & Information Sciences
King Saud University
Riyadh, SA

Abstract. Nowadays, videos are extensively used in different aspects of our lives such as entertainment, education and social networking. Therefore, video sharing is expected to dramatically grow in the future. There are important factors that affect the way of how to transmit videos over the clouds such as encoding techniques and compressions. In this paper, we study how distance affects QoS of real time applications such as video streaming services by means of simulation. The results show that distance has a significant impact on response time and packet end to end variation. However, throughput is not affected by far distances. In addition, the impact varies from one application to another as well for the same application.

Keywords. Cloud computing; video streaming; quality of service

1. Introduction

Cloud computing is a model to provide on-demand computing resources where usage is metered and these resources are provisioned in advance. Characteristics of cloud computing are on-demand services such that clients can request more resources whenever is needed and scale them down when it is no longer required. Resources pooling since clouds providers offer unlimited and different available resources for consumers such as storage, processing, and bandwidth services. Measured services where the user pay for what is used. Also elasticity and broad network access are considered as main characteristics. Hence, cloud computing facilitate consumers` work and reduce the cost of maintenance of traditional IT resources. [1].

Quality of Service (QoS) can be defined as the nonfunctional characteristics of the offered service. Due to the increasing number of cloud providers, quality of service is a main differentiation, especially for data intensive applications such as video streaming services. Cloud providers should insure high quality in order to fulfill customers` expectation and requirements. Therefore, quality of service is the main aspect that determines end user`s satisfactions. However, QoS is a major challenge in cloud computing environment, where resources allocating need to be done in such a way that overall performance is maintained. Quality of Service represents

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_23

performance, reliability and availability of the service in cloud computing environment. While its parameters vary from one application to another based on the nature of the service itself [2].

Video Streaming service is the service that provides watching videos online rather than downloading it. Videos sent in a compact form and the end user can watch it as it appears. People tend to watch videos online rather than downloading it. Studies show that more than 90% of internet traffic is caused by videos transmitting [3]. That is why, multimedia streaming applications are most critical to QoS constraints because it needs to be delivered in smooth way without disturbing clients [4]. Therefore, QoS measures are important to evaluate the service offered by providers. However, there are no standard measures for quality of service in cloud computing.

In order to help providers to ensure high quality of service for their consumers, this study analyzed distance impact on video streaming quality, which can help cloud providers to target consumers based on geographical locations.

Many studies stated number of factors that have an impact on video streaming services. Codec technique and compression are main factor that affect the videos quality. Bandwidth of networks plays major role when choosing the encoding method to transfer uncompressed videos, for example enterprise network with 2 – 5 Mbps uses H.246 video codec which use lossless and lossy compression methods in order to reach small bitrate videos [5]. However, most of the factors are not possible to take into consideration because of limited features of Riverbed Modular Academic Edition. As well some factors are not possible to simulate. Therefore, this study took into consideration response time, packet end to end delay and the net throughput. Therefore, this study took into consideration the following parameters:

- Response time
- Packet end to end delay
- Throughput

2. Literature Review

In [6] the paper purposed an experimental study by determining quality parameters: response time and requests time out. The experimental setup was built as private cloud architecture and they implemented HTTP-based application. They conclude that, the average response time and requests time out, are increased as the number of user is increased. While in [7] the paper studied the relations between latency and QoS parameters such as throughput and jitter in different geographical servers that are placed in different universities in Malaysia, Denmark, Brazil and Boland. Three parameters were studied: throughput, jitter, and delay. Authors monitor these parameters changes during 24 hours and smoothing functions were applied to throughput and delay. The result of their study was that there wasn't a fixed indicator if one parameter is changed the other parameter will change as well. This is because of parameters are not dependent thus their changes are not related. However, in some cases, there was positive relation between latency and file transfer time. In the other hand, jitter and latency were changing dependently. It is important before choosing time and place of a service in the cloud, to predict network QoS.

In [8] the paper reviewed important factors that play role when streaming videos over the clouds which are: storage, streaming, security and quality. Also it overviewed the characteristics that influent each of those factors. Nowadays, videos are extensively used in different aspects of our lives such as entertainment, education and social. Therefore, there is expected huge growth of video sharing in the future. There are important factors that affect the way of how to transmit videos over the clouds will be explained. The first factor is the storage which depends on availability, reliability, scalability and continuity. That's why cloud is the best framework to share videos through because it complies with storage characteristics. The second factor is streaming, which depends on video encoding technique, which has been determined to best use MP4 format, streaming protocols, which is advisable to use User Datagram Protocol (UDP) because it eliminate the delay compared with Transmission Control Protocol (TCP). The third factor is quality of service which depends on the network bandwidth allocation. The last factor is security where it is advisable to use hash code generator using MD5, advance encryption scheme by using AES algorithm for secure data authentication, JAR based log harmonizer which provides secure storage by observing contents` activities, and Java security framework for secure services.

3. Methodology

This section provides the important points that encompassed with the research methodology. This study used simulation methodology using Riverbed Modular Academic Edition. However, Due to the limited features of Riverbed Modular Academic Edition, many important elements cannot be added such as selecting encoding technique. Furthermore, video streaming application is not supported in this version. Since video conferencing application is one of the applications that considered as video streaming, hence it is simulated instead.

The aim of this paper is to compare QoS performance based on the parameters listed previously for database application and video conferencing application.As the distance cannot be directly determined by editing nodes x and y coordinates. This problem is solved by using the propagation delay formula in order determine the distance

Propagation delay = distance / speed of light.

Since distance and speed of lights are determined, we can easily calculate the propagation delay, then we specified this value for delay property for each link that connect the router with the cloud. It is important to put into consideration that each link had its own delay value based on its length. Also, quality of videos cannot be measured directly from simulation tools. So in order to measure it, this study referred to IUT Telecommunication Standards (IUT-IT) for real time applications. Which stated that packets end to end delay should not exceeds 150 milliseconds in order to preserve good quality [9].

It is important to note that if there is a relation between distance and QoS especially for real time applications and if the QoS will be affected by long

distances between routers and the cloud. This study aims to answer these questions. The hypothesis is that if the distance between routers and cloud increased, QoS will be decreased.

Five scenarios were created, with different distances between routers. For first scenario total distance is fixed to 1 kilometer (KM) in order to compare to other scenarios and study the impact of each KM. The second scenario, total distance is equal to 3262.5 KM where the cloud is located in Jeddah whereas east router and west router are located in Cairo and Dubai respectively, as shown in Figure 1. The third scenario, distance is determined by 17772 KM where the cloud is located in Riyadh .East router in Brasilia and west router in Beijing, China as shown in Figure 2. The fourth scenario, distance is determined by 28422 KM where the cloud is located in New Delhi .East router in Buenos Aires and west router in Wellington as shown in Figure 3. The last scenario, distance is determined by 37763 KM where the cloud is located in Xinghua, China .East router in Brasilia and west router in Rosario, Argentina as shown in Figure 4.

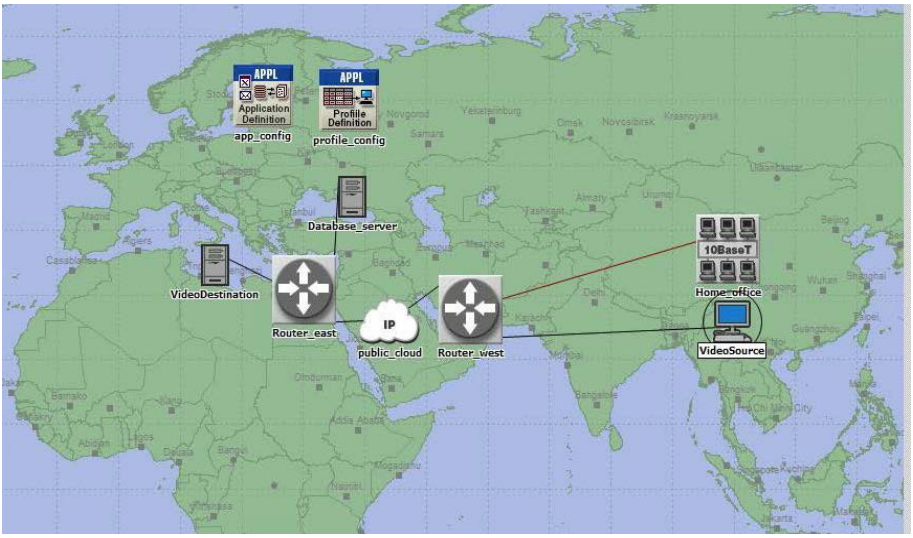


Fig. 1. Second Scenario Network

Fig. 2. Third Scenario Network

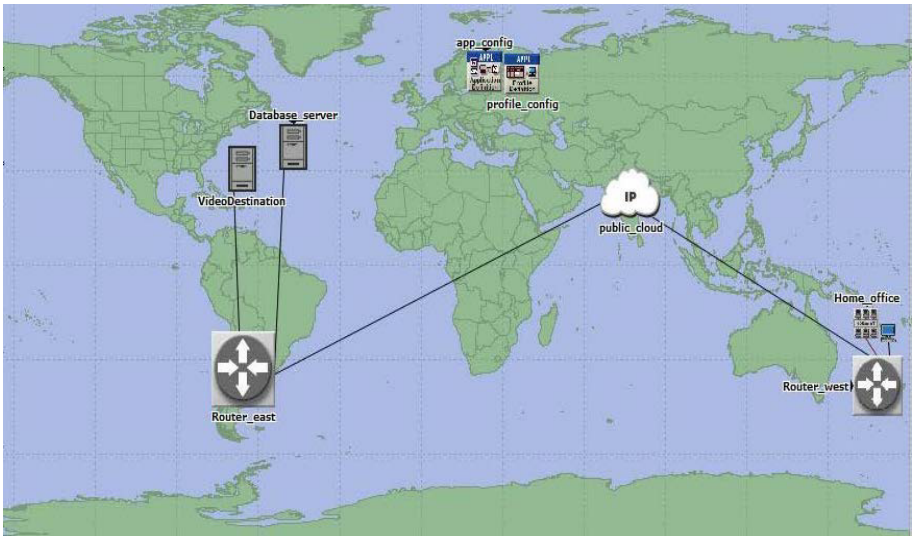


Fig. 3. Fourth Scenario Network

Fig. 4. Fifth Scenario Network

4. Results and Discussion

After simulation for 30 minutes, the obtained results of all scenarios are as follow: We started by database application and analyze response time parameters. The result is shown in Figure 5. Next, the results of video conferencing applications for end to end delay are illustrated in Figures 6. The result of total throughput is shown in Figure 7.

As seen in Figure 10, when the distance is only 1 kilometer (KM) the response time took about 250 milliseconds (mc) and 300 mc when the distance is 3000 KM which indicates that there is no huge difference. However, when in the 37K_scenario5 the response time took about 800 mc whereas 690 mc when the distance is 28 thousands which is represents the 28K_scenario 4. This indicated that 37K_scenario5 has an impact 13% higher than the 28K_scenario 4.

Next, when the distance is 1 KM, packet end to end delay was 50 mc. In the other hand, when the distance is 3000 KM, the delay was 61 mc. The important point, as shown in 28_scenario4 the distance is 28000 KM and the delay was 145 but when the distance is 37000 KM the delay was 180 mc. That's indicates that the distance in 37K_scenario5 has 19% higher impact then the distance in the 28K_scenario4. Hence, according to IUT standards the quality of videos will be affected as well as the delay exceed 150 mc.

For throughput results, as shown in Figure 13, the lines graph shows very clearly that the distance has no impact on. Therefore we can conclude that when the distance increased the QoS will be degraded. However, this doesn't imply on the throughput. Moreover, we can see that the impact of each kilometer is not stable.

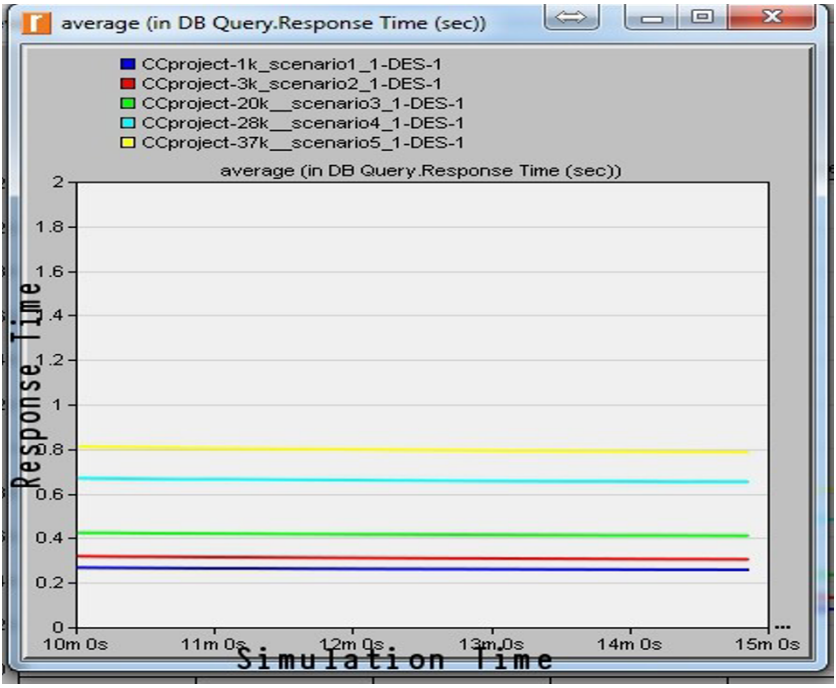


Fig. 5. Response Time Results

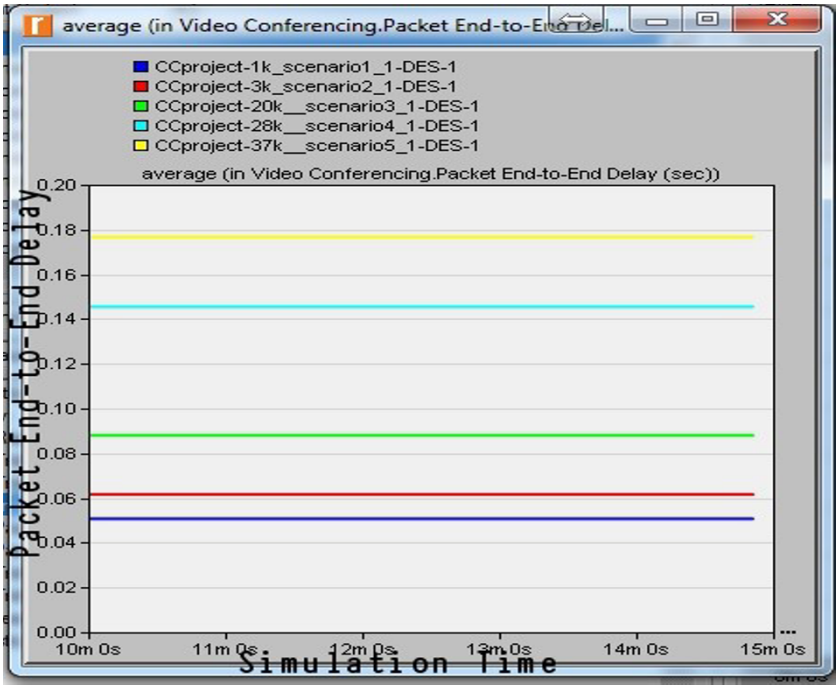


Fig. 6. Packet End to End Delay

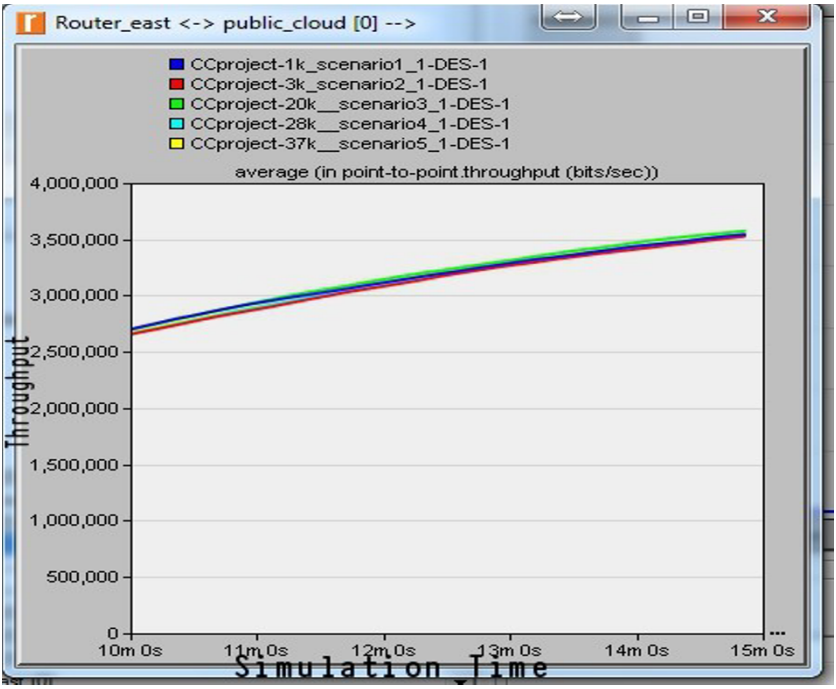


Fig. 7. Throughput Result

5. Conclusions

Video streaming applications preserve the highest traffic in networks as people prefer to watch videos online rather than downloading it. In order to offer such services, cloud provider have to ensure high quality of service since video streaming are most critical to deliver while preserving high quality at the same time. Therefore, this research studied the impact of distance on response time, packet end to end delay, and throughput.

After conducting the simulation, we analyze that there is a direct impact of distance on quality of service based on response time and packet end to end delay. As a result, video quality is degraded as well which indicates that when the distance increased, quality of service is degraded .Therefore, cloud services providers can target their consumers based on the geographical locations in order to deliver high quality services. In the other hand, distance has no impact on throughput based on simulation results. This is occurred because throughput concerns on how many packets transmitted regardless of the length from source to destination. Moreover, this study shows that the impact of each kilometer vary from one application to another as well in the application itself, which means the impact is not stable.

References

1. T. Erl, R. Puttini, and Z. Mahmood, *Cloud computing: Concepts, technology & architecture*. Prentice Hall, 2013.
2. Akpan H. A., and Vadhanam, B.R., "A Survey on Quality of Service in Cloud Computing", *International Journal of Computer Trends and Technology*, vol. 27, no. 1, pp. 58 – 63, 2015.
3. D.Kesavaraja, and A. Shenbagavalli, "Cloud Video as a Service [VaaSj] with Storage, Streaming, Security and Quality of service Approaches and Directions", in proceeding of *International Conference on Circuits, Power and Computing Technologies [ICCPCT-2013],* Nagercoil, India ,pp. 1093 – 1098, 201.
4. J. M. Pedersen, M. T. Riaz, J. C, B. Dubalski, D. Ledzinski, and A. Patel, "Assessing measurements of QoS for global cloud computing services," in Proc. *IEEE International Conference on Dependable, Autonomic and Secure Computing*, Sydney, New South Wales: IEEE, 2014, pp. 682–689.
5. Goel S.., "Cloud-Based Mobile Video Streaming Techniques", *International Journal of Wireless & Mobile Networks (IJWMN),* vol. 5, no. 1, pp. 85- 92, February 2013.
6. S. Suakanto, S. H. Supangka, Suhardi, and R. Saragih, "PERFORMANCE MEASUREMENT OF CLOUD COMPUTING SERVICES," *International Journal on Cloud Computing: Services and Architecture(IJCCSA)*, vol. 2, no. 2, pp. 9 – 20, Apr. 2012.
7. J. M. Pedersen, M. T. Riaz, J. C, B. Dubalski, D. Ledzinski, and A. Patel, "Assessing measurements of QoS for global cloud computing services," in Proc. *IEEE International Conference on Dependable, Autonomic and Secure Computing*, Sydney, New South Wales: IEEE, 2014, pp. 682–689.
8. D.Kesavaraja, and A. Shenbagavalli, "Cloud Video as a Service [VaaSj] with Storage, Streaming, Security and Quality of service Approaches and Directions", in proceeding of *International Conference on Circuits, Power and Computing Technologies [ICCPCT-2013],* Nagercoil, India ,pp. 1093 – 1098, 201.
9. VOIP-Info, QoS, " [Online]. Available: http://www.voip-info.org/wiki/view/QoS. Accessed: Apr. 27, 2016.

Application of Personalized Cryptography in Cloud Environment

Marek R. Ogiela[1], Lidia Ogiela[2]

[1] AGH University of Science and Technology
Faculty of Electrical Engineering, Automatics
Computer Science and Biomedical Engineering
30 Mickiewicza Ave., 30-059 Krakow, Poland
e-mail: mogiela@agh.edu.pl

[2] AGH University of Science and Technology
Cryptography and Cognitive Informatics Research Group
30 Mickiewicza Ave., 30-059 Krakow, Poland
e-mail: logiela@agh.edu.pl

Abstract. In this paper will be described the way of using personalized cryptography algorithms in cloud computing applications. Personal information may be used in creation advanced security protocols, which may be also applied for cloud data and services security and management. Such protocols may play important role in advanced secure management applications and intelligent access control to secure data.

1 Introduction

One of the most important computational paradigms, which expand modern security areas, is personalized cryptography, developed for using some personal data or features in security applications. Personalized cryptography [1], [2], [3] allow to make encryption process more personalized i.e. dependent on selected personal information, which may be associated with particular person or trusted user. In recent years have been proposed many interesting solutions for personalized cryptography. It seems that one of the most important applications of such procedures or protocols is security areas in Cloud Computing and distributed infrastructures. Cloud computing offers for user important computing resources and provide distributed applications which allow to facilitate quick information processing and accession to big data infrastructures [4], [5], [6].

Personalized cryptography is very promising area, which allow to use some special, very unique data for strong cryptographic purposes. Application of personal features additionally may allow to determine trusted sites, persons or participants of the protocols, whose data are used in cryptographic procedures.

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_24

In this paper will be presented some new possible applications of personalized cryptography approaches for distributed and cloud environment. Especially examples presented services and data management will be described.

2 The Concept of Personalized Cryptography

Personalized cryptography has been developed for several years [7], and is connected with application of personal or user features, biometrics or some special data for security purposes [8], [9], [10]. The most typical application of such techniques are connected with using different biometrics for cryptographic key generation, secret splitting, or hiding, creation of fuzzy vaults etc. Some important solutions within personalized cryptography were also connected with application of non-standard personal data like behavioral features or parameters extracted from personal medical records [11], [12].

Personalized cryptography requires application of special sensors or equipment which allows to determine or register some unique features. In case of using traditional biometric pattern it is possible to use standard sensors, which allow to evaluate requested patterns or features. For extraction more complex personal characteristics it is necessary to use dedicated systems like cognitive information systems which are enable to evaluate different, specific, rare, and very unique personal features [13], [14]. More detailed information connected with application of cognitive systems for personal feature extraction can be found in [15], [16]. It is worth noting that authors of this paper also suggest to use in encryption procedures some semantic information which may be extracted from encrypted information or from visual patterns connected with personal data (pictures, biometrics, visual characteristics) [17]. Such ideas are very promising, and in future should allow to define new areas of cryptography called cognitive cryptography [18], [19].

3 Application of Personalized Cryptography in Cloud Computing

Personalized cryptography offers secure and efficient solution for security purposes and distribution protocols. Such features may be very important in different and frequent applications for cloud computing security and distributed sensor networks.

It seems that one of the most important applications of such procedures is generation personalized cryptographic keys, which may be used in symmetric or asymmetric cryptosystems. To generate such keys it is necessary to encode some personal features into the random bit sequences. Random sequences will guarantee the strong security of such cryptographic keys, but personal features make them more personalized and associated with particular users or participants of protocols. Of course it may be necessary to hide personal data in key sequences in such manner that will guarantee its transparency for other users. For this purpose it is possible to applied different splitting or distribution strategies of personal data over random bit sequences. Having such personalized keys it is possible to use them in different cloud security application starting from pure encryption of strategic data stored in cloud

environment, over authenticating parties while transmitting data or providing services, till selected accessing grants for trusted users or consumers.

Second important application of personal features may be secure and efficient data distribution and management. For such protocols especially important may be application of biometric patterns, which allow not only encryption of confidential data but allow to hide or distribute them using fuzzy vaults approaches, or multi secret steganography [12]. Having big data infrastructures we can easily split over different places or repositories secret information divided using biometric patterns. Such key allow generate several different secret parts which may be also encoded using other personal characteristics. Finally only authorized person can restore the original information using his own biometric keys. One of the most important features of such solution is possibility to use personal data as steering information for secure data distribution over the cloud.

Additional for one participants we can consider application of his several different biometrics for securing or distribution several different secure information. Usually each biometrics also provides some redundant features, which in our application also may be applied for securing several different strategic data or classified information. In Fig. 1 is presented an idea of securing different data using several biometric patterns or using only one.

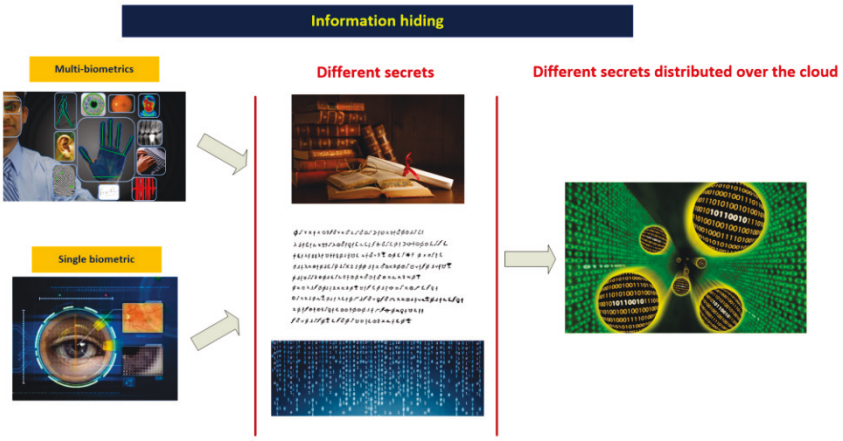


Fig. 1. An idea of distribution several different strategic data by application different biometric patterns or only one with redundant features.

Application of different crypto-biometric parameters has one very important feature. It allows to hide or split several different strategic data or secrets in totally independent way. This allows to reveal some of them without encryption of others. For non-trusted persons it will not be able to determine the exact number of secrets, so finally most of them are simply not visible for incidental persons. Different biometrics allows diffuse different data in independent way and according different schemas or algorithms.

Important aspect connected with biometric application is using of some non-standard biometrics or personal data which uniquely characterize different persons. Such non-standard characteristics may be obtained for particular persons from different information records, and may be connected with personal habits, behavioral features, health information etc. It may be also connected with personal morphometric describing external body features or internal body parts. The only one problem which may arise in using such not typical biometrics is connected with acquisition of such type of data. Usually it is necessary to apply special sensors or cognitive information systems, which are especially dedicated for analysis and extraction of such parameters.

4 Conclusions

In this paper were described the most important aspect of using personal information and biometric patterns to develop new solutions for security purposes and cryptography. Such new techniques and protocols are related to the new field of personalized cryptography, and may be successfully applied for several management tasks in the cloud environment. It may be also applied for secure and efficient services and data transfers in distributed environment. Personalized cryptography is very promising for such applications offering additional benefits by guaranteeing not only the security of transmitted data and services but also high efficiency of such protocols. Quite new idea related with personalized cryptography is application of cognitive techniques for acquisition of non-standard personal information or behavioral characteristics which also may be applied for cloud computing applications [15].

Acknowledgments. This work has been supported by the AGH University of Science and Technology research Grant No 11.11.120.329.

References

1. Ogiela, L., Ogiela, M.R.: Cognitive systems for intelligent business information management in cognitive economy. International Journal of Information Management, 34 (2014) 751-760
2. Hachaj, T., Ogiela, M.R.: CAD system for automatic analysis of CT perfusion maps. Opto-Electronic Review, 19 (2011) 95-103
3. Shi, J., Lam, K-Y.: VitaCode: Electrocardiogram Representation for Biometric Cryptography in Body Area Networks. 1st International Conference on Ubiquitous and Future Networks, Hong Kong, China, Jun 07-09 (2009) 112-115
4. Ogiela, L.: Semantic analysis and biological modelling in selected classes of cognitive information systems. Mathematical and Computer Modelling, 58 (2013) 1405-1414
5. Kumar, A., Kumar, A.: Adaptive management of multimodal biometrics fusion using ant colony optimization. Information Fusion 32 (2016) 49–63
6. Shi, J., Lam, K-Y.: VitaCode: Electrocardiogram Representation for Biometric Cryptography in Body Area Networks. 1st International Conference on Ubiquitous and Future Networks, Hong Kong, China, Jun 07-09 (2009) 112-115

7. Ogiela, L., Ogiela, M.R.: Cognitive systems and bio-inspired computing in homeland security. Journal of Network and Computer Applications, 38 (2014) 34-42
8. Ogiela, M.R., Ogiela, U.: Linguistic Approach to Cryptographic Data Sharing, FGCN 2008 – The 2nd International Conference on Future Generation Communication and Networking, December 13-15, 2008, Hainan Island, China, 1 (2008) 377–380
9. Bajwa, G., Dantu, R.: Neurokey: Towards a new paradigm of cancelable biometrics-based key generation using electroencephalograms. Computers & Security 62 (2016) 95-113
10. Hani, M.K., Marsono, M.N., Bakhteri, R.: Biometric encryption based on a fuzzy vault scheme with a fast chaff generation algorithm. Future Generation Comp. Syst. (2013) 800–810
11. Nandakumar, A.K.J.K., Pankanti, S.: Fingerprint-based fuzzy vault: Implementation and performance. In: IEEE Transactions on Information Forensics and Security. 2 (2007) 744–757
12. Ogiela, M.R., Ogiela, U.: Grammar Encoding in DNA-Like Secret Sharing Infrastructure. 2nd International Conference on Advanced Science and Technology (AST), Miyazaki, Japan, Jun 23-25, 2010. LNCS 6059, (2010) 175-182
13. Ogiela, M.R., Ogiela, U.: Shadow Generation Protocol in Linguistic Threshold Schemes. in: D. Slezak, T.H. Kim., W.C. Tang et all., Security Technology, Communications in Computer and Information Science 58 (2009) 35-42
14. Ogiela, L.: Towards cognitive economy. Soft Computing 18 (2014) 1675-1683
15. Ogiela, L., Ogiela, M.R.: Management Information Systems. in: J.J. Park, Y. Pan, H.C. Chao, et all., 2nd FTRA International Conference on Ubiquitous Computing Application and Wireless Sensor Network (UCAWSN), South Korea, 07-10 July 2014, Ubiquitous Computing Application and Wireless Sensor, Lecture Notes in Electrical Engineering 331 (2015) 449-456
16. Ogiela, L.: Cognitive informatics in image semantics description, identification and automatic pattern understanding. Neurocomputing 122 (2013) 58-69
17. Ogiela, L.: Cognitive Computational Intelligence in Medical Pattern Semantic Understanding. in: M. Guo, L. Zhao, L. Wang (Eds.), Fourth International Conference on Natural Computation, ICNC 2008, Jinan, Shandong, China, 18-20 October, 2008, 245-247
18. Ogiela, L.: Computational Intelligence in Cognitive Healthcare Information Systems. in: I. Bichindaritz, S. Vaidya, A. Jain et all., Computational Intelligence in Healthcare 4: Advanced Methodologies, Studies in Computational Intelligence 309 (2010) 347-369
19. Ogiela, L.: Data management in cognitive financial systems. International Journal of Information Management 33 (2013) 263-270

Optimizing Machine Learning based Large Scale Android Malwares Detection by Feature Disposal

Lingling Zeng, Min Lei, Wei Ren, Shiqi Deng

Abstract As a favorable opening platform for mobile terminals, android platform attracts close attentions from a large number of hackers. The great potential of security hazard makes the requirement of malicious software detection become effective, rapid and multitudinous. In recent years, a lot of machine learning based methods have been proposed. However, most of the focuses are searching for more effective feature information. In this paper, we propose an optimization method for machine-learning-based malware detection by focusing on the disposal of the feature information. We extract permission and intent information of malwares, and dispose them in a series of effectively methods. After the disposal, we use several machine learning algorithms to verify their effectiveness, and conclude a comparing list. After the comparing, we propose an optimization algorithm by combining several effective processing. The effectiveness of our proposal is illustrated and justified in extensive experimental results.

1 Introduction

Smartphones develop rapidly in recent years, especially with the Android mobile platforms. However, the popularity of Android also attracted vast hackers. According to the latest statistics from "G DATA", a network security company in Germany, the number of Android malware samples climbed to 2.333 million in 2015, which was a 50% increase over 2014 And the number reached 3.396 million in 2016 according to "360 Internet Security Center" in China. Meanwhile, problems the An-

Lingling Zeng, Wei Ren
School of Computer Science, China University of Geosciences, Wuhan, P. R. China, e-mail: weirencs@cug.edu.cn

Min Lei, Shiqi Deng
Information Security Center, Beijing University of Post and Telecommunications, Beijing, P. R. China, e-mail: leimin@bupt.edu.cn

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_25

droid ecosystem is facing are not only the increasing number of malicious apps, but also the new malwares whose internal structure is becoming more and more variable and complex.

The huge quantity and rapidly changes make it impossible to depend on only ordinary static and dynamic Android malware detections. Under this circumstance, the machine learning is introduced to solve this problem. To improve the efficiency and accuracy of malware detection, we pay attention to the optimization of data processing. The data we chose is extracted from AndroidManifest.xml. In order to test our scheme, we use machine learning algorithms, such as J48 and RandomForest etc., in Weka as the training model of processed data. The contributions of this paper are as follows:

1. We adopt a variety of data preprocessing methods in feature pretreatment disposal, and test their performance with multiple machine learning algorithms. Finally we display these performances and give our evaluations.

2. We integrate diversified data preprocessing method effectively, and improved the efficiency and accuracy of Android malware detection successfully.

The rest of this paper is organized as follows. We summarize the previous work in section 2. In section 3, we provide preprocessing for the feature data and propose our scheme in section 4. Experimental results are discussed in section 5. This paper is concluded in section 6.

2 Related Work

Machine learning was introduced into android malicious code detection since 2012. In recent years, a lot of malware detection algorithm was proposed.

At the beginning of the application, permissions and corresponding information were extracted as the feature data. Sahs and Khan [1] applied permissions and Control Flow Graphs as features, and used One-class Support Vector Machine to train the data. Most of training data are benign applications so that the classifier will classify a sample as malicious only if it is sufficiently different from the benign class. And Westyarian et al. [2] used 205 benign and 207 malicious apps and extracted API calls that are only related to the permission declared in manifest. xml file. The study concluded that telephonyManager and connectivityManager are the most important features. And random forest classification obtains 92.4% and SVM obtain 91.4%.

Along with the development of the research, more features were excavated from benign apps and malwares. Yerima et al. [3] features like permissions and code based properties such as API calls, both Java system based and Android system based, Linux and Android system commands. And presented a list of top 20 permissions and top 25 API calls used by benign and malicious applications.

These methods increased malware detection rate by searching for effective feature information, and mixing multiple feature data. The focus and difficulty of these methods lie in the feature extraction. Different types of characteristics require d-

ifferent processing methods, which costs great consumption. In comparisons, our work focuses on the feature disposal. We clean features of permission and action using a variety of feature data processing method. And optimize the most optimal processing method.

3 Preliminary

Feature classification algorithms that are commonly used in the field of text classification are: TF - IDF, information gain, chi-square, and etc. We will apply these three methods in disposing feature information in order to evaluate their performance and provide our optimization scheme. In this section, we will introduce these methods briefly as following.

3.1 TF-IDF

The main idea of the TF-IDF algorithm is that, the higher frequency a word or phrase in an article, and rarer appears in other articles, the better distinguish ability the word or phrase has, which indicates that its suitability for classification.

TF (Term Frequency), denotes the frequency the terms appear in a document "D".

IDF (Inverse Document Frequency), the less document contains word "t", the higher value of IDF, which means the greater distinguish ability of word "t".

We set n and N represent the frequency of word "t" and the total number of the feature respectively. And d is the number of document contain that word "t" while D is the total number of documents. The formula to calculate value of TF-IDF is:

$$V_{TF-IDF} = TF \times IDF = \frac{n}{N} \times \log_{10} \frac{D}{d+1} , \tag{1}$$

3.2 Information Gain

The central idea for information gain is to figure out how much information a feature can bring for classification system, the more information it can bring, the more important the feature is.

However, attributes selected by information gain can only denote its contribution for the whole system, rather than specific to a certain category. This specialty decides that it can only be used in the global feature selection, namely, all the classes have to use the same feature set.

$$V_{IG} = \sum_{i=1}^{n} P(C_i)\log_2 P(C_i) + P(t)\sum_{i=1}^{n} P(C_i|t)\log_2 P(C_i|t) + P(\bar{t})\sum_{i=1}^{n} P(C_i|\bar{t})\log_2 P(C_i|\bar{t}), \tag{2}$$

3.3 Chi-Square Test

Chi-square test, a method commonly used in mathematical statistics, is to decide whether the two variables are independent. The method quantify the correlation between features and categories, and the score of feature attributes increases with the correlation tightness. The high correlation feature should be retained.

Assuming the observed values are independent of each other, and the value of expectation is "E". And after finished the pretreatments, compare their values with predetermined threshold, if it is less than the threshold, then hypothesis tenability. The formula for chi-square value is as follows:

$$\chi^2 = \sum_{i=1}^{n} \frac{(x_i - E)^2}{E} \tag{3}$$

In the preliminary, we choose 814 malicious apps and 781 benign apps as the experimental samples. We extract permission and intent information from malicious and benign apps respectively. The corresponding operation permissions must be declared before a malware conducts malicious operations, and the intent fields reflect messages the app applies or transmits. For example, due to the destructive function, malicious apps prefer the one-off application for permissions and intents about sending messages, opening traffic backstage, getting location information and etc. Therefore, the permission and intent information, which accesses to the intention and operations of an author, can be used as the feature data for malware detection.

Permission comes in two forms: standard format and custom format. We intercept the substantial part of the permission statements to simplify and unitize the feature information. And we extract the substantial part of the action segment as the most representative data of intent. Meanwhile, the extraction of permission and intent at large scale could be implemented at the same time, which is proved to be efficient.

The total number of permission of malwares is 7654, and falls into 163 variety, while the 4097 action statements fall into 503 categories. The permission of benign apps are divided into 145 kinds from 3067 permission statements while the action statements in 230 kinds from 1556 action statements.

The experiments are divided into two parts. We firstly dispose the feature information using three text feature disposal methods mention before.

4 Proposed Approach

In this work, an optimization approach for machine learning based android malware detection is proposed. The proposed approach optimized the feature disposal according to the experimental results of several kinds of text feature information processing methods.

We can conclude from the result shown in Section 5.1 that TF-IDF has the best effect on the text feature of permission and action. On the one hand, TF-IDF doesn't consider the repetitive parts exist in both permission and action information between benign and malicious apps. In this case, we have to consider the influence on both sides at the same time. On the other hand, we can join the simple artificial intervention in text characteristic analysis. We select seven high-risk keywords contain dangerous intention extracted from the permission and action statements respectively, and each keyword will be given a corresponding weights. In this case, some low-frequency words appear only in the malicious code may be added into the learning model. The threshold number for the same attribute can be improved at the same time to ensure the stability for attribute amounts. The algorithm for calculating the corresponding weight can be described as the formula below:

$$V_{Weight} = \frac{|m-b|}{N} \times 10 + 1 \,, \tag{4}$$

Where m and b represent the frequency of keywords in the malicious and benign software representatively, while N is the total number of permissions or actions.

The danger keywords and corresponding weights are displayed in Table 1

Table 1 Weights of KeyWords

Permission	Weight
DISABLE	1.027
MMS	1.018
SHORTCUT	1.212
CONTACT	1.132
STORAGE	1.314
SETTING	1.258
SMS	2.048

Action	Weight
BOOT	1.702
REMOVED	1.060
PACKAGE	1.147
STATE	1.262
EDIT	1.049
CHANGED	1.060
SMS	1.501

4.1 Notations

For fast checking the short notation, we list the major notations used in the remainder of the paper in Table 2.

The algorithm of our proposed scheme can be briefly described in Algorithm 1.

Table 2 Notations

Symbol	Meaning
Prm_m	Permission extracted from malicious apps
Prm_b	Permission extracted from benign apps
Atn_m	Action extracted from malicious apps
Atn_b	Action extracted from benign apps
m_p, m_a	Statistical frequency of a certain permission or action in malicious apps
b_p, b_a	Statistical frequency of a certain permission or action in benign apps
M_P, M_A	Total number of the permission or action in malicious apps
B_P, B_A	Total number of the permission or action in benign apps
KyW	Keywords of permission or action in table 1
Wgt	Corresponding weight of keywords in table 1

Algorithm 1 Optimization based on TF-IDF

Require: $Prm_m, Prm_b, Atn_m, Atn_b, m, M, b, B, Kyw, Wgt$.
Ensure: *Feature Matrix*
1: find items Prm_s which are shared in Prm_m and Prm_b,
2: find items Atn_s which are shared in Atn_m and Atn_b,
3: **for all** Prm_s **do**
4: $TFIDF_{mi} - TFIDF_{bi}$
5: **end for**
6: **for all** rest of the Prm_m **do**
7: $TFIDF_{mi}$
8: **if** $prm_m i$ shares the same item in KyW **then**
9: $TFIDF_{mi} \leftarrow TFIDF_{mi} \times Wgt$
10: **end if**
11: **end for**
12: **for all** rest of the Prm_b and Atn_b **do**
13: $TFIDF_{bi}$
14: **if** prm_{bi} shares the same item in KyW **then**
15: $TFIDF_{bi} \leftarrow TFIDF_{bi} \times Wgt$
16: **end if**
17: **end for**
18: descending order prm_{si}, extract the item greater than $threshold_s$
19: descending order prm_{bi} and prm_{mi}, extract the item greater than $threshold_d$
20: repeat from step 3 on Atn_s, Atn_p, and Atn_b

After selected feature permissions and actions, we match it to each app, and set the value to "1" if the corresponding permission or action item exists in that app, otherwise to "0". And we'll set the final feature attribute as "M" or "B", representing malicious or benign app, to distinguish malicious and benign apps. On this account, we will build 1292 multi-dimensional vectors.

Then, we put feature vectors into Weka, a open source machine learning software based on JAVA environment, to establish the classification model. By comparing the result data in Weka, we can fairly evaluate feature processing performance of out proposed methods.

5 Result

We adopt four machine learning algorithms: J48, NaiveBayes, RandomForest and SMO. And the detail results about the time-consuming for model building , and details about accuracy of four feature disposal methods, including our proposed methods, are listed as follows:

5.1 Conventional Methods Results

Table 3 Result of TF-IDF

Algorithm	Time-Consuming	Accuracy	TP	FP	Precision	Recall	ROC Area
J48	0.46s	94.4961	0.945	0.059	0.945	0.945	0.952
NaiveBayes	0.12s	86.9767	0.87	0.162	0.872	0.87	0.956
RandomForest	4.81s	96.8217	0.968	0.03	0.969	0.968	0.994
SMO	1.15s	96.3566	0.964	0.042	0.964	0.964	0.961

Table 4 Result of IG

Algorithm	Time-Consuming	Accuracy	TP	FP	Precision	Recall	ROC Area
J48	0.42s	83.7984	0.838	0.192	0.838	0.838	0.946
NaiveBayes	0.07s	79.7674	0.798	0.268	0.813	0.798	0.926
RandomForest	9.94s	83.9535	0.84	0.18	0.839	0.84	0.958
SMO	1.05s	90.8527	0.909	0.08	0.914	0.909	0.914

Table 5 Result of Chi-Square

Algorithm	Time-Consuming	Accuracy	TP	FP	Precision	Recall	ROC Area
J48	0.33s	89.6124	0.896	0.131	0.899	0.957	0.957
NaiveBayes	0.06s	89.3798	0.894	0.138	0.898	0.894	0.966
RandomForest	6.96s	91.4729	0.915	0.098	0.915	0.915	0.985
SMO	1.13s	94.8062	0.948	0.056	0.948	0.948	0.946

Compare the three kinds of conventional feature data processing methods, RandomForest and SMO show higher accuracy. The rate of TP for TF-IDF and Chi-

Square are both higher than 0.9, among which the TF-IDF even higher than 0.96. On the other hand, these two kinds of algorithms are time-consuming, especially for RandomForest, which costs hundreds as much time over Naive Bayes. In this case, methods of reducing the testing time of high accuracy algorithm, and ensuring the detection rate at the same time is demanded.

5.2 Proposed Approach Results

Table 6 Proposed Scheme

Algorithm	Time-Consuming	Accuracy	TP	FP	Precision	Recall	ROC Area
J48	0.34s	94.7287	0.947	0.056	0.947	0.947	0.951
NaiveBayes	0.09s	87.907	0.879	0.15	0.882	0.879	0.972
RandomForest	2.67s	97.2868	0.973	0.03	0.973	0.973	0.994
SMO	1.17s	96.4341	0.964	0.039	0.964	0.964	0.963

We list the experimental results of our proposed scheme as above. Compares to the result of TF-IDF, we reduce the time taken for models building in 24.07% on average, especially for the RandomForests which acquires the highest accuracy and reduced 46.98% of time consuming. And the accuracy for the proposed scheme has been slightly improved in 0.5 percent on average. The comparison of accuracy and time taken for model build is shown below.

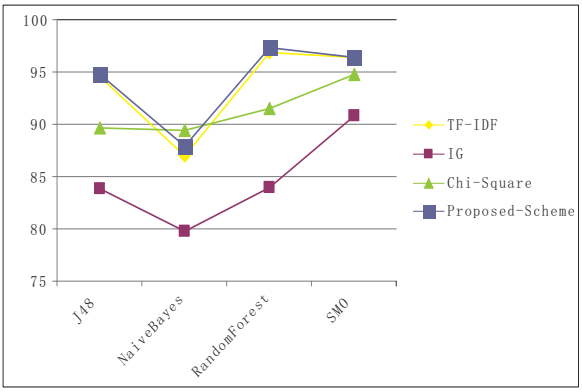


Fig. 1 Accuracy

In this way we can draw the conclusion that the approach we proposed is proved to be efficiency and significant.

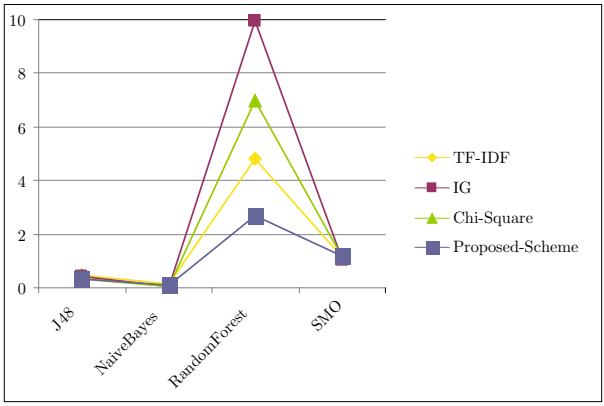


Fig. 2 Time Taken

6 Conclusion

In this paper, a optimizing machine learning based android malwares detection method has been successfully introduced. We optimize the algorithm based on TF-IDF, which, according to our experiment, is proved to be the most efficient method of three text feature processing method.

Experimental results show that the proposed approach achieves shortening the time of learning, modeling, and classification under the guarantee of classification quality, and even slightly raises the detection rate. In addition, this paper provides the performances of three algorithms under a variety of machine learning algorithms. As for the future work, we will find out the best algorithm for other malware detective features, and provide our optimized scheme for better efficiency and detection rate. The time complexity for the optimization method we proposed is $O(n)$ without any increase, which is also proved to be lightweight.

Acknowledgement

The research was financially supported by the National Natural Science Foundation of China under Grant No.s 61310306028.

References

1. J. Sahs and L. Khan.: A machine learning approach to android malware detection. In Intelligence and Security Informatics Conference (EISIC). (2012) doi: 10.1109/EISIC2012.34.
2. Westyarian, Y. Rosmansyah, and B. Dabarsyah.: Malware detection on android smartphones using api class and machine learning. In Electrical Engineering and Informatics (ICEEI).

(2015) doi: 10.1109/ICEEI2015.7352513.
3. S.Y. Yerima, S. Sezer, and I. Muttik.: Android malware detection using parallel machine learning classifiers. In Next Generation Mobile Apps, Services and Technologies (NGMAST). (2014). doi: 10.1109/NGMAST.2014.23.
4. Jit-Ping Siew, Heng-Chin Low, Ping-Chow Teoh.: An interactive mobile learning application using machine learning framework in a flexible manufacturing environment. In Next Generation Mobile Apps, Services and Technologies (NGMAST). (2014). doi: 10.1109/NGMAST.2014.23.
5. Javanmardi, S., Shojafar, M., Shariatmadari, S., Ahrabi, S. S. (2015). Frtrust: a fuzzy reputation based model for trust management in semantic p2p grids. International Journal of Grid & Utility Computing, 6(1), 57-66.
6. Brown, M. S., Tarverdian, E., Pelosi, M. (2015). Selection of direct and derived function point estimation methods. International Journal of Grid & Utility Computing, 6(1).
7. Hsu, T. Y., Kshemkalyani, A. D. (2015). Variable social vector clocks for exploring user interactions in social communication networks. International Journal of Space-Based and Situated Computing, 5.
8. Arour, K., Zammali, S., Bouzeghoub, A. (2015). Test-bed building process for context-aware peer-to-peer information retrieval evaluation. International Journal of Space-Based and Situated Computing, 5(1).
9. Thomas, A. M., Moore, P., Shah, H., Evans, C., Sharma, M., Xhafa, F., et al. (2013). Smart care spaces: needs for intelligent at-home care. International Journal of Space-Based and Situated Computing, 3(1), 35-44.

Research on Decisive Mechanism of Internet Financial Interest Rate

Shengdong Mu[1], Yixiang Tian[1], Li Li[1] , Xu An Wang[2]
121806967@qq.com
1 School of Management and Economics,
University of Electronic Science and Technology of China, Chengdu 614100
2 Engineering University of CAPF, Xi'an 710086

Abstract: This article studies the decisive mechanism of the Internet financial interest rate. Considering the possibility of basic symmetry for the credit information in the internet financial market with the application and development of the big data technology, a Nash Bargaining Module is established in the article, which includes the main decisive factors of the Internet financial interest rate (the program natural risk, program rate of return, loan period, loan scale, loan urgency, operating capacities of the borrower, the competition of the credit market, and the regime risk) as variables for analysis of their effects on the internet financial interest rates. According to relevant researches, the uncertainty of the loan projects in themselves and the costs for the implementation of secured credit contracts, rather than the reverse selection and moral risks arising from information asymmetry, form the main issues of the Internet financial interest rate.

Key words: Internet finance, decisive mechanism of the Interest rate, Nash bargaining solution

I Introduction

Interest rates are closely dependent on the development of Internet finance. Deposits, as the main passive liabilities for banks, constitute the main capital sources of commercial banks along with other active liabilities in the monetary market. The distortion of the deposit rates and the Internet technology development are the main reasons for the generation of Internet finance. As an essential impetus in economic development, Internet technology alters the financial operation modes significantly in aspects of payment modes, information processing and resource configuration. (Xie Ping, Zhou Chuanwei, 2012; Huang Xu et al, 2013). According to Chen Zhiwu (2014), Internet poses a great challenge for traditional banks and capital markets in the sense of channels, however, it has no difference from the products of banks and

[1] Fund program: fund program of "Reasons for the Lowered Sovereign Credit Rating Hurting the Global Economy- Exploration of International Mechanism and Relevant Policies" (12YJA790125) under the research project of human social science of Ministry of Education of the People's Republic of China

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_26

capital markets in structure and design; the essence of the finance stays the same, that is the time spread value exchange, and the credit exchange; it is the scope, number of people involved, amount and environment, rather than the essence of the financial transactions, that are changed by the Internet. Wu Xiaoqiu (2014), Xu Rong (2014) points out, Internet finance is an investment and financing structure built on the Internet platform with financing functions, basic and independent surviving spaces; in the aspect of generic matching, the Internet platform is characterized by higher coupling with the four main functions of the financial system, namely, payment and settlement, provision of price information, risk management, and resource configuration. In general, Internet finance covers both the financial businesses of non-financial Internet enterprises and businesses of financial institutions by the means of Internet in the board terms. While the Internet finance in the narrow terms refers only to financial businesses based on the Internet technologies by the Internet enterprises. The Internet finance in China currently is divided into Internet of finance and finance of Internet on the basis of the nature. Internet of finance refers to that traditional financial institutions utilizes Internet technologies to overcome the limits of time, space and physical network, so as to improve their businesses thoroughly, and provide various financial services by the Internet; finance of Internet, refers to that Internet enterprises which has nothing to do with the finance start to provide financial services. Exterior forces are affecting the finance irreversibly by the Internet, which causing significant reform in the financial industry (Li Haifeng, 2013). The theory of financial repression and deepening put forward by McKinnon and Shaw is one of the most important theories on liberalization of interest rates. It is common in developing countries to choose the wrong financial policies and systems. To be specific, it is featured by enforcement intervention over the financial activities by the government, imposing of low interest rates and exchange rates artificially, the so-called "financial regulation".

The interest rate control forms the impetus of the prosperity of the internet finance. However, there are few references on the decisive factors of the interest rates. The article aims to build a Nash Bargaining Module with general symmetry information, which includes the main decisive factors of the Internet financial interest rate, including the risk of credit default (such as program natural risks, adverse selection and moral risks), the transaction cost, the loan scale, the competition of the Internet credit market, the regime risk, the formal financial interest rate, the operating capacity of the borrower, the loan period, and the loan urgency in the model as variables to analyze the decisive mechanism of the Internet financial rates.

II Module

(I) Selection of Benchmark Model

Now that the information asymmetry is no longer an issue for the Internet finance, we will consider other factors affecting borrowing rates for the Internet finance, including: natural risks of the loan program, program rates of return, loan periods,

loan scales, loan urgency, operating capacities of the borrower, the competition of the credit market, and regime risks. These factors determine the capacity to borrow for the borrower, in other words, the Internet financial interests rates are mainly dependent on the strength gambling between the borrower and lender.

Nash (1950) put forward the following Symmetric Nash Bargaining Solution:

$$x^* = ArgMax_{x \in S} \left[U_b(x) - B_b\right]\left[U_l(x) - B_l\right]$$

Where $U_b(x)$ and $U_l(x)$ refer to utility functions of the bargaining parities under scheme x, B_b and B_l refer to the Utility of Breakdown Point of both parties, x^* is the Pareto-efficiency scheme and S is the sum of possible schemes.

Kalai and Smorodinsky (1975) presented the Drop Asymmetric Nash Bargaining Solution by relaxing conditions for the Symmetric Nash Bargaining Module.

$$x^* = ArgMax_{x \in S} \left[U_b(x) - B_b\right]^a \left[U_l(x) - B_l\right]^{(1-a)}$$

Where a and b refer to the relative bargaining power for the bargaining parties.

By comparing the above two bargaining models, general symmetry of the borrower and lender of the Internet finance is achieved through big data analysis. The Drop Asymmetric Nash Bargaining Model is selected as the theory model for the analysis of borrowing rates for the Internet finance.

(II) Establishment of the Module

The following fundamental assumptions are made for the Internet financial loaning transactions to simplify the analysis: (1) general symmetry is achieved for basic information of the borrower and the lender, (2) the loan is intended for operative investment, and the self-owned capital and the loans in the Internet financial market form the sources for all capital of operative investment; (3) the benefits from the loans are used to repay the principals and interests of the loans; (4) the issues of credit risks have been solved, that is to say the borrower will be certain to repay the principals and interests of the loans with the benefits of the loans; (5) there is no explicit mortgage or guarantee.

A specific Internet financial credit transaction model is established then. For example, an enterprise owner (or a farmer) owns an vestment project, and the rate of return for the project is a random variable P, $P \in [0, W]$, where W is a number larger than 0, and $f(P)$ is the density function of p, then the average rate of return for the program is $\overline{P} = \int_0^W Pf(P)dP$. Suppose the enterprise owner (or the farmer) doesn't have sufficient capital, a loan is necessary, and the loan ratio is r, $r \in [0,1]$, and a lender is willing to provide the loan by Internet and the bargaining loan rate is e, then the benefits distribution information of both the borrower and lender upon the end of the loan is as follows:

Where $P \le r(1+e)$, the benefit of the borrower is 0 and the benefit of the lender is P;

Where $P \geq r(1+e)$, the benefit of the borrower is $P - r(1+e)$ and the benefit of the lender is $r(1+e)$.

The loan rate is the result of bargaining between the borrower and the lender based on their bargaining positions respectively. Suppose the relative bargaining power of the borrower is a, and the relative bargaining power of the lender is $(1-a)$, and $a \in (0,1)$, and a is the function of the total number of borrowers n, and the total number of lenders m, the loan period t, the loan urgency u, the operation capacity of the borrower f and the regime risk level v, i.e., $a = a(n,m,t,u,f,v)$. Apparently, the less the lenders are in the market, the higher the relative bargaining power of the borrower is, and vice versa; and the extension of the loan term, the urgency the loan, and the increase of the regime risk will decrease the bargaining power of the borrower (Jiang Shuxia, Qin Guolou, 2000; Zhang Jianzhong, Yuan Zhonghong, Lin Ping, 2002), then:

$$\partial a/\partial n < 0,\ \partial a/\partial m > 0,\ \partial a/\partial t < 0,\ \partial a/\partial u < 0,\ \partial a/\partial f > 0,\ \partial a/\partial v < 0$$

Suppose U_b and B_b are the Utility Function and Utility of Breakdown Point of the borrower respectively, U_l and B_l are the Utility Function and Utility of Breakdown Point of the lender respectively. Considering that it is very difficult to quantify the utility, the Utility Function can generally be replaced by the benefit (or cost) function under the hypothesis of "economic person",

then:

$$U_b = \int_{r(1+e)}^{W} \left[P - r(1+e)\right] f(P)dP$$

$$U_l = \int_{0}^{r(1+e)} Pf(P)dP + \int_{r(1+e)}^{W} r(1+e) f(P)dP$$

The Utility of Breakdown Point B_b and B_l can be expressed by the opportunity cost and the transaction cost. Suppose s is the bank interest rate of the same period, and equals the Shanghai Interbank Offered Rate (hereinafter as Shibor), C_b is the transaction cost of the borrower, and C_l is the transaction cost of the lender, then B_b and B_l can be expressed by:

$$B_b = C_b + (1-r)(1+s),\ B_l = C_l + r(1+s)$$

III Analysis of Decisive Mechanism of Interest Rates

Based on the above analysis and assumption, the borrowing rate e^* under Pareto-efficiency situation is a Drop Asymmetric Nash Bargaining Solution, with the following algebraic programming expressions:

$$\underset{e}{Max}\left[U_b(e) - B_b\right]^a \left[U_l(e) - B_l\right]^{1-a}$$

$$s.t.\ U_b(e) - B_b \geq 0,\quad U_l(e) - B_l \geq 0$$

The first order optimal conditions are described as follows:

$$\left[\left(U_b - B_b\right)^a \left(U_l - B_l\right)^{1-a}\right]^l = a\left(U_b - B_b\right)^{a-1}\left(U_l - B_l\right)^{1-a} U_b^l + \left(1-a\right)\left(U_b - B_b\right)^a \left(U_l - B_l\right)^{-a} U_l^l + 0 \tag{1}$$

$$U_b = \int_{r(1+e)}^{W}\left[P - r\left(1+e\right)\right] f\left(P\right) dP \tag{2}$$

$$U_l = \int_0^{r(1+e)} Pf\left(P\right) dP + \int_{r(1+e)}^{W} r\left(1+e\right) f\left(P\right) dP \tag{3}$$

$$\frac{\partial U_b}{\partial e} = -r^2\left(1+e\right) f\left[r\left(1+e\right)\right] - r\int_{r(1+e)}^{W} f\left(P\right) dP + r^2\left(1+e\right) f\left[r\left(1+e\right)\right] = -r\int_{r(1+e)}^{W} f\left(P\right) dP$$

$$\frac{\partial U_l}{\partial e} = r^2\left(1+e\right) f\left[r\left(1+e\right)\right] + r\int_{r(1+e)}^{W} f\left(P\right) dP - r^2\left(1+e\right) f\left[r\left(1+e\right)\right] = r\int_{r(1+e)}^{W} f\left(P\right) dP \tag{4}$$

Considering $\frac{\partial U_b}{\partial e} = -\frac{\partial U_l}{\partial r}$, then $U_b^l = -U_l^l$

Plug expression (2), (3) and (4) into expression (1) to achieve the first order maximum conditions:

$$\int_{r\left(1+e^*\right)}^{W}\left[P - r\left(1+e^*\right)\right] f\left(P\right) dP - a\left(\overline{P} - B_b - B_l\right) - B_b + 0 \tag{5}$$

Plug expression $B_b = C_b + \left(1-r\right)\left(1+s\right)$, $B_l = C_l + r\left(1+s\right)$ into expression (1), then:

$$\int_{r\left(1+e^*\right)}^{W}\left[P - r\left(1+e^*\right)\right] f\left(P\right) dP - a\left(\overline{P} - C_b - C_l - s - 1\right) - C_b - \left(1-r\right)\left(1+s\right) + 0 \tag{6}$$

The first order conditions imply that the optimum interest rate e^* is dependent on the natural risk of the loan program, the expected rate of return of the program, the loan ratio, the transaction cost of the lender, the transaction cost of the borrower, the bank interest rate, and the relative bargaining power of the borrower(the loan period, the operating capacity of the borrower, the loan urgency, the competition of the credit market, and the regime risk), i.e., the explicit function under the first order conditions is $e^* = e^*\left(A, a, s, r, \overline{P}, C_b, C_l\right)$. Where, $A = \int_{r\left(1+e^*\right)}^{W} f\left(P\right) dP$, refers to the probability function of zero loss for the lender. Evidently, we can find a negative correlation between A and the program natural risks, and we can also get the natural risk level of the loan program.

The effects of the above factors on the Internet financial interest rates are further analyzed in accordance with the first order optimum conditions, i.e., the function $e^* = e^*\left(A, a, s, r, \overline{P}, C_b, C_l\right)$ in the following.

Proposition I. There is a positive correlation between the Internet financial interest rate and the natural risk level of the loan program. The lender is likely to demand corresponding risk premiums based on the natural risk level of the program.

Demonstration: The probability function of zero loss for the lender is $A=\int_{r(1+e^*)}^{W} f(P)dP$, take the derivative with respect to e^* in the function $A=A(e^*)$, and the derivative is less than 0:

Due to $\frac{\partial A}{\partial e^*}=-rf\left[r(1+e^*)\right]<0$, then: $\frac{\partial e^*}{\partial A}<0$

According to $\partial e^*/\partial A$, the Internet financial interest rate is a decreasing function of the probability of zero loss for the lender. The probability of zero loss for the lender is qualified to be used to evaluate the natural risk level of the loan program. In other words, the lower the program natural risk is, the higher the probability of zero loss for the lender is; and vice versa. As a result, there is a positive correlation between the Internet financial interest rate and the natural risk level of the loan program. The proposition I that the lender is likely to demand corresponding risk premiums based on the natural risk levels of the programs is correct.

In the Internet financial market, general information symmetry is achieved between the loaner and the borrower through the big data system. In the event of the loan, the lender has already had a basic understanding of the loan risks (i.e., ex-ante risks). Therefore, relevant strategies will be taken by the lender for the loan, that is, they will demand different borrowing rates on the ground of different natural risk levels of loan programs. Higher interests are demanded for loan programs with higher natural risks to be used as the premiums for the risk borne by the lender for the loan.

Proposition II. There is no simple positive correlation between the Internet financial interest rate and the expected rate of return for the loan program.

Demonstration: In expression (5) under the first order optimum conditions, the decisive function of the optimum interest rate is defined implicitly, $e^*=e^*(\bar{P})$. By using the implicit function theorem, we can get the relationship between the optimum interest rate e^* and the expected rate of return for the loan program $\bar{P}$. Considering $\bar{P}=\int_0^W Pf(P)dP$, it is inappropriate to take the derivative with respect to $\bar{P}$ in expression (5) only. It is also necessary take the derivative with respect to $\bar{P}$ in $f(P)$. As $f(P)$ is an abstract function, it is impossible to get the relationship between $f(P)$ and $\bar{P}$ directly. As a result, we will simplify $f(P)$ into a specific normal distribution function with a standard deviation of σ (normal distribution is very representative).

Suppose: $f(P)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(P-\bar{P})^2}{2\sigma^2}\right]$, and $W\to+\infty$, and plug the above expressions into expression (5) under the first order optimum conditions:

$$\int_{r(1+e^*)}^{+\infty}\left[P-r(1+e^*)\right]\left[\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(P-\bar{P})^2}{2\sigma^2}\right]\right]dP-a(\bar{P}-B_b-B_l)-B_b=0$$

Suppose:

$$F\left(\overline{P},e^*\right)=\int_{r\left(1+e^*\right)}^{+\infty}\left[P-r\left(1+e^*\right)\right]\left[\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{\left(P-\overline{P}\right)^2}{2\sigma^2}\right]\right]dP-a\left(\overline{P}-B_b-B_l\right)-B_b,$$

take the partial derivative with respect to $\overline{P}$ and e^* in $F\left(\overline{P},e^*\right)$, then:

$$F_{e^*}^{l}\left(\overline{P},e^*\right)=-r\int_{r\left(1+e^*\right)}^{+\infty}\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{\left(P-\overline{P}\right)^2}{2\sigma^2}\right]dP=-r\int_{r\left(1+e^*\right)}^{+\infty}f\left(P\right)dP=-rA$$

Suppose $D=r\left(1+e^*\right)$, and plug it into $F\left(\overline{P},e^*\right)$, then:

$$F_{e^*}^{l}\left(\overline{P},e^*\right)=\frac{1}{\sigma\sqrt{2\pi}}\left[\int_D^{+\infty}\left(P-D\right)\exp\left[-\frac{\left(P-\overline{P}\right)^2}{2\sigma^2}\right]dP\right]^l-a$$

$$=\frac{1}{\sigma\sqrt{2\pi}}\left\{\left[\int_D^{+\infty}\left(P-D\right)\exp\left[-\frac{\left(P-\overline{P}\right)^2}{2\sigma^2}\right]dP\right]^l\right.$$

$$\left.+\left[\int_D^{+\infty}\left(\overline{P}-D\right)\exp\left[-\frac{\left(P-\overline{P}\right)^2}{2\sigma^2}\right]dP\right]^l\right\}-a$$

$$=\frac{1}{\sigma\sqrt{2\pi}}\left[-\sigma^2\exp\left[-\frac{\left(P-\overline{P}\right)^2}{2\sigma^2}\right]\Bigg|_D^{+\infty}+\overline{P}\int_D^{+\infty}\exp\left[-\frac{\left(P-\overline{P}\right)^2}{2\sigma^2}\right]dP\right.$$

$$\left.-D\int_D^{+\infty}\exp\left[-\frac{\left(P-\overline{P}\right)^2}{2\sigma^2}\right]dP\right]^l-a$$

$$=\frac{1}{\sigma\sqrt{2\pi}}\left[\left(D-\overline{P}\right)\exp\left[-\frac{\left(D-\overline{P}\right)^2}{2\sigma^2}\right]+\int_D^{+\infty}\exp\left[-\frac{\left(P-\overline{P}\right)^2}{2\sigma^2}\right]dP\right.$$

$$\left.+\left(D-\overline{P}\right)\exp\left[-\frac{\left(P-\overline{P}\right)^2}{2\sigma^2}\right]\Bigg|_D^{+\infty}\right]-a$$

$$=\frac{1}{\sigma\sqrt{2\pi}}\left[\left(D-\overline{P}\right)\exp\left[-\frac{\left(D-\overline{P}\right)^2}{2\sigma^2}\right]+\int_D^{+\infty}\exp\left[-\frac{\left(P-\overline{P}\right)^2}{2\sigma^2}\right]dP\right.$$

$$\left.-\left(D-\overline{P}\right)\exp\left[-\frac{\left(P-\overline{P}\right)^2}{2\sigma^2}\right]\right]-a$$

$$= \int_D^{+\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{\left(P-\overline{P}\right)^2}{2\sigma^2} \right] dP - a = \int_D^{+\infty} f\left(P\right) dP - a = A - a$$

Suppose: $\int_{r(1+e_0)}^{+\infty} f\left(P\right) dP - a = 0$, then: as shown in Fig. 1, where $e^* < e_0$, $\partial e^* / \partial \overline{P} > 0$, the interest rate is low, and expected rate of return of the program is positively correlated to the loan rate. where $e^* > e_0$, $\partial e^* / \partial \overline{P} < 0$, the interest rate is high, and expected rate of return of the program is negatively correlated to the loan rate.

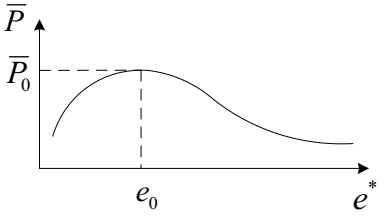


Fig. 1 Relationship of the expected rate of return of the program and the interest rate

In the Internet financial market, general information symmetry is achieved between the loaner and the borrower. The natural risk level of the loan program is generally limited within an acceptable range, and both the lender and borrower have certain pricing power. In cases that the program system risk is lowered, the success probability of the program will be increased, so is the expected rate of return of the program. The lender is certain to adopt relevant strategies to demand the borrower to confer parts of the benefits by virtue of the pricing power. The expected rate of return of the program is positively correlated to the loan rate subsequently.

Where the program system risk is lowered to the maximum extent, the expected rate of return of the program reaches its highest value $\overline{P}_0$, and the loan interest rate then is e_0. If the program system risk grows, the success probability of the program will drop, so is the expected rate of return of the program. The lender is certain to demand higher interest rate by virtue of the pricing power to maintain the stability of $A \times r\left(1+e\right)$, so as to safeguard its own interests. The expected rate of return of the program is negatively correlated to the loan rate subsequently.

As a result, there is no simple positive correlation between the Internet financial interest rate and the expected rate of return for the loan program.

Proposition III. There is a positive correlation between the Internet financial interest rate and the transaction cost of the lender, and a negative correlation with the transaction cost of the borrower.

Demonstration: In expression (6) under the first order optimum conditions, the decisive function of the optimum interest rate is defined implicitly, $e^*\left(C_b, C_l\right)$. Take the derivative with respect to C_b, C_l in expression (6), then:

$$\frac{\partial e^*}{\partial C_l} = \frac{a}{r\int_{r(1+e^*)}^{W} f(P)dP} = \frac{a}{rA} > 0\,,\ \frac{\partial e^*}{\partial C_b} = \frac{1-a}{r\int_{r(1+e^*)}^{W} f(P)dP} = \frac{1-a}{rA} < 0$$

$\partial e^* / \partial C_l > 0$, and $\partial e^* / \partial C_b < 0$ implies that the optimum interest rate is an increasing function of the transaction cost of the lender, and is a decreasing function of the transaction cost of the borrower. In the Internet financial market, both the lender and borrower enjoy certain bargaining power ($0 < a < 1$). Where the transaction cost of the borrower is excessive high, the borrower will take advantage of its bargaining position to lower the loan interest rate, and transfer part of the cost to the lender; similarly, the loan interest rate can be increased to transfer part of the cost to the borrower. Therefore, there is a positive correlation between the Internet financial interest rate and the transaction cost of the lender, and a negative correlation with the transaction cost of the borrower.

Proposition IV. There may not be a positive correlation between the Internet financial interest rate and the bank interest rate S of the same period.

Demonstration: In expression (6) under the first order optimum conditions, the function of $e^*(s)$ is defined. Take the derivative with respect to s in expression (6), then:

$$\frac{\partial e^*}{\partial s} = \frac{a-(1-r)}{r\int_{r(1+e^*)}^{W} f(P)dP} = \frac{a-(1-r)}{rA}$$

Where $a > 1-r$, $\partial e^* / \partial s > 0$, the bargaining power of the borrower is higher. If the loan ratio is high, the Internet financial interest rate is positively correlated to the bank interest rate S of the same period. For instance, if the bank interest rate S of the same period declines, the overall capital environment is loose. The borrower is likely to demand the lender to reduce the interest rate by virtue of its relative strong bargaining power; if the bank interest rate S of the same period grows, the overall capital environment is tense, and the loan ratio of the borrower itself is higher, the borrower is still likely to accept the higher interest rate posed by the lender even if it enjoys a relative strong bargaining power.

Where $a < 1-r$, $\partial e^* / \partial s < 0$, the ratio of self-owned capital of the borrower is high, the bargaining power of it is lower, the Internet financial interest rate is negatively correlated to the bank interest rate S of the same period. For example, in cases that the bank interest rate S of the same period declines, the capital demander with low risk and low interest rate in the Internet financial market turns to the formal financial market, while the capital demander with high risk and high interest rate remains in the Internet financial market. To achieve the expected profits, the capital suppliers with high bargaining power in the Internet financial market increase the interest rates, and the Internet financial interest rates will be on the rise conversely; in cases that interest rate S of the same period increases, the formal financial market pushes out some capital demanders with high security, and they turns to informal financial market, which demands the Internet financial market to lower the interest rates. Meanwhile, the original formal financial capital which is supposed to be loaned to the capital demanders with low interest rates is ready to be loaned to capital

demanders in the Internet finance who are willing to pay higher interest rates, and they will flow to the formal financial market. Subsequently, the capital suppliers in the Internet financial market are forced to lower the interest rates, while the lower risk in the Internet financial market makes the capital suppliers in the Internet financial market willing to reduce the interest rates. The interest rates in Wenzhou private lending drops rather than grows when the bank interest rates are raised slightly (Wang Yiming, Li Minbo, 2005), which is an excellent example of the above theories.

IV Conclusion

The Internet finance discloses the credit information of the borrower and the loan program information to the lender. As a result, the uncertainty of the loan projects in themselves and the costs for the implementation of secured credit contracts, rather than the reverse selection and moral risks arising from information asymmetry, form the main issues faced by the Internet finance. They also constitute the decisive factors of the loan interest rate of the Internet finance.

According to the analysis in the article, the Internet financial interest rate is a function of the natural risk of the loan program, the expected rate of return of the program, the loan ratio, the transaction cost, the bank interest rate S of the same period, and the relative bargaining power of the borrower and the lender. Effects of these variants on the Internet financial interest rate are also illustrated. Both the natural risk function and bargaining power function are introduced in the article with multiple factors affecting the interest rate included into the analysis framework of the model so as to illustrate the following propositions:

(1) There is a positive correlation between the natural risk level of the loan program and the interest rate. (2) The expected rate of return of the program is positively correlated to the loan rate where the program risk declines; while the expected rate of return of the program is negatively correlated to the loan rate where the program risk increases. (3) There is a negative correlation between the interest rate and the transaction cost of the borrower, and a positive correlation with the transaction cost of the lender. (4) There may not be a positive correlation between the Internet financial interest rate and the bank interest rate S of the same period. (5) The interest rate is positively correlated to the loan ratio if the interest rate is low; while the interest rate is negatively correlated to the loan ratio if the interest rate is high; (6) The loan interest rate is negatively correlated to the competition extent of lenders, and is positively correlated to the competition extent of borrowers. (7) There is a positive correlation between the loan urgency, the loan period, the regime risk level and the interest rate.

To sum up, the interest rate risk is common in the whole financial industry which includes the Internet financial industry. The marketization of the Chinese financial system, and the acceleration of the interest rate liberalization pose a great challenge for the emerging Internet financial industry. As for now, the Internet finance has suffered no severe impact of the loan interest rate liberalization. However, the Internet financial enterprises will certainly face more competition and challenges after the deposit interest rate liberalization. The operation cost is the advantage of Internet financial enterprises. They also provide financial products with higher rates of return.

The advantage of the rates of return will be weakened after the fully liberalization of the interest rates, which will result in loss of customers, increase in sensibility to the variations of interest rates, and amplification of the impact of the liberalization of the interest rates on the Internet financial industry.

References

[1] Tian Suning. Review: Reform by "Cloud Calculation" [J]. China Finance Review, 2012, (1).

[2] Wan Jianhua. Review: Innovation of Internet Financial Model and Future Financial Variations [J]. China Finance Review, 2012, (1).

[3] Xie Ping, Zou Chuanwei. Research on Internet Financial Model [J]. Journal of Financial Research, 2012, (12).

[4] Hamilton A. The financial revolution [M].Penguin, 1986.

[5] Hamilton J D.A new approach to the economic analysis of nonstationary time series and the business cycle[J].Econometrica,1989,57(2),357-384.

[6] Hester D D. Financial disintermediation and policy[J].Journal of Money, Credit and Banking,1969,1(3):600-617.

[7] Krolzig H M. Markov-switching vector autoregressions: Modeling, statistical inference and applications to business cycle analysis[M].Springer,1997.

[8] Mishkin F S. The economics of money, banking, and financial markets [M]. Addison-Wesley ,2001.

[9] Roldos J. Disintermediation and monetary transmission in Canada[R]. IMF Working Paper,2006.

[10] Tan A C, K Goh.Financial disintermediation in the 1990s:Implications on monetary policy in Malaysia[R].Paper Prepared for the 15th Annual Conference on Pacific Basin Finance, Economics, Accounting and Management,2007.

[11] Xu Rong, Liu Yang, Wen Wujian and Xu Zhao. Research on Potential Risks of Internet Finance, Financial Regulation Research, 2014, (3), 40-56.

[12] Zhang Xiaopu. Principles on Internet Financial Regulation: Exploration on New Financial Regulation Modes, Financial Regulation Research, 2014, (2), 6-17.

[13] Zheng Liansheng, Liu Liang, and Xu Jianjun. Statuses, Modes and Risks of Internet Finance – Research on American Experiences, Financial Market Research, 2014, (2), 41-49.

[14] He, D., and H. Wang, Dual-track Interest Rates and the Conduct of Monetary Policy in China, China Economic Review, 2012, 23(4), 928-947.

[15] Porter, N., and T. Xu, What Drives China's Interbank Market? IMF Working Paper, 2009, No. WP/09/189.

[16] Tarhan, F., N. Porter, and E. Takáts, Interest Rate Liberalization in China, IMF Working Paper, 2009, No. WP/09/171.

[17] Cheng Kun, Decisive Mechanism for Informal Financial Interest Rates- Drop Asymmetric Nash Bargaining Solution, Shanghai Journal of Economics, 2006, (5), 37-45

[18] Shamsul Kamal Ahmad Khalid; Mustafa Mat Deris; Kamaruddin Malik Mohamad. Anti-cropping digital image watermarking using Sudoku DOI: http://dx.doi.org/10.1504/IJGUC.2013.056253.178-186

[19] Jinson Zhang; Mao Lin Huang; Doan Hoang. Visual analytics for intrusion detection in spam emails DOI: http://dx.doi.org/10.1504/IJGUC.2013.056254.187-196

[20] Sujuan Li; Futai Zhang Leakage-resilient identity-based encryption scheme . DOI: http://dx.doi.org/10.1504/IJGUC.2013.056255.197-203

[21] Yuanyuan Zhang; Marine Minier. How network coding system constrains packet pollution attacks in wireless sensor networks DOI: http://dx.doi.org/10.1504/IJGUC.2013.056256.204-211

[22] Guangye Sui; Mohamed Mejri. FASER (Formal and Automatic Security Enforcement by Rewriting) by BPA algebra with test DOI: http://dx.doi.org/10.1504/IJGUC.2013.056257.212-218

Toward Construction of Encryption with Decryption Awareness Ability for Cloud Storage

Xu An Wang, Fatos Xhafa, GuangMing Wu, Wei Wang

Abstract In cloud platforms, nowadays more and more data owners prefer to first encrypt their personal files and then outsource them to the cloud storage. This however, poses several challenges such as those related to access control and rights. Most previous research results care more about the fine-grained access to the encrypted private files, while little research has concentrated on the following basic question: have the stored ciphertexts been successfully decrypted and when did this happen? Aiming at tackling this problem, we propose a new notion of encryption, namely, public key encryption with decryption awareness ability, which we denote as PKE-DAA. Concretely speaking, PKE-DAA can realize the following: after a ciphertext has been created by the encrypter, once this ciphertext has been successfully decrypted, the encrypter or some other party could be aware of the decryption! To the best of our knowledge, until now there are no such notions or encryption schemes. Own to the recent cryptographic community research result on obfuscation, especially indistinguishable obfuscation, we try to give such a construction in a weak form. We believe that this new primitive has many potential applications, such as against strongest illegal eavesdropping launched by NSA in the post-Snowden era etc.

Xu An Wang
Key Laboratory of Cryptology and Information Security, Engineering University of CAPF, Xi'an 710086, China, wangxazjd@163.com

Fatos Xhafa
Department of Computer Science, Technical University of Catalonia, Spain, fatos@cs.upc.edu

Guangming Wu
Key Laboratory of Cryptology and Information Security, Engineering University of CAPF, Xi'an 710086, China

Wei Wang
Engineering University of CAPF, Xi'an 710086, China

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_27

1 Introduction

1.1 Background

Since the introduction of public key encryption[7][22] [2][24][15], researchers are paying great attention on how to design encryption schemes, which can correctly decrypt the ciphertext and simultaneously achieve some other interesting properties. The first interesting property is how to achieve the strong security notion such as indistinguishable under chosen plaintext attack and chosen ciphertext attack, many wonderful results have been achieved in this field, such as the CS98, KD04, CKS08, HK09 schemes etc. The second interesting property is how to achieve flexible access control on the decryption of the ciphertexts, in this field many advanced encryption paradigms have been proposed such as fuzzy identity based encryption, attribute based encryption, predicate encryption etc. [20, 21, 19] [9] [6, 18] [14]. In these advanced encryption paradigms, the decryptor can only get the contents corresponding to his credentials, the encrypter can have flexible control on which one/role/party can decrypt the ciphertexts, these advanced paradigms break the one-to-one barrier and extend our ability to the advanced one-to-many field. The third interesting property is how to implement complex homomorphic computation on the ciphertexts, such as "Xor", "ADD", "Multiplication" , "Compare", "Average", "Sort" etc., fully hommorphic encryption is a breakthrough research result in this field[4] [5][10][11] [17], which can support any arbitrary complex computation.

1.2 Motivation

It should be noted all these kinds of schemes pay great attention on how to deal with the encrypted content, but pay little attention on the "event-driving property". A basic such property on public key encryption is the following: *"After a ciphertext has been created, can the encrypter be aware of whether it has been decrypted successfully? Or even, can anyone publicly be aware of this?"*

At a first glance, this task may seem impossible. After the decryptor get the correct secret key, he can decrypt the corresponding ciphertext correctly at any time. After the encrypter creates the ciphertext, he can play no more role on the decryption. We denote this paradigm of encryption as public key encryption with decryption awareness ability (PKE-DAA). This notion of encryption is somewhat orthogonal with the traditional public key encryption/attribute based encryption/functional encryption/fully homomorphic encryption, it emphasis on the decryption event instead of the decrypted content. Fig.1 describes the traditional public key encryption while Fig. 2 describes a basic public key encryption with decryption awareness ability.

Here we compare the features of these kinds of PKE schemes, the comparison results can be seen in Table 1.

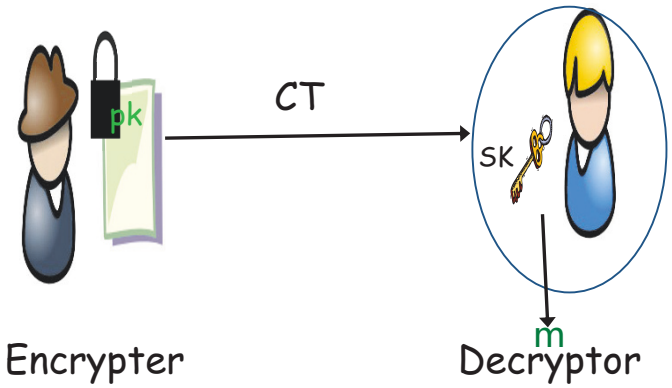


Fig. 1 Traditional Public Key Encryption.

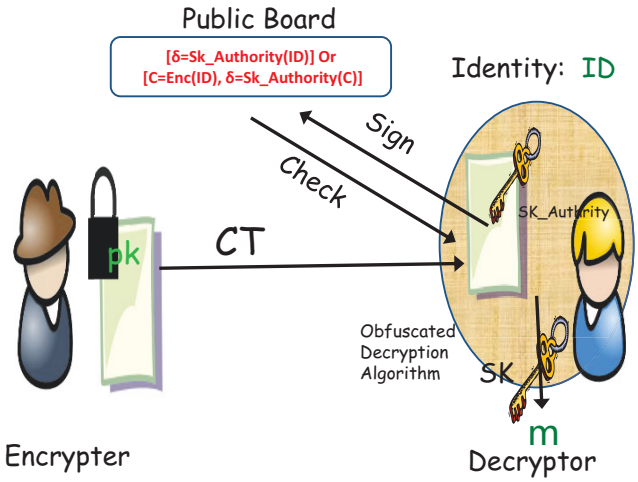


Fig. 2 Public Key Encryption with Decryption Awareness Ability.

Table 1 Comparison among several existing kinds of PKE schemes.

Categorizes	Emphasis on	From the point view of
IND-CPA/IND-CCA PKE	Level of security	Adversaries
FIBE/ABE	Access control on the decrypted contents	Decryptors
FE	Functional results of the decrypted contents	Decryptors
FHE	Full homomorphic properties of encryption	Operators
PKE-DAA	Awareness of the successful decryption	Encrypter or other auditors

1. For PKE schemes with IND-CPA, IND-CCA properties, the adversary cannot distinguish, which one corresponding to the challenge ciphertext from two candidate plaintexts under chosen plaintext or chosen ciphertext attack. This notion of PKE emphasises on the strong level of security, from a point view of adversaries.
2. For advanced PKE schemes, like fuzzy identity based encryption/attribute based encryption (FIBE/ABE), the encrypter can selectively control on which subset of potential receivers can decrypt the ciphertexts. This notion of PKE emphasises on the access control of the decryption of ciphertexts, from a point view of potential decryptors.
3. For advanced PKE schemes like functional encryption (FE), besides the encrypter can selective control on which can decrypt the ciphertexts, the decryptor can obtain the function result on the decrypted content. This notion of PKE emphasises on the functional result of the decrypted content, from a point view of different decryptors.
4. For advanced PKE schemes like fully homomorphic encryption (FHE), any one can manipulate on the ciphertexts just like on the plaintexts but blindly, and get the encrypted manipulated result. This notion of PKE emphasises on the homomorphic properties of encryption, from a point view of any operators.
5. For our proposed PKE-DAA schemes, the encrypter or anyone can be aware of the event of successful decryption. This notion of PKE emphasises on the decryption event-driven property, from a point of the encrypter or other auditors.

Roughly we can categorize PKE-DAA under the following properties:

1. The first property, any one can learn whether the ciphertext has been decrypted correctly or not;
2. The second property, only the encrypter can learn whether the ciphertext has been decrypted correctly or not;
3. The third property, any one can learn who has decrypted the ciphertext;
4. The fourth property, only the encrypter can learn who has decrypted the ciphertext.

Of course, there other variants of PKE-DAA can be designed.

1.3 Our Contribution

1. We first propose the concept of public key encryption with decryption awareness ability, and give such a construction. We also discuss the difference of this new primitive with other kinds of public key encryption schemes.
2. We first construct such a primitive based on indistinguishable obfuscation [16][1] [12] [13] [3] [8] [23] and roughly analyse its security.
3. We also discuss several interesting applications of PKE-DAA, especially in cloud computing context. We also believe that there are many other applications to be exploited by using this new primitive.

1.4 Related works

The concept of obfuscation was first proposed by soft engineers, its main purpose is to protect the software against invalid copying. Hada [16] first introduced the concept of obfuscation to the cryptographic community in Aisacrypt'01, then this first attempt drew great attention from many theoretical cryptographers, they tred to give a provable secure construction of obfuscation based on hard cryptographic problems, but this try failed. Barak et al. [1] proved that there can be no existing such an obfuscation if the security proof reduction is a black box type. However considering the great potential use of obfuscation, cryptographers try to give a weak construction but still be useful enough. In 2005, Goldwasser et al. [12] proposed the concept of best possible obfuscation and proved that such obfuscation are stronger than ever existing obfuscation schemes and can achieve the best possibility. This is an amazing concept which point out a new goal for obfuscation construction. In 2007, Goldwasser et al. [13] proposed a new kind of obfuscation called indistinguishable obfuscation (iO), this obfuscation can be defined as following: any two program having same functionality can be obfuscated to two indistinguishable unreadable programs. At first sight, this kind of obfuscation is not strong, but Goldwasser et al. proved that this kind of obfuscation can be reduced to best possible obfuscation and the other way also holds. In 2013, Garg et al.[8] first gave a plausible construction of indistinguishable obfuscation based on the multilinear map, which also was achieved only in recent years. Since then, more concrete construction of iO have been proposed and more security analysis results were also presented on multilinear map. Here, we concentrate on the application of iO. Boneh and Zhandry [3] proposed a very efficient multi-party non-interactive key agreement protocol, high efficient broadcast encryption and traitor tracing scheme with almost optimal parameters. Sahai and Waters [23] use iO and a new cryptographic primitive-punctured program to solve the open problem of construction of undeniable encryption, which has been left open for many years, they also gave several other constructions of CCA-secure PKE, non-interactive zero knowledge, signature etc. based on iO. Their result shows that iO is a very powerful tool, although at first seemed an impossible concept, it can be concretely constructed, therefore this observation also inspires our exploration of PKE-DAA, which seems to be a very difficult primitive to be achieved.

1.5 Organization

We organize the paper as the following. In section 2, we review the preliminaries which need to understand our work. In section 3, we give the PKE-DAA's formal definition, we present our construction of PKE-DAA and roughly analysis our construction security. In section 4, we discuss PKE-DAA's several interesting applications, especially on secure health monitoring cloud storage. Finally in section 5, we conclude our work with some interesting open problems.

2 Definition

In this section, we give a very basic definition of public key encryption with decryption awareness ability (PKE-DAA) and its security model. PKE-DAA consists of the following algorithm PKE-DAA.KeyGen, PKE-DAA.Enc, PKE-DAA.Dec, PKE-DAA.Ver:

1. PKE-DAA.KeyGen$(1^k) \rightarrow (PK, SK)$: The randomized key generation algorithm KeyGen takes as input a security parameter 1^k and outputs a public key PK and a secret key SK. We write $(PK, SK) \leftarrow KeyGen(1^k)$.
2. PKE-DAA.Enc$(PK, m) \rightarrow C$: The randomized encryption algorithm Enc takes as input a public key PK and a message $m \in \{0,1\}^*$, and outputs a ciphertext C, We write $C \leftarrow \mathsf{Enc}(\mathsf{PK}, \mathsf{m})$.
3. PKE-DAA.Dec$(PK, m) \rightarrow C$: The decryption algorithm $\mathscr{D}$ takes as input a ciphertext C and a secret key SK. It returns a message $m \in \{0.1\}^*$ or the distinguished symbol $\perp$. We write $m \leftarrow \mathscr{D}_{SK}(C)$. However, this algorithm also has the ability of notifying the encrypter (or other ones) the ciphertexts have been decrypted.

We require that for all (PK, SK) output by Gen, all $m \in \{0,1\}^*$, and all C output by $\mathscr{E}_{PK}(m)$ we have $\mathscr{D}_{SK} = m$.

3 Our Proposal

At first sight, PKE-DAA seems to be impossible to be constructed. A natural way to construct PKE-DAA is the following: when the decryptor decrypts the ciphertexts by using his secret key, besides outputting the plaintext, the algorithm also outputs a SIGNAL to indicate if the decryption is whether successful or not. However the following malicious attacker scenarios needs to be considered:

- The outside attacker can forge such a false SIGNAL and thus give the encrypter or the outsider checkers a false decision. Thus how to design an unforgeable SIGNAL is necessary for achieving security.
- The malicious decryptor, such as a hacker who although has a correct secret key but do not want the encrypter or the outside checkers to know this SIGNAL, he can hide such a SIGNAL when decryption or let the decryption algorithm do not generate such a SIGNAL.

Thus we need to solve the above two challenges if we want to design meaningful PKE-DAA schemes. The intuition of our construction is the following:

1. Our first attempt is this: the decryption algorithm needs to implement the standard decryption functionality, additionally there is needed to be embedded the secret key of the encrypter or the outsider authortity (the SIGNAL checker), there is also needed to be added the program of signing the ciphertext (or the identity

of the decryptor) by using the embedded secret key. Finally, the decryption algorithm is obfuscated by the indistinguishable obfuscation to be unreadable, thus the attacker cannot get the embedded secret key and thus to forge the SIGNAL. This design intuition is that the decryption algorithm needs to output the plaintext and the SIGNAL simultaneously and the encrypter or the outsider checker can check the SIGNAL to indicate whether the decryption event has happened.

2. However the first attempt omits the following attack: the malicious decryptor can hide the SIGNAL or prevent the SIGNAL to be communicated to the encrypter or to the outsider checker. Let us suppose the following scenario: A high confidential file has been stored in a very well protected server, a malicious hacker wants to decrypt this high confidential file. Suppose the malicious hacker has obtained the correct secret key. He thus first downloads the ciphertext from the server, he stops to access to the internet while the server resides, and runs the decryption algorithm totally locally. This time, although the decryption algorithm outputs the plaintext and the SIGNAL simultaneously, but the encrypter or the outsider checker does not receive any SIGNAL, and the hacker successfully completes his job!
3. The second attempt: besides the first attempt, we also need a public board which can be accessed by the decryption algorithm and the SIGNAL checker. In the decryption algorithm, besides the signature signing by the embedded secret key, it also needs to add this program: the signature needs first to outsouce and record to the public board and later it also to check this signature has been indeed recorded on the public board, only after that the decryption algorithm can output the plaintext, if it checks that there is no such a signature on the public board, the decryption algorithm output $\perp$.

3.1 A Concrete Proposal

The figure 3 is the description of the program of decryption algorithm of our PKE-DAA scheme, the real decryption algorithm is the obfuscated result of this program by using indistinguishable obfuscation. In this code, the "Authority" means the outside SIGNAL checker. In this program, the embedded secret key is $sk_{Authority}$, which is used to generate the SIGNAL of successful decryption, the signature on the decrypted ciphertext. If this signature has been outsourced and recorded successfully on the trustful public board, the decryption algorithm output m, otherwise it outputs $\perp$ and the decryptor cannot get any information. The above pesudo-code can describe three different runs. The first running is the construction in the weakest security model, which does not run step c, d, e, in this model we do not consider the decryptor malicious to hide the output SIGNAL; the second running is the construction in the strong security model but not outsourcing the identity of the decryptor to the public board, who runs step c, d, e but the signature form is $\sigma(c)$; the third running is the construction in the strong security model but also outsourcing the

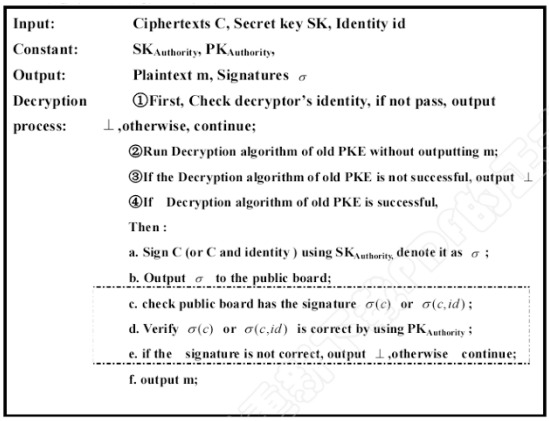


Fig. 3 The iO of decryption algorithm of our PKE-DAA scheme

identity of the decryptor to the public board, who runs step c, d, e and the signature form is $\sigma(c, id)$ where id is the identity of the decryptor.

4 Application

We consider the following applications for PKE-DAA scheme.

1. PKE-DAA can be used to protect high confidential files against strongest illegal eavesdropping launched by NSA in the post-Snowden era. Suppose National Defense Department has a high confidential file, which encrypted by PKE-DAA, and the ciphertexts are stored in a trusted server, which will be protected very well. The ciphertexts can be decrypted by only very few commanders in a presumed period. But if one day the encrypter (the officer who is responsible to protect this file) can be aware the encrypted file has been decrypted by some commander A but not in the presumed period, then the encrypter can be suspicious the encrypted file has been broken, he can be assured the file was indeed by the commander A by first contacting with A. If not, he can do something to resist this attack and alert others this file has been broken, also he can be aware commander A's secret key has been stolen.
2. PKE-DAA can be used to implement high confidential communication between the spy and the intelligence agency. Suppose spy A aboard wants to transfer a high confidential file to his own country's intelligence agency. He can use PKE-DAA to encrypt this confidential file and upload the ciphertexts to a public server in the Internet. Later when the intelligence agency decrypt the confidential file, the spy A will be aware of this, thus he can delete the confidential file after this decryption. In this way the risk of being caught will be greatly reduced.
3. PKE-DAA can also be used to efficiently manage the finite memory of restricted resource equipments such as mobile phones, embedded information processing

unit in vehicles, etc. Suppose mobile phone user A wants to download some confidential files from the cloud, then the cloud can use PKE-DAA to encrypt the files. The ciphertexts will be first stored in the finite memory of the mobile phones. When the mobile phone user A decrypts the ciphertexts, the cloud can be aware of this, thus it can push out the decrypted the ciphertext from the finite memory and load the new ciphertexts to the finite memory.

5 Conclusion

In this paper, for the first time we propose the concept of public key encryption with decryption awareness ability, which we denote as PKE-DAA scheme. We also try to realize such a primitive in a weak form based on indistinguishable obfuscation. We remark here we only give a very basic construction just for demonstrating such a possibility. Of course, more work is needed to be done, such as more rigorous definition and security model of PKE-DAA, more carefully design and security proof, etc., which we plan to address in our future work.

Acknowledgements This work is supported by the Natural Science Foundation of Shaanxi Province (Grant No. 2014JM8300).

References

[1] Barak B, Goldreich O, Impagliazzo R, Rudich S, Sahai A, Vadhan SP, Yang K (2001) On the (im)possibility of obfuscating programs. In: Kilian J (ed) CRYPTO 2001, Springer, Berlin, Germany, Santa Barbara, CA, USA, LNCS, vol 2139, pp 1–18

[2] Boneh D, Franklin MK (2001) Identity-based encryption from the Weil pairing. In: Kilian J (ed) CRYPTO 2001, Springer, Berlin, Germany, Santa Barbara, CA, USA, LNCS, vol 2139, pp 213–229

[3] Boneh D, Zhandry M (2014) Multiparty key exchange, efficient traitor tracing, and more from indistinguishability obfuscation. In: Garay JA, Gennaro R (eds) CRYPTO 2014, Part I, Springer, Berlin, Germany, Santa Barbara, CA, USA, LNCS, vol 8616, pp 480–499, DOI 10.1007/978-3-662-44371-2_27

[4] Brakerski Z, Vaikuntanathan V (2011) Efficient fully homomorphic encryption from (standard) LWE. In: Ostrovsky R (ed) 52nd FOCS, IEEE Computer Society Press, Palm Springs, California, USA, pp 97–106

[5] Brakerski Z, Gentry C, Vaikuntanathan V (2012) (leveled) fully homomorphic encryption without bootstrapping. Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, pages 309–325

[6] Chase M, Chow SSM (2009) Improving privacy and security in multi-authority attribute-based encryption. In: Al-Shaer E, Jha S, Keromytis AD (eds) ACM CCS 09, ACM Press, Chicago, Illinois, USA, pp 121–130
[7] Diffie W, Hellman ME (1976) New directions in cryptography. IEEE Transactions on Information Theory 22(6):644–654
[8] Garg S, Gentry C, Halevi S, Raykova M, Sahai A, Waters B (2013) Candidate indistinguishability obfuscation and functional encryption for all circuits. In: 54th FOCS, IEEE Computer Society Press, Berkeley, CA, USA, pp 40–49
[9] Garg S, Gentry C, Halevi S, Sahai A, Waters B (2013) Attribute-based encryption for circuits from multilinear maps. In: Canetti R, Garay JA (eds) CRYPTO 2013, Part II, Springer, Berlin, Germany, Santa Barbara, CA, USA, LNCS, vol 8043, pp 479–499, DOI 10.1007/978-3-642-40084-1_27
[10] Gentry C, Halevi S, Smart NP (2012) Fully homomorphic encryption with polylog overhead. In: Pointcheval D, Johansson T (eds) EUROCRYPT 2012, Springer, Berlin, Germany, Cambridge, UK, LNCS, vol 7237, pp 465–482
[11] Gentry C, Sahai A, Waters B (2013) Homomorphic encryption from learning with errors: Conceptually-simpler, asymptotically-faster, attribute-based. In: Canetti R, Garay JA (eds) CRYPTO 2013, Part I, Springer, Berlin, Germany, Santa Barbara, CA, USA, LNCS, vol 8042, pp 75–92, DOI 10.1007/978-3-642-40041-4_5
[12] Goldwasser S, Kalai YT (2005) On the impossibility of obfuscation with auxiliary input. In: 46th FOCS, IEEE Computer Society Press, Pittsburgh, PA, USA, pp 553–562
[13] Goldwasser S, Rothblum GN (2007) On best-possible obfuscation. In: Vadhan SP (ed) TCC 2007, Springer, Berlin, Germany, Amsterdam, The Netherlands, LNCS, vol 4392, pp 194–213
[14] Gorbunov S, Vaikuntanathan V, Wee H (2013) Attribute-based encryption for circuits. In: Boneh D, Roughgarden T, Feigenbaum J (eds) 45th ACM STOC, ACM Press, Palo Alto, CA, USA, pp 545–554
[15] Goyal V, Pandey O, Sahai A, Waters B (2006) Attribute-based encryption for fine-grained access control of encrypted data. In: Juels A, Wright RN, Vimercati S (eds) ACM CCS 06, ACM Press, Alexandria, Virginia, USA, pp 89–98, available as Cryptology ePrint Archive Report 2006/309
[16] Hada S (2000) Zero-knowledge and code obfuscation. In: Okamoto T (ed) ASIACRYPT 2000, Springer, Berlin, Germany, Kyoto, Japan, LNCS, vol 1976, pp 443–457
[17] Halevi S, Shoup V (2014) Algorithms in HElib. Cryptology ePrint Archive, Report 2014/106, http://eprint.iacr.org/2014/106
[18] Lewko AB, Waters B (2011) Decentralizing attribute-based encryption. In: Paterson KG (ed) EUROCRYPT 2011, Springer, Berlin, Germany, Tallinn, Estonia, LNCS, vol 6632, pp 568–588
[19] Lewko AB, Waters B (2012) New proof methods for attribute-based encryption: Achieving full security through selective techniques. In: Safavi-Naini R, Canetti R (eds) CRYPTO 2012, Springer, Berlin, Germany, Santa Barbara, CA, USA, LNCS, vol 7417, pp 180–198

[20] Lewko AB, Okamoto T, Sahai A, Takashima K, Waters B (2010) Fully secure functional encryption: Attribute-based encryption and (hierarchical) inner product encryption. In: Gilbert H (ed) EUROCRYPT 2010, Springer, Berlin, Germany, French Riviera, LNCS, vol 6110, pp 62–91

[21] Okamoto T, Takashima K (2010) Fully secure functional encryption with general relations from the decisional linear assumption. In: Rabin T (ed) CRYPTO 2010, Springer, Berlin, Germany, Santa Barbara, CA, USA, LNCS, vol 6223, pp 191–208

[22] Rivest RL, Shamir A, Adleman LM (1978) A method for obtaining digital signature and public-key cryptosystems. Communications of the Association for Computing Machinery 21(2):120–126

[23] Sahai A, Waters B (2014) How to use indistinguishability obfuscation: deniable encryption, and more. In: Shmoys DB (ed) 46th ACM STOC, ACM Press, New York, NY, USA, pp 475–484

[24] Sahai A, Waters BR (2005) Fuzzy identity-based encryption. In: Cramer R (ed) EUROCRYPT 2005, Springer, Berlin, Germany, Aarhus, Denmark, LNCS, vol 3494, pp 457–473

Elastipipe: On Providing Cloud Elasticity for Pipeline-structured Applications

Rodrigo da Rosa Righi, Mateus Aubin, Cristiano André da Costa, Antonio Marcos Alberti and Arismar Cerqueira Sodre Junior

Abstract Although offering clear benefits for Web and business-critical demands, the use of cloud elasticity still imposes challenges when trying to reap its benefits over complex applications such as those modeled as pipelines. This often happens because replication, the standard technique for resource reorganization, by default, doesn't address both function-level parallelism and communication between working VMs. Taking into account this background, here we are proposing a model named Elastipipe to provide automatic elasticity over pipelines-structured applications. Our main goal is to reduce total execution time for a set of tasks in a way that is effortless to cloud users, eliminating the need for any preconfiguration. Elastipipe's contribution consists in a framework designed to provide a new concept named flexible superscalar pipelines, in which scaling operations and load balancing take place over different elasticity units. An elasticity unit refers to a set of one or more stages of a pipeline that will be grouped together under the same elasticity rules. Using a real workload from an IT Brazilian company, the Elastipipe prototype presented performance gains of up to 60% when confronted with a non-elastic approach. In addition, we demonstrated that the functional decomposition of pipeline stages (CPU-bound, memory-bound, and so on) in corresponding elasticity units was responsible for the best results in terms of performance and cost.

Rodrigo da Rosa Righi, Mateus Aubin, Cristiano André da Costa
Applied Computing Graduate Program - Unisinos - Brazil. e-mail: {rrrighi, aubin, cac}@unisinos.br

Antonio Marcos Alberti and Arismar Cerqueira Sodre Junior
Instituto Nacional de Telecomunicações - INATEL - Brazil. e-mail: {alberti, arismar}@inatel.br

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_28

1 Introduction

Cloud computing has been gaining increasing adherence particularly because its ability to auto-scale systems [5]. The elasticity service is better represented on Web-based business-critical systems, which must satisfy certain SLA (Service Level Agreement), *e.g.*, upper bounds on user perceived response time. Besides acting on performance and energy saving issues, horizontal and/or vertical elasticity actions are also especially pertinent in dynamic environments, where human intervention is becoming increasingly difficult or even impossible [3]. The standard technique relies on using a single load balancer that accepts incoming requests and performs scaling in (consolidation) and out (addition) operations over VM (virtual machine) replicas. This interaction is loosely-coupled by nature, where replicas do not establish communication among themselves and only cooperate in a request-response manner to the dispatcher.

Web applications like e-commerce and EFT (Electronic Funds transfer) take advantage from the aforementioned loosely-coupled cloud elasticity mechanism to offer a better experience to end users [13]. In the same way, this idea is being increasingly explored in the context of HPC (High Performance Computing) systems, so it is desirable to provide automatic resource reorganization on Master-Slave, Bag of Tasks and MapReduce applications [12]. In this scope, the idea is to explore data-parallelism since each VM, or replica, executes the same code over different incoming requests. However, this kind of elasticity design is unable to address complex applications where unpredictable load behaviors and interactions between the processes takes place. Such interactions, for example, may include many data dependencies and trigger process communication in which source and destination entities are only resolved at runtime.

Aiming to mitigate this problem, the state-of-the-art presents as its main alternative the use of elasticity-specific APIs, forcing users to change and insert directives in their source code to get the benefits from this cloud capability [2, 6, 12, 15]. Although feasible, this solution could be classified as time-consuming and not portable (addressing a particular application and infrastructure duet). In this context, we are investigating how to efficiently integrate complex applications and cloud elasticity; more precisely, we are focusing on pipeline-structured applications. A pipeline is a set of processing elements, or stages, connected in series where the output of one element serves as the input to the next one. In addition to being used by the CPU and GPU industries, pipelines are also present in HPC workflows and production line procedures to explore task-level parallelism along different stages [7].

Concerning the aforementioned background, this article presents Elasitipipe: an elasticity model that provides on-the-fly resource reorganization for pipeline applications at the SaaS (Software-as-a-Service) level. The user simply input tasks to be processed in the pipeline, which is managed by the cloud with scaling in and out operations. The resource reorganization is relevant

for avoiding executing in both under- and over-loaded states, while obtaining equal or better execution time when compared to non-elastic pipelines. Elastipipe's scientific contribution consists in a framework to explore the notion of flexible superscalar pipelines in which the application administrator can design elasticity units for an arbitrary set of stages. For example, suppose stages $s1$ and $s2$ are CPU-bound, in this case we can map both to a single elasticity unit: each scaling operation will create or consolidate a CPU-optimized VM that better accommodates them. Based on the proposed model, we developed a prototype that was tested with real workload traces from an IT Brazilian company, both in terms of stages and input tasks definitions. The results are promising, with performance gains of up to 60%, encouraging the company to migrate to the cloud and to adopt this elasticity model to handle future workloads.

2 Related Work

Cloud elasticity is increasingly explored both in public and private cloud providers (including for example Amazon Web Services, Azure and RightScale for the first group, and Eucalyptus, OpenStack and OpenNebula, for the latter). Many research initiatives are now using these providers and platforms, exploring cloud elasticity through horizontal (scaling in and out VM instances) with replication techniques (producing templates that will serve as building blocks for scaling operations) [2, 6, 8, 9, 10, 11, 12, 13, 14, 15]. Among the previous considered articles, we selected four of them which are strongly directed to handling complex workloads and complement our strategy [6, 9, 14, 15].

Marshall, Keahey and Freeman [9] proposed Elastic Site which efficiently adapts services provided within a site, such as batch schedulers, storage archives, or Web services to take advantage of elastically provisioned resources. The idea is to securely extend existing physical clusters into the cloud, where additional worker nodes are dynamically acquired or released from the cloud based on changes to the cluster's job queue. Tran, Skhiri and Zimányi [14] developed the EQS (Elastic Queue Service), an elastic message queue architecture and scaling algorithm which can be adapted to any message queue in order to make it scale elastically. According to them, existing message queues are only able to scale for a certain number of clients and are not able to scale out elastically. The EQS architecture is layered onto multiple distributed components, each one with its own elasticity rules. The authors also proposed a user-level load balancing algorithm, which is capable of rebalancing work after elasticity actions.

Imai, Chestna and Varela [6] presented COS (the Cloud Operating System), a middleware framework to support autonomous workload elasticity and scalability based on application-level migration as a reconfiguration strategy. While other frameworks (for example, MapReduce or Google App En-

gine) force application developers to write programs following specific APIs, COS provides scalability in a general-purpose programming framework based on an actor-oriented programming language. This strategy is pertinent to enhance the migration time, since only the application state is actually migrated. Zhang et al. [15] proposed an elastic application model that enables the use of cloud resources to augment the capability of resource-constrained mobile devices. Their main idea consists in the partition of a single application into multiple components called weblets, which can be executed on a mobile device or migrated to the cloud. Acting similarly to SOA (Service-Oriented Architecture) concepts, weblets allow user-level migration and, assuming they operate over HTTP, can facilitate the interaction with cloud computing providers. Thus, an elastic application can augment the capabilities of a mobile device including computation power, storage, and network bandwidth usage with dynamic cloud execution, configured according to the device's status, including CPU load, memory, battery level, network connection quality, and user preferences.

3 Elastipipe Model

This section presents the Elastipipe model, which was developed at the SaaS level of a cloud to address elasticity over pipeline-structured applications. As a performance-driven initiative, our idea is to exploit reorganization to faster compute a set of tasks in an effortless way to the users who won't need to change the application's code. Consequently, in the system's viewpoint, such time enhancement also contributes to achieving a better throughput. In addition to scaling out operations that are linked directly to the application performance, Elastipipe also supports scaling in actions, focusing on better resource usage and energy consumption. Figure 1 depicts the Elastipipe architecture. It can be seen as a framework that facilitates the migration of pipeline applications to the cloud, offering reductions both in costs and risks perspectives [11]. In Figure 1, Elasticity Management and Resource Monitoring modules are commonly supported by private and public cloud providers through a graphical or command line interface, while the Task Input, Control and Execution Modules are Elastipipe's responsibility.

3.1 Architecture

This subsection describes Elastipipe's architecture, detailing its modules and interactions. The input element for the system is a task (represented by t_z, numbered from $1 \leq z \leq l$). Each task has a type that indicates the pipeline with the appropriate set of stages to handle it. In addition, a task can present a set of dependencies (for task t_a, we have dependency set D_a composed by arbitrary tasks enclosed in the interval $[1, a-1]$). Thus, task t_a will only enter

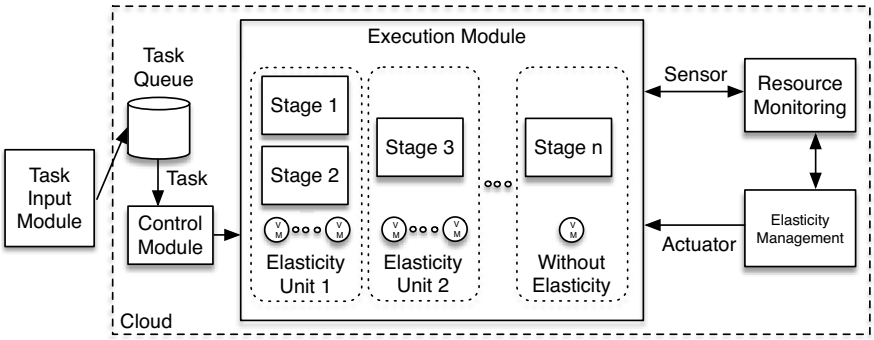


Fig. 1 Elastipipe's architecture. We are detailing an example within the Execution Module: stages can be grouped under elasticity units or do not use elasticity.

in the pipeline when all others in D_a are finished. A pipeline (represented by P_i, numbered from $1 \leq i \leq m$) is a logical unit that is composed by a set of stages (represented by s_j, numbered from $1 \leq j \leq n$) that may share stages with other pipelines. Since our model is of general purpose, we do not define the functionality of each stage, this procedure is conducted later, at implementation time, according to each application's needs.

3.2 Task Input and Control Modules

The Task Input Module is responsible for inserting a task into the queue (here enabled by a message-queue system), so that the Control Module is able to read data from this shared space afterwards. In our case, the queue acts as an element to control the Jitter effect[1], smoothing demand spikes and enabling regular delivery of tasks to the Control Module and fast storage to the Task Input Module. The Control Module is responsible for the following functionalities: (i) getting tasks from the queue; (ii) task validation; (iii) parallelism control; (iv) error handling and; (v) task dispatching and monitoring. After capturing a task to be computed, functionality ii takes place to ensure task integrity and adherence to the Execution Module. The idea is to reduce the quantity of errors in the application due to malformed or unknown messages, avoiding problems which might interfere with the execution of tasks along the pipeline. The parallelism control is associated to a QOS (Quality of Service) definition to manage the pipeline superscalarity. Here we can define minimum and maximum number of concurrently executing tasks, which impacts the number of virtual machines used to handle each elasticity unit (which may contain a single stage or a set of them, more details in Subsection 3.3). Error handling addresses exceptions, as well as retrying tasks that failed

[1] See details about this network metric in: http://howdoesinternetwork.com/2013/jitter

due to communication or infrastructure problems during, for instance, a VM being consolidated with running tasks.

To accomplish the fifth functionality, we designed a dispatcher that identifies which pipeline a particular task belongs to, in addition to controlling pipeline flows and task dependencies. To cope with such cases the strategy adopted is similar to that applied in microprocessor's pipelines, where a data or control hazard might cause a pipeline stall, effectively stopping a subset of the tasks. When situations like these are detected we set the current task at a waiting state until all of its dependencies are completely satisfied. The execution flow of each task is as follows: the Control Module decides the correct pipeline and dispatches a task to the first stage, receiving the computation output and subsequently dispatching the task (and its set of output data) again to the next stage and so on. In other words, based on [11] we designed a master-slave architecture to manage task execution, where the Control Module represents the master and the processing stages, the slaves.

3.3 Elasticity Model

Elastipipe was built to take advantage from resource monitoring and elasticity management already available in cloud providers, therefore, this section aims at classifying Elastipipe's elasticity strategy while also presenting the cloud provider requirements to implement it. The adopted elasticity strategy is horizontal, automatic and reactive (we are adopting the taxonomy proposed by Galante et al. [4]). Concerning the reactive keyword, we are using: (i) upper and lower load thresholds that trigger elasticity operations when violated; (ii) replication based on multiple templates (according to the number of elasticity units); (iii) a load balancer per elasticity unit. This combination was selected while taking into account the following tradeoffs: ease of configuration, accuracy and latency of elasticity actions and availability in major cloud computing environments.

The horizontal elasticity approach is known by scaling in and out operations, *i.e.*, the consolation or addition of worker VMs. The vertical approach (know as resource resizing) was refused because it is limited to the capabilities of a single compute node (CPU, memory, disk and network), not being a scalable and cost effective solution to support pipeline applications. To the user's point of view, resource reorganization takes place automatically, following computational demand, without any intervention at runtime. The only parameters needed for our model are threshold-based elasticity rules and actions which are parameterized in the cloud environment by the system administrator or even by the user himself. Basically, "if" a determined rule is satisfied, "then" an elasticity action is triggered to address the objectives established by the user, which may be, performance, energy, cost, and so on, maybe even a combination of them.

Thresholds, rules and actions are used to control elasticity units. Figure 2 illustrates an example of four deployments, three of them with elasticity units. An elasticity unit is the grouping of one or more stages under the same VM template, then all scaling operations will always instantiate or consolidate this functional element. An elasticity unit has its own rules and actions, in addition to a particular load balancer that decides the best replica (VM instance) to handle a request and its stage (when there are more than a single stage in the template). Load balancers update their connections to the replicas each time the underlying infrastructure changes. Using elasticity units as building blocks, Elastipipe proposes the concept of Flexible Superscalar Pipeline as follows:

- Definition 1. *Flexible Superscalar Pipeline*: It is a concept related to the ability of the user or application administrator to configure as many elasticity units as required to run a pipeline-structured application. Besides performance isolation among elasticity units, a flexible superscalar pipeline presents the following benefits: (i) the number of stage replicas is totally malleable during execution, so presenting different VM deployments per elasticity unit; (ii) and it is possible to to explore functional decomposition by grouping stages according to their resource dependency nature (CPU-, memory-, disk- or network-bound), so creating different VM templates to support particular demands.

It is important to emphasize here that the notion of pipelines is logical, *i.e.*, an elasticity unit can serve multiple pipelines. For instance, we can accommodate stages s_1, s_2 and s_3 in the same elasticity unit, with stages s_1 and s_2 belonging to pipeline P_1 while s_2 and s_3 belong to P_2. Finally, considering that complex applications my comprise many stages, each one with its own resource requirements, the elasticity unit and flexible superscalar pipeline concepts are useful to balance each stage's duration. Such equilibrium is essential as to avoid bottlenecks in the pipelines, since the slowest stage will be responsible for most of the pipeline's runtime [7]. As a welcome side-effect, the usage of elasticity units further enhances the system reliability because they result in replicas that are able to account for crash-failures. Thus, failures over a single component won't take down the whole application [15].

4 Evaluation Methodology

To evaluate our proposal we developed a prototype in which input tasks are represented by text files. As our cloud provider we've selected Amazon Web Services (AWS) and a set of Microsoft technologies to develop the Elastipipe modules: C# compiler, .NET framework and SQL Server database. We are using the following components from AWS: (i) EC2: virtual machines; (ii) CloudWatch: monitoring; (iii) Autoscaling: elasticity management; (iv) Elastic Load Balancing (ELB): load balancing; (v) RDS: relational database; (vi)

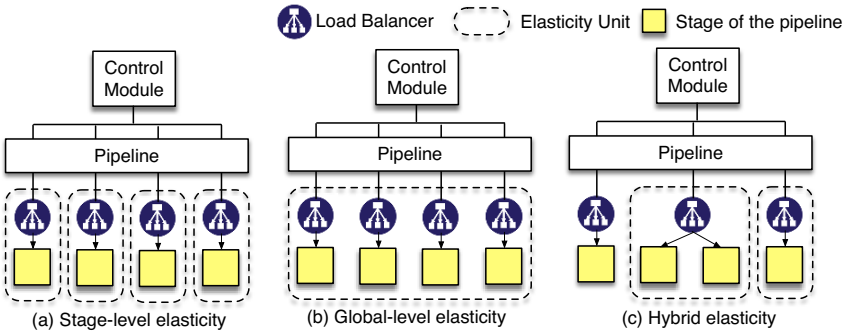


Fig. 2 Elasticity unit and load balancer: Enabling different elasticity strategies according to the behavior of each stage.

SQS: distributed-message queue; and; (vii) S3: storage. The prototype works according to the following execution path: (a) a component first detects the presence of files in S3 and generates a notification message that is inserted into the message queue; (b) a second component, responsible for scanning the queue, detects this notification and inspects the file to analyze which pipeline the task belongs to; (c) taking this information, the Control Module dispatches the task to the appropriate web services to execute each stage of the pipeline.

We implemented two pipelines: P_1 and P_2. P_1 has 5 stages as follows: (i) layout validation (CPU-bound); (ii) layout conversion (CPU-bound); (iii) task consistency and data cross checking (memory-bound); (iv) datastore persistence (disk-bound); (v) distribution stage, which is responsible for sending data reports to other data sources and applications (network-bound). P_2 is comprised of only two stages: (i) schema, which is responsible for inspection and validation according to a predefined layout specified as a XML schema (CPU- and memory-bound); (ii) datastore persistence (disk-bound). Concerning the definition of the stages, we defined 5 elasticity units using the AutoScaling package from AWS: (a) CPU-optimized AutoScaling for validation, conversion and schema stages; (b) Memory-optimized AutoScaling for the consistency stage; (c) Network-optimized AutoScaling for the distribution stage; (d) Disk-optimized AutoScaling for for the datastore persistence stage; (e) global AutoScaling in which all stages are placed together in the same elasticity unit.

4.1 Defining the Workload and Evaluated Metrics

We used a workload that represents a sample from a Brazilian enterprise located in the Tecnosinos tech-park. The workload refers to traces captured during a single week, for a total of 11.459 input tasks (in the form of text files). Each file ranges from 3,500 to 480,000 lines of data, accounting for up

Table 1 Proposed scenarios and their respective elasticity rules. Scenario (i): without elasticity; Scenario (ii): global elasticityl; (iii) functionally decomposed elasticity.

Scen.	VM description	Elasticity rules	
1	m4.large (for all stages)	—	
2	m4.large (single elasticity unit for all stages)	$\uparrow CPU > 75\%$	$\downarrow CPU < 35\%$
3	c4.large (CPU-optimized elasticity unit)	$\uparrow CPU > 75\%$	$\downarrow CPU < 35\%$
	r3.large (Memory-optimized elasticity unit)	$\uparrow RAM > 75\%$	$\downarrow RAM < 35\%$
	i2.xlarge (Disk-optimized elasticity-unit)	$\uparrow IOPS > 3k$	$\downarrow IOPS < 1k$
	m4.large (Network-opt. elasticity unit)	$\uparrow MBps > 40$	$\downarrow MBps < 10$

to 110 megabytes of data. We created three scenarios to evaluate different elasticity deployments, as can be seen in Table 1. This table also presents the elasticity rules for each elasticity unit. Three metrics are used to evaluate the Elastipipe prototype: time, resource and cost. The former refers to performance, presenting the total execution time for the whole dataset. Through periodical observations configured in the CloudWatch tool, we estimate the resource metric taking into account a cloud configuration with i VMs and $Obs(i)$ as the number of observations taken in such configuration. Finally, we are following the notion of cost used in parallel systems that is computed by multiplying time and consumed resources, not to be confused with monetary cost. Equations 1 and 2 compute the values of the resource and cost metrics. In Equation 1, v represents the maximum number of VMs that can be instantiated in the elasticity units.

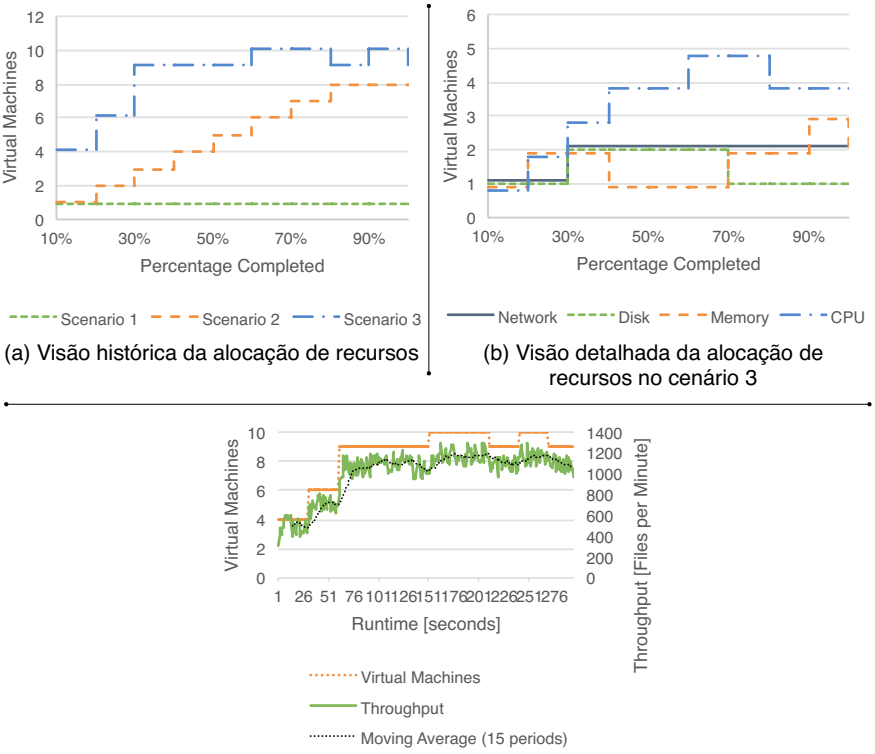
$$resource = \sum_{i=1}^{v} (i \times obs(i)) \tag{1}$$

$$cost = resource \times time \tag{2}$$

5 Results

Table 2 presents the results. The total execution time in this table assures the benefits of using cloud elasticity to handle the demands of pipeline-structured applications. The strategy of mapping services (stages of the pipeline) according to their main resource dependency (CPU-bound, network-bound, memory-bound or disk-bound), in addition to well configured elasticity rules, was responsible for the fastest performance rates and favor scenario 3. Considering the performance of scenario 1 as a normalizing factor, we obtained improvements of 62% and 72% when enabling elasticity. Comparing only scenarios 2 and 3, the later achieved 27% in performance gains while using 39% more VMs to accomplish it.

Figure 3 (a) depicts the resource allocation behavior along the execution of each scenario. We can observe a performance bottleneck on scenario ii: for most of the execution time it uses a saturated number of VMs, stabilizing only near the end, thus explaining the continuous “staircase” behavior on such scenario: a new VM is allocated but it becomes overloaded soon, so



(a) Visão histórica da alocação de recursos

(b) Visão detalhada da alocação de recursos no cenário 3

(c) Relação entre alocação de recursos e vazão no cenário 3

Fig. 3 Observing resource allocation and throughput along the execution.

Table 2 Results: scenario 1 — without elasticity; scenario 2 — elasticity through a single elasiticty unit; scenario 3 — elasticity through multiple elasticity units

Scen.	Mean Time per Task (s)	Throughput per Minute	Total Time	Mean Number of VMs	Resource Usage	Total Cost
1	5.52	108.70	17:34:121.0		1,054*1 = *1,054*	18,518
2	2.62	594.74	06:45:365.2		40*1 + 40*2 + 40*3 + 40*4 + 40*5 + 40*6 + 40*7 + 120*8 = *2,080*	14,056
3	2.13	1,007.45	04:55:128.5		29*4 + 29*6 + 145*9 + 87*10 = *2,465*	12,131

implying in a new elasticity request. Such a bottleneck is mitigated in scenario 3 through the distribution of elasticity units since the beginning, so that we can perceive earlier resource utilization equilibrium (see Figure 3 (b)).

Scenario i uses a single VM to accommodate the four sets of stages. In this context, we are using an AWS server from the *m*4 family which is classified as a general-purpose server that balances processing, disk, memory and network requirements. This static deployment is responsible for the lowest energy consumption in our tests. However, it presents the worst cost since its

execution time was much longer than those observed in the following scenarios. Scenario 2 also starts with a single VM, but the elasticity is in charge of increasing energy consumption while also enabling better execution time. In average, we had 26 simultaneous tasks being executed concurrently in the pipelines when enabling elasticity over a single elasticity unit that hosts all of the stages.

Scenario 3 works by grouping the stages in different elasticity units according to their resource requirements. Here, the fact that we are using more nodes in this scenario was responsible for achieving the highest value for the resource consumption metric. In particular, Baliga et al. [1] claim that the number of VMs in a node is not a major influential factor on the energy consumption; instead, the fact that the node is on or off is influential. Surpassing scenario 2, in scenario 3 we measured 36 parallel tasks in the pipelines, almost doubling the throughput of tasks per minute (from 594.74 in scenario 2 to 1,007.45 in scenario 3). Figure 3 (c) illustrates the throughput of tasks per minute and its relationship to resource allocation. Finally, we observed that functional decomposition of the pipeline stages to design the elasticity units is a major factor influencing performance while also facilitating infrastructure maintenance and diagnostics.

6 Conclusion

The concept of assembly lines, originated in the automotive industry and popularized in microprocessor designs, is present in complex workloads in a variety of knowledge areas. In this context, our article proposed a model named Elastipipe to address such applications while also presenting the concept of flexible superscalar pipeline to support on-the-fly resource reorganization. Such flexibility is achieved through the definition of elasticity units by the application's users or administrators, which can include one or more stages of a pipeline. Thus, each scaling in or out (consolidation or allocation of VMs) action considers this unit as resource grain to increase or reduce the resource pool that best fits the current needs of the application. The results obtained through our prototype proved the feasibility of Elastipipe's approach to cloud elasticity, where elastic scenarios present better cost rates when compared to a non-elastic one. In addition, the functional decomposition of stages in elasticity units according to their characteristic (CPU-, memory-, network- or disk-bound) was decisive to obtain the best results. This deployment scenario achieved gains of 27% when compared with a situation in which all stages are concentrated in a single elasticity unit.

Acknowledgement

The authors would like to thank to: CNPq, CAPES and FAPERGS.

References

1. J. Baliga, R.W.A. Ayre, K. Hinton, and R.S. Tucker. Green cloud computing: Balancing energy in processing, storage, and transport. *Proceedings of the IEEE*, 99(1):149–167, 2011.
2. S. Dustdar, Yike Guo, Rui Han, B. Satzger, and Hong-Linh Truong. Programming directives for elastic computing. *Internet Comp, IEEE*, 16(6):72–77, Nov 2012.
3. Soodeh Farokhi, Pooyan Jamshidi, Ivona Brandic, and Erik Elmroth. Self-adaptation challenges for cloud-based applications : A control theoretic perspective. In *10th Int. Workshop on Feedback Computing*. ACM, 2015.
4. Guilherme Galante, Luis De Bona, Antonio Mury, Bruno Schulze, and Rodrigo Righi. An analysis of public clouds elasticity in the execution of scientific applications: a survey. *J. Grid Comput.*, 14(2):193–216, 2016.
5. Nikolas Roman Herbst, Samuel Kounev, Andreas Weber, and Henning Groenda. Bungee: An elasticity benchmark for self-adaptive iaas cloud environments. In *Proc. of the 10th Int. Symp. on Software Engineering for Adaptive and Self-Managing Systems*, SEAMS '15, pages 46–56, Piscataway, NJ, USA, 2015. IEEE Press.
6. Shigeru Imai, Thomas Chestna, and Carlos A. Varela. Elastic Scalable Cloud Computing Using Application-Level Migration. In *2012 IEEE Fifth Int. Conf. on Utility and Cloud Computing*, pages 91–98, Honolulu, nov 2012. IEEE.
7. Janmartin Jahn, Santiago Pagani, Sebastian Kobbe, Jian-Jia Chen, and Jörg Henkel. Optimizations for configuring and mapping software pipelines in many core systems. In *Proc. Design Automation Conference*, DAC '13, pages 130:1–130:8, New York, NY, USA, 2013. ACM.
8. Joao Loff and Joao Garcia. Vadara: Predictive Elasticity for Cloud Applications. In *2014 Int. Conf. on Cloud Computing Technology and Science*, pages 541–546, Singapore, dec 2014. IEEE.
9. Paul Marshall, Kate Keahey, and Tim Freeman. Elastic Site: Using Clouds to Elastically Extend Site Resources. In *2010 10th IEEE/ACM Int. Conf. on Cluster, Cloud and Grid Computing*, pages 43–52, Melbourne, 2010. IEEE.
10. Patrick Martin, Andrew Brown, Wendy Powley, and Jose Luis Vazquez-Poletti. Autonomic management of elastic services in the cloud. In *2011 IEEE Symp. on Computers and Communications (ISCC)*, pages 135–140, Kerkyra, jun 2011. IEEE.
11. Dinesh Rajan, Anthony Canino, Jesus A. Izaguirre, and Douglas Thain. Converting a high performance application to an elastic cloud application. In *Proc. of Int. Conf. on Cloud Computing Technology and Science*, CLOUDCOM '11, pages 383–390, Washington, DC, USA, 2011. IEEE Computer Society.
12. A. Raveendran, T. Bicer, and G. Agrawal. A framework for elastic execution of existing mpi programs. In *Parallel and Distributed Processing Workshops and Phd Forum (IPDPSW), 2011 IEEE International Symposium on*, pages 940–947, May 2011.
13. R. Righi, V. Rodrigues, C. Andre daCosta, G. Galante, L. Bona, and T. Ferreto. Autoelastic: Automatic resource elasticity for high performance applications in the cloud. *Cloud Computing, IEEE Transactions on*, PP(99):1–1, 2015.
14. Nam-Luc Tran, Sabri Skhiri, and Esteban Zimányi. EQS: An Elastic and Scalable Message Queue for the Cloud. In *2011 IEEE Third Int. Conf. on Cloud Computing Technology and Science*, pages 391–398, Athens, nov 2011. IEEE.
15. Xinwen Zhang, Sangoh Jeong, Anugeetha Kunjithapatham, and Simon Gibbs. Towards an Elastic Application Model for Augmenting Computing Capabilities of Mobile Platforms. In *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecomm. Eng.*, volume 48, pages 161–174. Springer, 2010.

Semantic Summarization of News from Heterogeneous Sources

Flora Amato, Antonio d'Acierno, Francesco Colace, Vinenzo Moscato, Antonio Penta, Antonio Picariello

Abstract Summarization techniques are becoming an essential part of everyday life, basically because summaries allow users to spend less time making effective access to the desired information. In this paper, we present a general framework for retrieving relevant information from news articles and a novel summarization algorithm based on a deep semantic analysis of texts. In particular, a set of triples (subject, predicate, object) is extracted from each document and it is then used to build a summary through an unsupervised clustering algorithm exploiting the notion of semantic similarity. Finally, we leverage the centroids of clusters to determine the most significant summary sentences using some heuristics. Several experiments are carried out using the standard DUC methodology and ROUGE software and show how the proposed method outperforms several summarizer systems in terms of recall and readability.

1 Introduction

Seeking bits of information from a large amount of data still remains a difficult and time consuming task for a wide range of people such as stu-

Flora Amato, Vincenzo Moscato and Antonio Picariello
DIETI - University of Naples, e-mail: flora.amato,vmoscato,picus@unina.it

Antonio d'Acierno
ISA - Consiglio Nazionale delle Ricerche (CNR) e-mail: antonio.dacierno@isa.cnr.it

Antonio Penta
University of Turin, Department of Computer Science, Via Pessinetto, 12, 10149, Torino (Italy) e-mail: penta@di.unito.it

Francesco Colace
Dipartimento di Ingegneria dell'Informazione, Ingegneria Elettrica e Matematica Applicata, University of Salerno, Fisciano (Italy) e-mail: fcolace@unisa.it

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_29

dents, news reporters, and many other types of professionals. The exponential growth of the Web has made the search and track of information apparently easier and faster, but the huge information overload requires algorithms and tools for a fast and easy access to the specific *desired* information, discriminating between "useful" and "useless" information, especially in the era of *Big Data*.

We can consider as an example a typical workflow related to a news reporter that has just been informed of a plane crash in Milan and that would quickly like to gather more details about this event[1, 2, 3]. We suppose that the following textual information [1] on the accident, extracted from some web sites already reporting the news, are publicly available:

"*A Rockwell Commander 112 airplane crashed into the upper floors of the Pirelli Tower in Milan, Italy. Police and ambulances are at the scene. The president, Marcello Pera, just moments ago was informed about the incident in Milan, he said at his afternoon press briefing*". "*It was the second time since the Sept 11 terror attacks on New York and Washington that a plane has struck a high-rise building. Many people were on the streets as they left work for the evening at the time of the crash. Ambulances streamed into the area and pedestrians peered upward at the sky. The clock fell to the floor. The interior minister had informed the senate president, Marcello Pera, that the crash didn't appear to be a terror attack*".

Our basic idea consists in finding, from each sentence, a sequence of relevant information that are then clustered into subsets with a *similar* information content; thus we would like to discard repeated sentences, and to consider only the relevant ones for each cluster. More in details, we can obtain that by reducing the information, that is contained into a sentence, to a regular and simple form: for example, using NLP algorithms [4, 5, 6] we retrieve a list of triples formed by $\langle subject, verb, object \rangle$, where the verb is reported in infinitive form while subject and object are nouns. For the previous air crash example, the extracted triples are[2] : ⟨airplane, crash, tower⟩, ⟨police, be, scene⟩, ⟨ambulance, be, scene⟩, ⟨president, inform, incident⟩*, ⟨terror, attack, New York⟩, ⟨terror, attack, Washington⟩, ⟨plane, strike, building⟩, ⟨people, be, street⟩, ⟨ambulance, stream, area⟩, ⟨pedestrian, peer, sky⟩, ⟨clock, fall, floor⟩, ⟨minister, inform, president⟩, ⟨crash, appear, terror attack⟩*.

Successively, a clustering algorithm creates the clusters reported in Table 1. The clusters are obtained by computing the *semantic similarity* between each couple of triples[7, 8]. Finally, we assume that it is possible to obtain a useful summary by just considering the sequence of centroids and eventually re-loading the associated original sentence, as in the following: "*A Rockwell Commander 112 airplane crashed into the upper floors of the Pirelli Tower in Milan, Italy. Police and ambulances are at the scene. The interior minister had informed the senate president, Marcello Pera, that the crash didn't appear to be a terror attack.*"

[1] The analyzed texts come from news articles, thus they are generally quite simple as concerning grammar and syntax.

[2] Particular cases in which verb is a modal verb or is in a passive or negation form - see triples marked with '*' - have to be opportunely managed.

Cluster 1	⟨ambulance, be, scene⟩ ⟨police, be, scene⟩ ⟨ambulance, stream, area⟩ ⟨people, be, street⟩ ⟨pedestrian, peer, sky⟩
Cluster 2	⟨minister, inform, president⟩ ⟨president, inform, incident⟩
Cluster 3	⟨airplane,crash,tower⟩ ⟨plane, strike, building⟩ ⟨clock, fall, floor⟩ ⟨terror, attack, Washington⟩ ⟨terror, attack, New York⟩ ⟨crash, appear, terror attack⟩

Table 1: The clustering results on a set of triples (centroids are underlined).

Summarizing, we take advantages of some NLP facilities in order to propose a novel approach for text summarization based on the extraction of semantic descriptors of documents, avoiding approaches that are time-consuming and require domain dependent settings[9, 10, 11, 12].

In particular, the semantic content of documents is captured by a set of triples and we then propose a methodology to semanitcally cluster "similar information", thus a summary may be decribed as the sequence of sentences that are associated to the most representative clusters. The idea of using triples as semantic units for representing content of web documents is well studied in the Resource Description Framework (RDF) [3] in the Semantic Web Community.

2 A model for automatic multi-document summarization

2.1 Basic elements of the model

The summarization problem may be stated as follows: given a set of source documents, let us produce an accurate and all-sided summary that is able to reflect the *main concepts* expressed by the original documents, matching some *length restrictions* and without introducing *additional* and *redundant* information.

Our idea is inspired by the text summarization models based on *Maximum Coverage Problem* ([13, 14]), but differently from them we design a methodology that combines both the syntactic and the semantic structure of a text. In particular, in the proposed model, the documents are segmented into several linguistic units (named as *summarizable sentences*) in a preprocessing

[3] http://www.w3.org/TR/WD-rdf-syntax-971002/

stage, and each linguistic unit is then characterized by a set of conceptual units (named as *semantic atoms*) containing the meaning of a sentence. Our main goal is to *cover* as many conceptual units as possible using only a small number of sentences.

Let us give some preliminary definitions about our idea of "summarizable sentence" and "semantic atoms", and for the sake of clarity, we introduce the example sentences in Table 2 to better explain the introduced definitions.

Document	Sentence	Text of Sentence
1	1.a	People and ambulances were at the scene.
	1.b	Many cars were on the street.
	1.c	A person died in the building.
2	2.a	Marcello Pera denied the terror attack
	2.b	A man was killed in the crash.
	2.c	The president declared an emergency.

Table 2: Example Sentences

Definition 1 (Summarizable Sentence and Semantic Atoms). A *Summarizable Sentence* σ defined over a document D is a couple:

$$\sigma = \langle s, \{t_1, t_2, \ldots, t_m\} \rangle \tag{1}$$

s being a sentence belonging to D and $\{t_1, t_2, \ldots, t_m\}$ being a set of atomic or structured information that expresses in some way the semantic content related to s.

In particular, t_i is defined as a set of couples $\langle A_i, \mathcal{V}_i \rangle$, where A_i is any relevant attribute on which a generic classifier is trained and $\mathcal{V}_i$ is a string or a set of strings in the sentence that is classified as a value for A_i[15, 16]. Let us call t_i as *Semantic Atom.*

Example 1 *In Table 3 there are depicted some possible summarizable sentences that can be extracted from the sentences introduced in Table 2. We introduces attributes like "Sub" (Subject), "Verb", "Obj" (Object) and "Person" to describe the syntactic structure or the named entities of a sentence.*

Given a set of documents, a *summary* is a set of summarizable sentences. The set has to satisfy some length restriction conditions and include the most representative content of all input data. For now, let us assume the existence of a similarity function $sim(t_i, t_j) \in [0, 1]$ able to compute the semantic similarity between two semantic atoms and another function able to *score* the semantic atoms based on their importance. We will give more details on both these aspects in the following subsections.

Document 1	
Summarizable Sentences	Semantic Atoms
$\sigma_{1.1}=\langle 1.a, t_1^{1.1}, t_2^{1.1}\rangle$	$t_1^{1.1}=\{\langle Sub, people\rangle, \langle Verb, were\rangle, \langle Obj, scene\rangle\}$, $t_2^{1.1}=\{\langle Sub, ambulances\rangle, \langle Verb, were\rangle, \langle Obj, scene\rangle\}$,
$\sigma_{1.2}=\langle 1.b, t_1^{1.2}\rangle$	$t_1^{1.2}=\{\langle Sub, cars\rangle, \langle Verb, were\rangle, \langle Obj, street\rangle\}$.
$\sigma_{1.3}=\langle 1.c, t_1^{1.3}\rangle$	$t_1^{1.3}=\{\langle Sub, person\rangle, \langle Verb, died\rangle, \langle Obj, building\rangle\}$.
Document 2	
Summarizable Sentences	Semantic Atoms
$\sigma_{2.1}=\langle 2.a, t_1^{2.1}, t_2^{2.1}, t_3^{2.1}\rangle$	$t_1^{2.1}=\{\langle Sub, Marcello\ Pera\rangle, \langle Verb, denied\rangle, \langle Obj, attack\rangle\}$ $t_2^{2.1}=\{\langle Person, Marcello\ Pera\rangle\}$,
$\sigma_{2.2}=\langle 2.b, t_1^{2.2}\rangle$	$t_1^{2.2}=\{\langle Sub, man\rangle, \langle Verb, was\ killed\rangle, \langle Obj, crash\rangle\}$.
$\sigma_{2.3}=\langle 2.c, t_1^{2.3}\rangle$	$t_1^{2.3}=\{\langle Sub, president\rangle, \langle Verb, declared\rangle, \langle Obj, emergency\rangle\}$

Table 3: Example of Summarizable Sentences extracted from the documents in Table 2.

Now we are in position to introduce the concept of "Summarization Algorithm" as follows.

Definition 2 (Summarization Algorithm). Let $\mathcal{D}$ be a set of documents, a *Summarization Algorithm* is formed by a sequence of two functions ϕ and χ. The semantic partitioning function (ϕ) partitions $\mathcal{D}$ in K sets $\mathcal{P}_1, \ldots, \mathcal{P}_K$ of summarizable sentences having similar semantics in terms of semantic atoms and returns for each set the related *information score* by opportunely combining the score of each semantic atom:

$$\phi : \mathcal{D} \rightarrow \mathcal{S}^* = \{\langle \mathcal{P}_1, \hat{w}_1\rangle, \ldots, \langle \mathcal{P}_K, \hat{w}_K\rangle\} \tag{2}$$

s. t. $\mathcal{P}_i \cap \mathcal{P}_j = \emptyset, \forall i \neq j$.

The *Sequential Sentence Selection* function (χ):

$$\chi : \mathcal{S}^* \rightarrow \mathcal{S} \tag{3}$$

selects a set of the sentences $\mathcal{S}$ from original documents containing the semantics of most important clustered information sets in such a way that:

1. $|\mathcal{S}| \leq L$,
2. $\forall \mathcal{P}_k, \hat{w}_k \geq \iota, \nexists t_j, t_j \in_\sigma \mathcal{S}^* : sim(t_i, t_j) \geq \gamma, t_i \in_\sigma \mathcal{S}$.

ι and γ being two apposite thresholds. With abuse of notation, we use the symbol $\in_\sigma$ to indicate that a semantic atom comes from a sentence belonging to the set of summarizable sentences $\mathcal{S}$.

Note that we can select any unsupervised clustering algorithm that is able to partition the documents space into several clusters using the semantic similarity among semantic atoms. Once obtained a partition of the space in terms of clusters of semantic atoms, we select a number of sentences, trying to: (i) maximize the semantic coverage - the most representative sentences of each cluster should be considered starting from the most important clusters, i.e. those having the highest average information score; (ii) minimize the redundancy by selecting one sentence for each cluster that is most representative in terms of semantic content and not considering similar sentences.

Now, we are going to explain the following points of our model are: (i) how to represent and extract semantic atoms of a document, (ii) how to evaluate the similarity between two semantic atoms, (iii) how to calculate a score for each semantic atom, and finally, (iv) how to define suitable semantic partitioning and sentence selection functions.

2.2 Extracting semantic atoms from a text

We adopted the principles behind the RDF framework used in the Semantic Web community to semantically describe web resources. The idea is based on representing data in terms of a triple $\langle subject, verb, object \rangle$. In the Semantic Web community subjects and objects are web resources while verbs are predicates/relations defined in schemata or ontologies, in our case instead we attach to the elements of triples the tokens extracted by processing documents using NLP techniques. Thus, the semantic content of a document can be modeled by a set of structured information $\mathcal{T}=\{t_1, \ldots, t_n\}$, where each element t_i is a semantic atom described by the following couples $t_i=\{\langle sub, val \rangle, \langle verb, val \rangle, \langle obj, val \rangle\rangle\}$, and we can call t_i as *summarization triple*. Since now, we use in the rest of the paper the name *triple* and *summarization triple* with the same meaning.

The triples are extracted from each sentence in the documents by applying a set of rules on the *parse tree* structure computed on each sentence. In our rules, subjects and objects are nouns or chunks while the verbs are reported in the infinitive form.[4] It is worth to be noted that in the case of long sentences more triples may be associated to a sentence. The rules are obtained by defining a set of patterns for subject, verb and object which includes not only part of speech features but also parse tree structures. In particular, we start from the patterns described in [17] in order to include not only relations

[4] We do not consider any other grammatical units for a sentence such as adjective, preposition and so on. This because we are not interested in detecting the sentiment or opinion in the text or to provide just triples themselves as final summarization results to the user, but we exploit them like" pointers" to original sentences that are the real components of our summaries. Thus, we want to ensure that the quality of these pointers is enough to get together similar sentences.

but also subjects and objects and we add to the pattern expressions features related to the sentence linguistic structure (parse tree).

2.3 Semantic similarity function

In our model, we decided to compare two semantic atoms based on the similarity measure obtained by the comparison of the elements hosted by a summarization triple, namely subject, predicate, and object. In particular, let us consider two sentences and assume to extract from them two triples t_1 and t_2; we define as similarity between two t_1 and t_2 the function:

$$sim(t_1, t_2) = F_{agr}(F^{sim}(sub_1, sub_2), F^{sim}(pred_1, pred_2), F^{sim}(obj_1, obj_2)) \quad (4)$$

The function F^{sim} is used to obtain the similarity among values of the semantic atoms, while F_{agr} is an aggregation function. The following functions are some examples of the easiest combination strategies:

$$F_{agr}(t_1, t_2) = \sum_{i=1}^{3} \alpha_i \cdot F^{sim}(t_1[i], t_2[i]); \quad (5)$$

$$F_{agr}(t_1, t_2) = \max_{i \in \{1,2,3\}} \left[\alpha_i \cdot F^{sim}(t_1[i], t_2[i])\right] \quad (6)$$

where the constraint $\alpha_1 + \alpha_2 + \alpha_3 = 1$ is used to obtain a convex combination of the results and we consider the same function for each element of a triple.

If the F^{sim} takes into account the information stored in a knowledge base, the function is computed based on the "*semantic*" aspects of its input, otherwise we can apply any similarity among words like the one based on the well-known edit distance.

In particular, we use the *Wu & Palmer* similarity ([18]) for computing the similarity among elements of our triples. This similarity is based on the *Wordnet Knowledge Base*, that lets us compare triples based on their semantic content.

3 The proposed framework

In Figure 1, we show at a glance the summarization process that consists of the following steps: (i) *Web Search* - this activity has the task of retrieving a set of HTML documents that satisfy some search criteria using a *Search Engine* external component; (ii) *Text Extraction* - the sentences are extracted from the several web sources by parsing the related HTML pages and analyzing the HTML tags; (iii) *NLP and Triples Extraction* - NLP pro-

cessing techniques are performed on the input sentences, in particular *Named Entities Recognition* (NER), *Part Of Speech* (POS) tagging, Parse Tree Generation, Anaphora and Co-Reference Resolutions, successively in this stage, semantic triples are detected for each sentence, analyzing the related Parse Tree and using appropriate heuristics; (iv) *Similarity Matrix Builder* - a matrix containing the similarity values for each couple of triples is computed; (v) *Clustering* - a proper clustering algorithm is applied on the input matrix; (vi) *Sentence Selection* - this activity performs a sentence selection to generate the summary using our proposed algorithm; (vii) *Summary Building* - this activity performs a sentence ordering to generate the final summary[19].

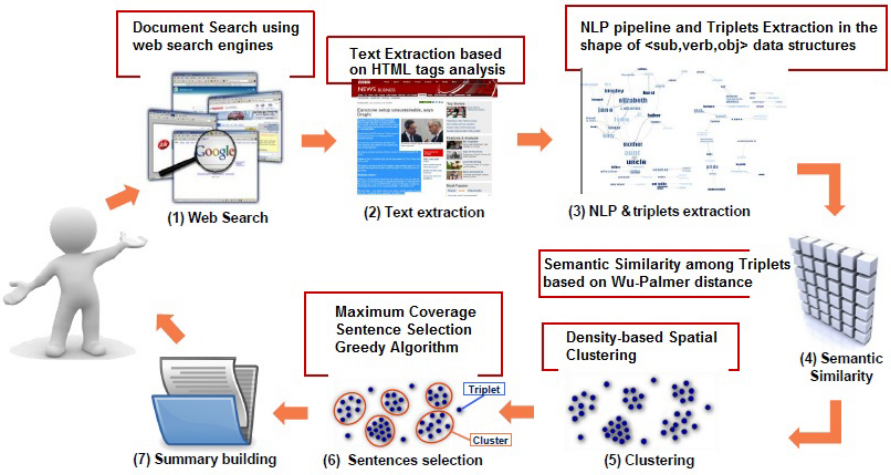


Fig. 1: The summarization process

The system is named *iWIN* (information on the Web In a Nutshell) and provides a graphical user interface that allows the user to configure some system settings, thus allowing to tune the system and evaluate its performances under several conditions and situations.

4 Conclusions

In this paper, we described a novel approach for summarizing web documents based on semantic extraction and description of documents.

We first proposed a model of summarization based on the semantic content of a document, captured and modelled by a set of triples, i.e. a subject, verb, object.

We then proposed a novel methodology based on cluster analysis of triples, thus obtaining a summary as the sequence of sentences that are associated to the most representative clusters' triples: in particular, the centroids of the clusters are used to detect the main representative topics, then they are properly combined for producing non-repetitive and brief summaries. To the best of our knowledge, this is the first work that uses a semantic-driven approach between triples for text summarization applications[20, 21, 22, 23].

Based on this approach, we implemented a system called iWin, that provides all the functionalities of a multi-document summarization tool. We tested iWin using some well-known data sets, showing good and sometimes excellent performances with respect to classical evaluation measures in the summarization literature. In addition, we made a performances comparison with open sources and commercial summarizer systems, obtaining promising results with respect to other approaches both for query-based and for generic summaries.

Future work will be devoted to improve the current research into main directions: i) extend the proposed methodology to the query-based approach; ii) consider multimodal summaries able to combine different data coming from unstructured (image, video, audio data) or structured (linked data) repositories; iii) improve efficiency of our approach (by caching similarity values between terms that have to be computed more times, using concurrent computation threads for subjects, verbs and objects and defining a proper indexing to access to the distances matrix); iv) implement the algorithms in a parallel computing environment; v) test several and more advanced density-based clustering approaches.

References

1. Leonard Barolli and Fatos Xhafa. Jxta-overlay: A p2p platform for distributed, collaborative, and ubiquitous computing. *Industrial Electronics, IEEE Transactions on*, 58(6):2163–2172, 2011.
2. Fatos Xhafa, Raul Fernandez, Thanasis Daradoumis, Leonard Barolli, and Santi Caballé. Improvement of jxta protocols for supporting reliable distributed applications in p2p systems. In *Network-Based Information Systems*, pages 345–354. Springer, 2007.
3. Leonard Barolli, Fatos Xhafa, Arjan Durresi, and Giuseppe De Marco. M3ps: a jxta-based multi-platform p2p system and its web application tools. *International Journal of Web Information Systems*, 2(3/4):187–196, 2007.
4. Mario Sicuranza, Angelo Esposito, and Mario Ciampi. An access control model to minimize the data exchange in the information retrieval. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–12, 2015.
5. Aniello Minutolo, Massimo Esposito, and Giuseppe De Pietro. A fuzzy framework for encoding uncertainty in clinical decision-making. *Knowledge-Based Systems*, 98:95–116, 2016.
6. Aniello Minutolo, Massimo Esposito, and Giuseppe De Pietro. Design and validation of a light-weight reasoning system to support remote health monitoring applications. *Engineering Applications of Artificial Intelligence*, 41:232–248, 2015.

7. F. Amato, A.R. Fasolino, A. Mazzeo, V. Moscato, A. Picariello, S. Romano, and P. Tramontana. Ensuring semantic interoperability for e-health applications. pages 315–320, 2011.
8. F. Amato, A. Mazzeo, V. Moscato, and A. Picariello. A framework for semantic interoperability over the cloud. pages 1259–1264, 2013.
9. Tim French, Nik Bessis, Fatos Xhafa, and Carsten Maple. Towards a corporate governance trust agent scoring model for collaborative virtual organisations. *International Journal of Grid and Utility Computing*, 2(2):98–108, 2011.
10. Valentin Cristea, F. Pop, C. Stratan, A. Costan, C. Leordeanu, and E. Tirsa. A dependability layer for large-scale distributed systems. *International Journal of Grid and Utility Computing*, 2(2):109–118, 2011.
11. Soichi Sawamura, Admir Barolli, Ailixier Aikebaier, Makoto Takizawa, and Tomoya Enokido. Design and evaluation of algorithms for obtaining objective trustworthiness on acquaintances in p2p overlay networks. *International Journal of Grid and Utility Computing*, 2(3):196–203, 2011.
12. Evjola Spaho, Gjergji Mino, Leonard Barolli, and Fatos Xhafa. Goodput and pdr analysis of aodv, olsr and dymo protocols for vehicular networks using cavenet. *International Journal of Grid and Utility Computing*, 2(2):130–138, 2011.
13. H. Takamura and M. Okumura. Text summarization model based on maximum coverage problem and its variant. In *Proceedings of the 12th Conference of the European Chapter of the AC*, pages 781–789, 2009.
14. Dan Gillick and Benoit Favre. A scalable global model for summarization. In *Proceedings of the Workshop on Integer Linear Programming for Natural Language Processing*, ILP '09, pages 10–18. Association for Computational Linguistics, 2009.
15. Salvatore Cuomo, Pasquale De Michele, Ardelio Galletti, and Giovanni Ponti. *Intelligent Interactive Multimedia Systems and Services 2016*, volume 55 of *Smart Innovation, Systems and Technologies*, chapter Influence of Some Parameters on Visiting Style Classification in a Cultural Heritage Case Study, pages 567–576. Springer International Publishing, 2016.
16. Salvatore Cuomo, Pasquale De Michele, Ardelio Galletti, and Giovanni Ponti. *Data Management Technologies and Applications: 4th International Conference, DATA 2015, Colmar, France, July 20-22, 2015, Revised Selected Papers*, volume 584 of *Communications in Computer and Information Science*, chapter Classify Visitor Behaviours in a Cultural Heritage Exhibition, pages 17–28. Springer International Publishing, 2016.
17. Oren Etzioni, Anthony Fader, Janara Christensen, Stephen Soderland, and Mausam Mausam. Open information extraction: The second generation. In *IJCAI*, volume 11, pages 3–10, 2011.
18. Zhibiao Wu and Martha Palmer. Verb semantics and lexical selection. In *32nd. Annual Meeting of the Association for Computational Linguistics*, pages 133 –138, 1994.
19. A. D'Acierno, V. Moscato, F. Persia, A. Picariello, and A. Penta. iwin: A summarizer system based on a semantic analysis of web documents. pages 162–169, 2012.
20. G. Sannino, I. De Falco, and G. De Pietro. An automatic rules extraction approach to support osa events detection in an mhealth system. *IEEE Journal of Biomedical and Health Informatics*, 18(5):1518–1524, 2014.
21. Angelo Chianese, Fiammetta Marulli, Francesco Piccialli, Paolo Benedusi, and Jai E Jung. An associative engines based approach supporting collaborative analytics in the internet of cultural things. *Future Generation Computer Systems*, 2016.
22. A. Chianese, F. Picciali, and I. Valente. Smart environments and cultural heritage: a novel approach to create intelligent cultural spaces. *Journal of Location Based Services*, 9:209–234, 2015.
23. Giuseppe Caggianese, Luigi Gallo, and Giuseppe De Pietro. Design and preliminary evaluation of a touchless interface for manipulating virtual heritage artefacts. In *Signal-Image Technology and Internet-Based Systems (SITIS), 2014 Tenth International Conference on*, pages 493–500. IEEE, 2014.

Self Planning in Critical Multi-Agent Systems

Flora Amato, Nicola Mazzocca, Francesco Moscato

Abstract Novel architectures like Cloud and Internet of Things (IoT) make available several resources like computing nodes, environmental sensors etc. that enable the introduction of more and more intelligent systems able to face complex situations. In particular management of critical and dangerous situations may take advantage of those systems whose complexity is growing up faster and faster. In this scenario, it is difficult to orchestrate different autonomous systems in order to face with new, previously unmanaged emergencies. In this work we present a modeling methodology and a planning techniques based on a multi-agent model. Agents describe capabilities of each available IoT element in an area where a critical situation has occurred; the planning methodology exploit both classical and a new counter-example based approaches to build a coordination plan of resources in order to achieve given goals like traffic management or people flight during a terrorist attack.

1 Introduction

Cloud Architecture and Internet of Things (IoT) are nowadays a *de facto* standard for providing *any* kind of service on Internet and to collect and share data from many kind of sensors and devices. Agent-based models are

Flora Amato
DIETI - University of Naples "Federico II", via Claudio 21, Naples, Italy, e-mail: flora.amato@unina.it

Francesco Moscato
Second University of Naples, Dipartimento di Scienze Politiche, Caserta, Italy, e-mail: francesco.moscato@unina2.it

Nicola Mazzocca
DIETI - University of Naples "Federico II", via Claudio 21, Naples, Italy, e-mail: nicola.mazzocca@unina.it

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_30

the most used in description of devices and sensors in IoT environments [1, 2]. Modeling elements by using agents allows for the description of their autonomous behaviours and interactions. Anyway, interactions have usually limited goals: they are pre-defined tasks and critical situations have to be managed manually by human operators. The problem is that in these cases, management is intrinsically and hard real-time process [3] and having a *true* autonomous system able to reconfigure in case of necessity is appealing. This is possible only if agents themselves are able to build new communication workflows in order to face new situations.

This is a planning problem where agents defines their own actions in a high level plan involving all available agents in the environment. State Space Informed Search is one of most applied method for planning actions in intelligent systems. Many works have been proposed for the definition of new strategies and algorithms able to improve results of classical planning approaches anyway, planning literature usually neglects a common problem in planning: the control of the existence of a plan. Approaches based on State Space Search are not efficient in detecting if a Goal is not reachable in any way.

This problem is very important in critical systems where, the impossibility of reaching a safe state in a critical situations should be detected as fast as possible, in order to allow the definition of new plans to face the problem. In particular, this is important when dealing with real-time systems[4], resource sharing and interaction may lead to problem in reaching the goals. In these cases, a way to determine in a short time if a goal is not reachable is really appealing since agents may choose to change goals without wasting time.

In order to change goals agents have to realize that any plan achieving current goal does not exist. The work in [5] for example, proposes an efficient way to determine if for a given goal, a plan of length n exists, examining the satisfiability of properly defined formulae. In our work we propose a Framework composed by a Modeling Tool, that provides the means to model Multi Agents Systems, and a Planner based on Counter Example. In the Counter Example approach, the planning problem is defined trying to answer the following question: "is it true that from the initial state, only states where the goal conditions are not satisfied are reachable?". If the answer of the previous question is "yes" than no plan for the current goal exist. Otherwise it is possible to find a "counter example" showing that at least one path exists from the initial state to a state where the goal conditions are satisfied: this counter-example as we will show, represent a plan for the current goal.

The approach we propose is particularly efficient in finding the non existence of a plan for a given goal. We will show that our approach has good performance in computational expensive planning problems where a single agent is involved[6, 7, 8, 9]. The agent modeling approach is based on the Beliefs, Desires and Intentions(BDI) model and the action modeling is based on First Order Logic (FOL) STRIPS [10] formalism: for each action we define a)a list of preconditions; b) a list of postconditions (in spite of add and

delete lists of STRIPS) c) the domain definition for each of the variables in the previous lists. If the number of beliefs is large, generation of the whole automaton in a single run could be infeasible; for this reason, automata generation follows a Self-Learning approach: each time an agent finds a new state, that state is added to the automaton and expanded considering the possible agent actions.In this way the automata generation is made step by step while the agent is acting instead of a unique step at the beginning.

2 Agent Based Planning Problem

In this section we provide a simplified version of the model we use to define our planning problem. The model is based on a Multi-Agent system representation of systems. We consider a variant of Beliefs, Desires, Intentions (BDI) logics[11] for our Multi-Agent System model that is a quadruple:

$$(Agents, World, \mathcal{TS}, \mathcal{F})$$

Where *Agents* is the set of all agents in the system; $World$ represent the environment where agents executes; $\mathcal{TS}$ is a Transition System that resumes possible state transitions of agents in the environment and, finally, $\mathcal{F}$ is a a set of formulas expressed in first order logics that characterize each state in $World$.

We use a triple $\langle n, d, v\rangle$ to define variables evaluations that describes states in the World, where n is the name of a variable, d represents its domain and v a value assigned to the variable. A State $s \in World$ is a set of variable evaluation.

In addition, we call *State Conditions* of a state, the set of all formulas in $\mathcal{F}$ that holds in a state s:

$$StateCondition(s) = \{\phi \in \mathcal{F}, s \in World : s \models \phi\}$$

In addition, ϕ cannot be a subformula of other formulas holding in s. In this work we consider $\mathcal{TS}$s and States with only one State Condition per state. If $s \models \psi$; $s \models \phi$ and $s \models \psi \wedge \phi$; then we consider only the last formula as State Condition in s. An Agent is in turn a triple:

$$(Actions, Beliefs, Goals)$$

Actions is a set of possible actions an agent is able to perform. Actions modify the environment changing $World$ representation. They can also require the intervention of other agents in order to achieve common goals and in general, they include *communication* and *execution* actions. In addition, an action can be *reactive* if its execution depends on external events or messages; or *proactive* if its execution is decided directly by the agent. We call

Proactive Agent an agent with at least one proactive action; an agent with no proactive actions and with at least one reactive action is a *Resource Agent*; an agent with no reactive or proactive actions is classified simply as a *Resource*. $Beliefs$ include the knowledges the agents have about: the $World$; the Agent itself; other Agents. Notice that an agent may have a belief about the $World$ which in turn is *not* true in the environment: in general, beliefs of each agents may not be exact. $Goals$ is a set of states in $\mathcal{TS}$ that represent goals an agent want to reach. Since $\mathcal{TS}$ is not available when agents are defined, abusing notation we identify Goals with formulas in $\mathcal{F}$ that are satisfied in goal states. We call these formulas: *Goal Conditions*. Notice that a goal condition is a State Condition for a goal state. Beliefs are manages as Worlds variables and states (they are practically local World representation in each agent). Agents define the $\mathcal{TS}$ Transition System on $World$ states by means of $Actions$. An action $\alpha \in Actions$ is a triple:

$$(name, Precondition, Effects)$$

where *name* is trivially the name of the action;*Effects* is a set of formulas that hold in the new state; *Precondition* is a formula that *evaluates* true in a state s in order to *apply*(execute) the action and to produce a state transition:

$$s \overrightarrow{\alpha} s'$$

produces a transition from the state s to s'. If Precondition of α evaluates true in s, s' will be the same of s, except for variables involved in Effects evaluation. The values of these variables have to change in order to satisfy *all* effects in s':

$$\forall \phi \in Effects \; s' \models \phi$$

In addition we must consider that an agent executing an action can access only to its local representation of $World$, i.e. to its beliefs. Hence, if agent's beliefs and World State are not synchronized (i.e., if the agents has a wrong belief about the world), it is possible that Precondition is evaluated true on beliefs, but *not* on $World$ state. In order to apply an action, we must execute the following two steps: (1) an agent tries to apply Effects in a state s producing a transition from s to s' if precondition is evaluated true on its *beliefs*; (2) if Precondition evaluates true in *World too*, then $s \overrightarrow{\alpha} s'$ both in agent's Beliefs and in $World$ too. $\mathcal{TS}$ is then the Transition System defined by the application of all actions in any state of $World$, performed by *all* Agents in the model. The execution of an action to build $\mathcal{TS}$ must follow the two steps previously defined. State transitions apply both to agents $Belief$ and to $World$. Anyway Precondition control is enacted on Beliefs first. If evaluation fails on Beliefs, the action is not applied even if Precondition would evaluate true on $World$. In this model, a **Plan** to reach a Goal $\mathcal{G}$ with given Goal Condition is a *path* from a starting state to a state where the Goal Condition holds. Notice that in a Multi Agent System, actions in a transition System

can be executed by different agents, even concurrently. We consider here a path as a linear scheduling of concurrent applications of actions. A Planning problem hence, is the problem of finding such a path or to state that the requested goal is unreachable in the current environment.

3 Methodology and Framework

The methodology we are going to illustrate is based on the model presented in the previous section and it comprises a set of phases fully summarized in Fig.1. In the *Modeling* phase, we provide a representation of systems that is compliant with the model in section 2.

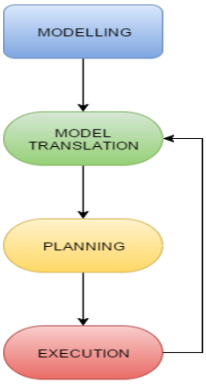


Fig. 1 Methodology phases

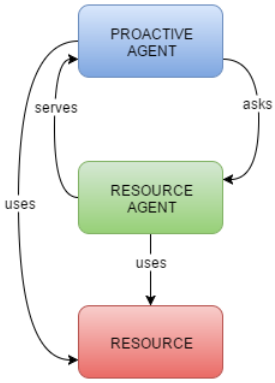


Fig. 2 Agents interaction model

The environment ($World$) is modeled in terms of states defined by means of $\langle n, d, v \rangle$ triples.

In addition, Agents are modeled in terms of Beliefs and Actions. Actions requires the definition of $(name, Precondition, Effects)$ triples, but we introduce in this phase an extension of the model since we need some additional information during planning. Hence we extend Actions definition to the quadruple:

$$(name, Precondition, Effects, VarInfo)$$

where $VarInfo$ set is necessary to specify variable domains, quantification(universal or existential), and eventually other properties.

During *Model Translation* and *Planning Phase*, a *Planner Engine* processes the model of the system. This produces a scheduling of actions representing

the plan to achieve the requested goal. In the last phase, Agents perform plans action by action at run time. The *Execution Environment* monitors correct behaviors of agents: if for any reason the current goal is no longer reachable, or if scheduled actions cannot be executed, (because preconditions no longer meet or because Beliefs were different from $World$ real configuration), the Execution Environment enacts a *Replanning* action that tries to retrieve new plans for the same goals (if they exist). Interactions among agents follow a *hierarchical orchestration* model as summarized in Fig.2.

In this work we consider communication among agents only at different layers. The interaction model from Proactive Agents and Resource Agents to Resources is very simple: They can use a resource if it is *free* (i.e. if it is not used by other high level agents). Interactions between Proactive and a Resource Agents is more complex: a Proactive Agent can *ask for a service* to a Resource Agent (and this latter *serves* this kind of request). When a proactive agents asks for a service, it assigns a schedule of actions to a resource agent, providing a scheduling for a plan it want to execute. Hence, in this work, a planning actions execute with the following steps: (1) A Proactive Agents finds a plan in an environment with one ore more Resource Agents and resources. Actions in the plan are related to Resource Agents actions which in turn may use simple Resources. (2) The proactive Agent with the new plan, assigns sub-goals to Resource Agents that, in turn, executes planning actions at a different layer of abstraction to reach sub-goals. (3) Resource Agents executes their sub-plans to reach sub-goals assigned by proactive agents. This usually involves the use of simple Resources.

Proactive Agents are grouped into *clusters*. A Cluster is the set of Resource Agents that are serving the proactive agent at the moment. Proactive Agents are able to *ask* for a service only to Resource Agents in their clusters. Communication between agents is guaranteed by a very simple messages exchange protocol. It uses four kinds of messages: *Join*, used by Proactive Agents to ask for a Resource to join its cluster;*Leave* used by a Proactive to notify a Resource that it is no longer in its cluster;*SendGoal* to ask for a service; *OKJoin* an acknowledgement message.

The framework is developed by using Unity3D Engine[1] so it easily allows 3D simulations of planning problems. The framework uses the UPPAAL Reasoning Engine to perform counter-example searching. Two kinds of planners are implemented: a *Classical Planner* that is a simple implementation of the Breadth First and A* state space search algorithms, and *Counter Example Planner*. The presence of different planning strategies enables the implementation of a *Multi Expert* system. Multiple planners can start in parallel processes improving performances or providing solutions at different layer of abstraction. The Plan search algorithm in the Counter Example Planner works in four steps: (1) The TA (Timed Automata) translator implements an algorithm for Agent models translation from the model in sec.2 to Timed Au-

[1] `https://unity3d.com/`

tomata formalism; (2) The TA is passed to the UPPAAL Reasoning Engine in order to produce a counter example (if any) to the following formula: "**A [] !'GoalCondition'**" that is *"is it true that from initial state, only states where the goal condition is not satisfied are reachable?"*; (3) If the previous property is satisfied, then no plan exists for the current goal; otherwise, UPPAAL returns a *counter example* that is a path from the initial state to a state where the goal condition is satisfied, in form of an *Uppaal trace*; (4) The Uppal trace translator perform the translation from trace to a sequence of actions that is the agent plan.

4 Case Study

We will now show how agent systems can be modeled with our framework, through the definition of a case study. We consider a simulation of a critical system with a multi agent environment. this example simulates a warzone with wounded soldier and where robots automate rescue actions. The Environment ($World$) is a maze where there are a fixed number of points named *HotSpots*, where soldiers could be present, and where Search Robots can move. In the maze there are M wounded soldiers, N *Search Robots*, and K *Aid Robots* with N, M, K variable. At the beginning of the simulation, the M soldiers are randomly disposed on M different HotSpots; The role of the Search Robots is to find soldiers: in a single step a Robot will move from the current HotSpot to an adjacent one, with the condition that the target spot is not currently occupied by other robots and that it cannot go back to the previous position. If the robot finds a soldier in the new position, it will call an available Aid Robot to rescue him. An Aid Robots is in an idle state until it receives a rescue call from a Search Robot: when a call is received, it moves to the soldier position, rescues the soldier (that will disappear from the maze) and returns to its initial position. Simulation terminates when robots rescue all the soldiers. The environment is modeled by the following formulas:

- isAdjacent(HS_i,HS_j) : a facts defining the adjacence of the HotSpots (a Search Robot can only move from an HS to an adjacent one)
- isBusy($AidRobot_i$): a fact defining if an Aid Robot is currently serving a Search Robot
- isBusy(HS_i): a fact defining if an HotSpot is occupied

The Search Agents are modeled as Proactive Agents with the following Beliefs:

- position() ; the current HotSpot occupied by the agent
- isSoldierInPosition(); defines if there is a soldier in the current position
- prevpos(); the previous position occupied

The actions these agents can perform are of two kinds: *Move* and *Call*;the first one is modeled as follows:

- Precondition: (position()==X) AND (X!=Y) AND (prevpos!=Y) AND (isAdjacent(X,Y)==true) AND (isBusy(Y)==false)
- Effects: position()=Y,prevposition()=X, isBusy(Y)=true, isBusy(X)=false
- Variables: (X,"HotSpot","Existential"), (Y,"HotSpot","Existential")

where obviously "HotSpot" is the name of the domain defining all possible HotSpots names. The Call action is so modeled:

- Precondition: (isSoldierInPosition()==true) AND (isBusy(X)==false)
- Effects: (isSoldierInPosition()=false)
- Variables: (X,"AidRobot","Existential")

In this belief *AidRobot* is the name of the domain defining all Aid Robots in the environment. Search agents manage two goals at different priority: (1) *MoveGoal*, whose condition is (pos(X)!=X), of priority 1; (2) *CallAidRobot*, whose condition is (isSoldierInPosition()==false), of priority 2; this goal will be activated whenever the belief isSoldierInPosition() evaluates "true" Aid Robots are modeled as Resource agents, whose only belief is *soldierPosition()*, the position of the soldier currently healed (whose value can be the HotSpot name where the soldier is located, or null if the robot is in idle state) and with no goals in the initial (idle) state. The only action available is to move to the position of a soldier previously located by a Search Agent:

- Precondition: (soldierPosition!=null)
- Effects: (soldierPosition=null)

When a call is received by the Aid Robot agent, a Rescue goal whose condition is (soldierPosition()==null), will be activated. In order to analyze planner performances, we consider two kinds of simulations; in the first one, the number of the two kind of robots is the same(N=K), so that a Search Robot will always find an available Aid Robot when it locates a soldier position: in this simulation we'll study the average planning time employed by the planners in reaching the two different goals of the Search Robots in the environment (Move and CallAidRobot). In the second simulation, the number of Search Robots is greater then the Aid robot one(N=2,K=1); in this case, a Search Robot will not always find an available Aid Robot when it locates a soldier: in this situation the planner must notify the absence of a solution for the current goal.

In Fig.3 and Fig.4, we compare average planning time of two planners for the Search Robot goals when N and K increase (for the Counter Example Planner, we always consider the sum of the automata generation and planning time). Notice that planning time increases with N for both planners,but for Counter Example Planner this growth is much more evident. This lack of performances of the Counter Example Planner is caused by the *automaton dimension growth* : every Search Robot agent in the system, has to consider the position of all the other agents, so the number of possible states generated in the automaton grows very rapidly increasing the number of Search Robots. In the second kind of simulation, we consider the presence of a single Aid

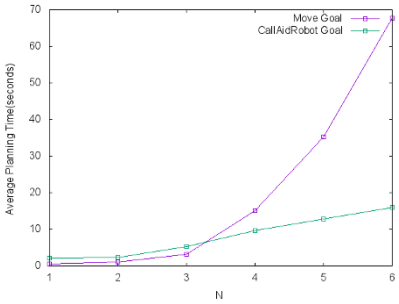


Fig. 3 Classical Planner Average Planning Time

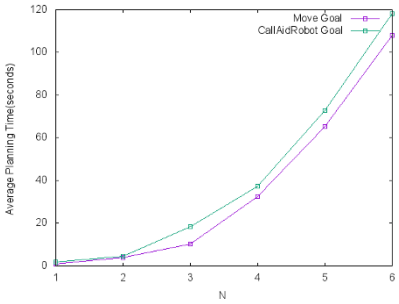


Fig. 4 Counter Example Planner Average Planning Time

Robot and of two Search Robots; in this case there is a clear difference of behaviour for the two planners: whenever a Search Robot finds a soldier and the only Aid Robot is serving the other one, the Classical Planner is not able to notify the absence of a solution for the CallAidRobot goal after hours of computation, while the Counter Example Planner notifies this situation within few seconds. This second example highlights the advantage of a Multi Expert approach: the Counter Example Planner can work in parallel with other planners (implemented to show the same interface and so inserted in the framework architecture) to stop any planning activity when it notifies the impossibility of achieving the current goal.

5 Conclusions and Future Works

In this paper we described a an autonomous planning methodology that is based on multi-agent models. The planning methodology exploits both classical and a new counter-example based approaches in order to build an effective multi-expert system able to face with increasing complexity of these systems. We presented a framework able to enact the methodology. The methodology and the framework have been applied to a case study of a distributed critical system[12, 13, 14]. Results shows how the counter-example planner executes faster than classical ones in some cases, motivating the use of a multi-expert system in the framework. We also show the ability of the new planner methodology when dealing with unreachable goals. Future works include the improvement of the counter-example planner algorithm and the extension of the methodology to conflicting interactions among agents and to choreographed execution of plans. Moreover we plan to test the use of embedded technology[15, 16, 17] in order to improve performances in state space exploration, by means of proper hardware accelerators.

References

1. Min Chen. Towards smart city: M2m communications with software agent intelligence. *Multimedia Tools and Applications*, 67(1):167–178, 2013.
2. Anas M Mzahm, Mohd Sharifuddin Ahmad, and Alicia YC Tang. Agents of things (aot): An intelligent operational concept of the internet of things (iot). In *2013 13th International Conference on Intellient Systems Design and Applications*, pages 159–164. IEEE, 2013.
3. Giovanna Sannino, Ivanoe De Falco, and Giuseppe De Pietro. Monitoring obstructive sleep apnea by means of a real-time mobile system based on the automatic extraction of sets of rules through differential evolution. *Journal of Biomedical Informatics*, 49:84 – 100, 2014.
4. Francesco Moscato, Flora Amato, Alba Amato, and Rocco Aversa. Model–driven engineering of cloud components in metamorp (h) osy. *International Journal of Grid and Utility Computing*, 5(2):107–122, 2014.
5. Jussi Rintanen, Keijo Heljanko, and Ilkka Niemelä. Planning as satisfiability: parallel plans and algorithms for plan search. *Artificial Intelligence*, 170(12):1031–1080, 2006.
6. Tim French, Nik Bessis, Fatos Xhafa, and Carsten Maple. Towards a corporate governance trust agent scoring model for collaborative virtual organisations. *International Journal of Grid and Utility Computing*, 2(2):98–108, 2011.
7. Evjola Spaho, Gjergji Mino, Leonard Barolli, and Fatos Xhafa. Goodput and pdr analysis of aodv, olsr and dymo protocols for vehicular networks using cavenet. *International Journal of Grid and Utility Computing*, 2(2):130–138, 2011.
8. Soichi Sawamura, Admir Barolli, Ailixier Aikebaier, Makoto Takizawa, and Tomoya Enokido. Design and evaluation of algorithms for obtaining objective trustworthiness on acquaintances in p2p overlay networks. *International Journal of Grid and Utility Computing*, 2(3):196–203, 2011.
9. Valentin Cristea, F. Pop, C. Stratan, A. Costan, C. Leordeanu, and E. Tirsa. A dependability layer for large-scale distributed systems. *International Journal of Grid and Utility Computing*, 2(2):109–118, 2011.
10. Daniel Burfoot, Joelle Pineau, and Gregory Dudek. Rrt-plan: A randomized algorithm for strips planning. In *ICAPS*, pages 362–365, 2006.
11. Michael Wooldridge. Agent-based software engineering. In *IEE Proceedings on Software Engineering*, pages 26–37, 1997.
12. Leonard Barolli and Fatos Xhafa. Jxta-overlay: A p2p platform for distributed, collaborative, and ubiquitous computing. *Industrial Electronics, IEEE Transactions on*, 58(6):2163–2172, 2011.
13. Fatos Xhafa, Raul Fernandez, Thanasis Daradoumis, Leonard Barolli, and Santi Caballé. Improvement of jxta protocols for supporting reliable distributed applications in p2p systems. In *Network-Based Information Systems*, pages 345–354. Springer, 2007.
14. Leonard Barolli, Fatos Xhafa, Arjan Durresi, and Giuseppe De Marco. M3ps: a jxta-based multi-platform p2p system and its web application tools. *International Journal of Web Information Systems*, 2(3/4):187–196, 2007.
15. A. Cilardo. New techniques and tools for application-dependent testing of FPGA-based components. *IEEE Transactions on Industrial Informatics*, 11(1):94–103, 2015.
16. A. Cilardo. Efficient bit-parallel $GF(2^m)$ multiplier for a large class of irreducible pentanomials. *IEEE Transactions on Computers*, 58(7):1001–1008, 2009.
17. A. Cilardo and E. Fusella. Design automation for application-specific on-chip interconnects: A survey. *Integration, the VLSI Journal*, 52:102–121, 2016.

Towards a coaching system for daily living activities: the use of kitchen objects and devices for cognitive impaired people

Alba Amato, Antonio Coronato, Giovanni Paragliola

Abstract Europeans live longer and the increase of life expectancy is naturally linked to the growth degenerative disease. ICT can contribute in maintaining cost efficient and high quality health and social care by providing European citizens with better and cheaper services for health and ageing well. The overall objective is to promote greater autonomy and individual psychosocial integration. Personalized coaches bear benefit to dementia patient in several ways: they help to remain independent, they maintain and stimulate his mental and physical capacity, they help him to consume excess energy and to facilitate sleep in the night. They also can encourage the patient to have the interest, to remain active and to do movement in order to help him to lead a life as normal as possible. With this aim the paper proposes a novel representation approach, based on a cognitive model, to design and develop a coaching system for daily living activities which relies on solving a problem as an human so modeling human cognition. The idea is to handle the process of acquiring skills as a transition from use of declarative knowledge of procedures that can be applied quickly and automatically in specific situations.

Alba Amato
National Research Council (CNR) - Institute for High-Performance Computing and Networking (ICAR), Italy e-mail: alba.amato@icar.cnr.it

Antonio Coronato
National Research Council (CNR) - Institute for High-Performance Computing and Networking (ICAR), Italy e-mail: antonio.cornato@icar.cnr.it

Giovanni Paragliola
National Research Council (CNR) - Institute for High-Performance Computing and Networking (ICAR), Italy e-mail: giovanni.paragliola@icar.cnr.it

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_31

1 Introduction

Europeans live longer, hence the costs of health and social care will rise in he next years. The increase of life expectancy is naturally linked to the growth degenerative disease. Among the diseases that most affect people over 65 are senile dementia, among these, the most common is Alzheimer's disease, that represents approximately 50-60% of all the dementia syndromes. Dementias are a group of complex diseases characterized by impaired cognitive performance (memory, reasoning , language ...) as to interfere with normal work and social activities of the patient. There are also non-cognitive symptoms involving mostly the behavior. ICT can contribute in maintaining cost efficient and high quality health and social care by providing European citizens with better and cheaper services for health and ageing well. As the global telecare and telehealth market is forecast to grow from 7.6 to 17.6 billion Euros already by 2017, therefore European economy can benefit from such growing market. We can achieve a triple win: a better quality of life for European citizens, innovation and growth for a competitive EU industry and more sustainable healthcare systems for society [1]. One of the main objectives is to increase the comfort and autonomy of life of the elderly or persons problems with using a user-centered and individualized approach. Another important objective is to increase the stay as long as possible at home, in the elderly independently. It is also, minimizing hospital stay, doing rehabilitative work at home. It is also important to avoid wrong lifestyle and provide tools for cognitive rehabilitation, coaching and training with the overall objective of promoting greater autonomy and individual psychosocial integration. Personalized coaches bear benefit to dementia patient in several ways: they help to remain independent, they maintain and stimulate his mental and physical capacity, they help him consume excess energy and to facilitate sleep in the night. They also can encourage the patient to have the interest, to remain active and to do movement in order to help him to lead a life as normal as possible. With this aim the paper proposes a novel representation approach,based on a cognitive model, to design and develop a coaching system for daily living activities which relies on solving a problem as an human so modeling human cognition. The idea is to handle the process of acquiring skills as a transition from use of declarative knowledge of procedures that can be applied quickly and automatically in specific situations. Computational cognitive models of human memory can provide a basis for modeling changes in self-efficacy and adherence to personalized behavioral goals in e-health behavior change systems [2]. The paper does not claim to present an implementation but the main objective of this paper is to present an idea, highlighting the potentialities and the novelty of the approach also showing the early-stage of the design and development of the prototype. The paper is organized as follows: Section 2 describes some some other solutions presented in the related literature; in Sections 3 and 4 the problem is described together

with the tool chosen to solve it. In Section 5 the proposed approach is shown. Conclusions and future work are in Section 6.

2 Related Work

There are several methodologies and applications related to the field of information and communication technologies for ageing well and e-health. It is possible to define five main topic areas, covering different aspects, including ambient-assisted living, telemedicine and e-health, monitoring, accessibility and user interfaces, robotics and devices for independent living, and HCI for ageing populations. The literature on activity tutoring and coaching is extensive, we focus here on those systems that offer real-time coaching. A detailed analysis of current use of tailoring techniques in real time physical activity coaching applications is presented in [3] where authors is also define a conceptual framework. In this paper the focus is on coaching that can be seen as a relationship between a coach and an user. The coach's task is to unleash the quality of the user to help him go to its full capacity and achieve maximum performance both at work and personally. The work is developed mainly by activating the individual motivations of the person, encouraging and stimulating the development of blocked potential. The similarities between coach and tutor seem to be many: the purpose, the roles or functions in the accompaniment of the subject, reflexivity . However, there are also marked differences. In fact, while the tutor's main task is to give instruction or to alert the user when something dangerous happens, the coaching has as main task to orient, support and develop user's potential, promoting active and proactive attitudes and stimulating his ability. The Health Coaching is a way to work with patients which aims to raise awareness and to encourage greater accountability in the management of their health . Thus coaching can effectively integrate the more traditional medical models of health education and the indication directive by the professional. The Health Coaching increases the ability of patients to improve their health, helping them to manage independently their own conditions rather than just listen to what they should do. In the last years many smartphone application and platform have been designed and developed to motivate and support users during several activities of different kind. In [4] MOPET is presented. It is a mobile guide designed to help users in fitness trail sessions where the user has to alternate jogging and exercising. The user runs along an indicated path and has to stop when he arrives at an exercise station. In each exercise station, the user finds an exercise tool that he has to use to perform a specific fitness exercise. MOPET helps users in navigation, to allow the user to follow the correct path in the fitness trail. MOPET also motivate users using audio and visual feedback of users speed. The feedback motivates the user to maintain an adequate speed during all the session, according to her feelings. Moreover when

the user reaches an exercise station, MOPET shows how to correctly perform the exercise. This application can also be used in rehabilitation for individuals with dementia with the aim of affect the way he performs an action or movement. A noticeable example of platform is The Move2Play system described in [5] where is presented a conceptual design of an innovative platform to stimulate healthy living and improve quality of life. A key aspect of the systems design is the mechanisms of activity recommendation and the concept of motivation obtained integrating different motivational facilities delivered in a non-obtrusive manner, taking into account the current context of the user. The main novelty of this approach is the introduction of gaming elements and cooperation in order to motivate people. A personal coach, designed and developed for old people is described in [6]. The development of pairs of perceivable tactile icons which could be used to communicate commands with the goal of reducing the visual burden experienced when users interact with mobile applications is presented. Tactile feedback can offer considerable potential to mobile users with situationally-induced impairments and disabilities (SIIDs). The findings informed guidelines for tactile interface developers and researchers regarding the importance of designing evaluation conditions such that appropriate real world characteristics are addressed. As stated in [3] those kind of systems are based on seven key concepts: feedback, inter-human interaction, adaptation, user targeting, goal setting, context awareness, and self learning. In this paper we present a coaching system for daily living activities based on a cognitive architecture. The novelty of our approach is in the utilization of a cognitive architecture, named ACT-R, in order to build the model so that the individual decision-making mechanism are 'immersed' within the environment in which they act.

3 Scenario

In order to design a coaching system for cognitive impaired people, it is necessary to identify the main task that those kind of system have to perform. Those activities represent the core functionalities and can be grouped in four set as shown in 1.

- Emergency and Safety: the prediction and the recovery of anomalous situations resulting from anomalous behaviors;
- Communication: the transmission of patient's data to the medical centre to be evaluated by physician or access to medical services in emerging situation;
- Health Monitoring: multi-parameter patient monitoring (blood pressure, glucose, etc);
- Training: mindbody training program personalized for the improvement of the patient's outcome.

Fig. 1 Coaching Activities

In this paper we focus on Emergency and Safety issue and we describe as an example [7] an home scenario concentrating on the cooking activities that a cognitively impaired patient could perform in a kitchen. In the cooking activities scenario, we focus on object misplacement as anomalous behavior that could happen modelling all those anomalous situations where an object is placed in an inappropriate place presenting present the classification of user behaviors and the model for the prediction of user actions.

4 Cognitive Architecture and ACT-R

The idea that is at the basis of our work is the utilization of models that are based on the structure of the brain and has been supported that provide excellent tools for the study of the mind and its workings. The development of this kind of models has played a very important role in research debate, particularly cognitive science that definitely represent a very interesting and fascinating research field. Cognitive architectures can be divided into several categories, which correspond to different goals of long-term research:

- Construction of cognitive models or the development of human behavior models, taking into account factors such as reaction time, errors. . . The use of cognitive architectures in this area has the advantage of using processes and structures consistent with human behavior. ACT-R is an architecture applied to many phenomena and tasks of this type.
- Development of an agent, when the objective is the development of agents in dynamic environments, even taking into account combinations of tech-

niques such as reactivity or planning, even if the main purpose is not to integrate purely human activities such as natural language or vast quantities of knowledge.
- Human-level agents, or agents whose range of cognitive ability has many, if not most, of the skills associated with humans. The difference with the previous point is that these architectures try to work with the higher number of problems.

Adaptive control of thoughtrational (ACT-R) [8] is the latest in a family of cognitive architectures, concerned primarily with modeling human behavior, that has seen continuous development since the late 1970s. The main architecture of the ACT-R[9] consists of 7 modules which are integrated and coordinated from a centralized system based on productions. Each module within ACT-R was constructed for the purpose of processing a different type of information. Modules (Figure2) are:

- The goal module: This module takes care of keeping track of the current goals. The initial goals can be initialized at system startup due to a specific command and then all requests to + goal buffer will result in instantaneous creation of a new goal chunk.
- The imaginal module: This module is connected with the homonymous buffer that can be used by a production to create a new chunk. These chunks are the internal representation of the information and will be saved within the declarative memory when the buffer cleaning. The module can only create one chunk at a time and each creation requires a certain amount of time during which the module is engaged.
- The perceptual-motor modules; Those modules are a set of modules that deal to interact with the external environment. In this category fall so the module vision, the auditory module, the speech module and motor module. The modules related to vision and listening are two buffers each one, instead of one as the standard modules. The first of the two buffers (aural or visual) will contain the chunk that encodes the information perceived by the environment, while the second (aural or visual-location-location) will contain a chunk that represents the position of an object or a sound to 'interior of the scene. A production may take one of the two modules to perceive an external information through a request-type + to visual or aural-buffer location-location. Within these requests can be specified via parameter: attended, if you want to hear an unknown element, instead of one recently observed, or an item appeared recently within the scene. At any given time only a limited number of chunks can be labelled as "attended" and maintains that state only for a small period of time, these labels represent the attentional and are called finsts marker. When talking about engine modules in ACT-R is considered alone hand movement. The manual buffer, associated with the motor module, contains no chunk and is only used as a means to make a request to the underlying module. The same observation also applies to the speech module and its vocal buffer. It

is very important to check the status of perceptual-motor before making a new application forms to avoid jamming a form that can only handle one request at a time.

- The procedural and declarative module module: These two modules dealing with store and activate the knowledge of the system. The procedural module is based on a match-select-fire mechanism: all productions whose premises are verified form a conflict-sets within which will be selected from a single production run. Within the declarative memory, compared to a request it returns a certain chunk if the content of its slot corresponds to the content of the chunk slot required.

Each module communicates to the outside only through its own buffer and the central system chooses the next action to be performed taking into account the contents of the buffer at the time of choice. Each buffer can contain a single unit of information, also called chunks. This choice limits the knowledge available to the productions system, following the intuition that the information available on any individual in a given instant is always in some way reduced: for example you are not always aware of all the information within its field of view, but only those to which you are devoting attention, as one is not aware of the entirety of the information in long-term memory, but only portions of which are from time to time recovered. Declarative knowledge is the kind of knowledge that is accessible to consciousness, which can be described, on which you can reason. A single declarative information units in the ACT-R is said chunk and consists of a name, a type and of the slots with the respective values. Procedural knowledge is, however, in all skills that can be demonstrated by the behavior but usually we can not describe, and we can not use it in a conscious way. This type of knowledge is encoded within the procedural module in the form of if-then type rules, called production rules. Inside the ACT-R each chunk must specify its type. The chunk-type operator allows you to define the types of chunk that you want to represent in the system, defining the slot that characterize them, any default values for these slots and, if present, the hierarchy relationships between types. ACT-R is in fact supported single inheritance between types of chunk thus allowing hierarchically organize the content of the declarative memory. Each buffer can be made requests of different types (identified by the symbol before the name of the buffer within the productions), which have a different behavior depending on the module that is involved, and the production area where it is used. These symbols are called in ACT-R action symbols. This software consists of a core LISP implementation, but there are many theoretically-motivated extensions and alternative practicality-motivated alternative implementations. In some cases the software provides the best definitions of what the theoretical claims are.

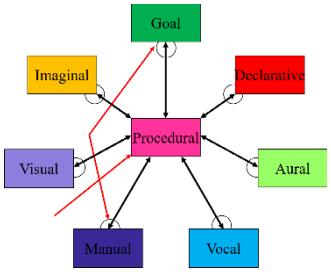


Fig. 2 ACT-R 7 Modules

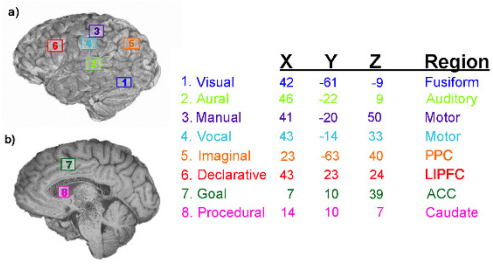


Fig. 3 ACT-R 7 Module-Region Mappings

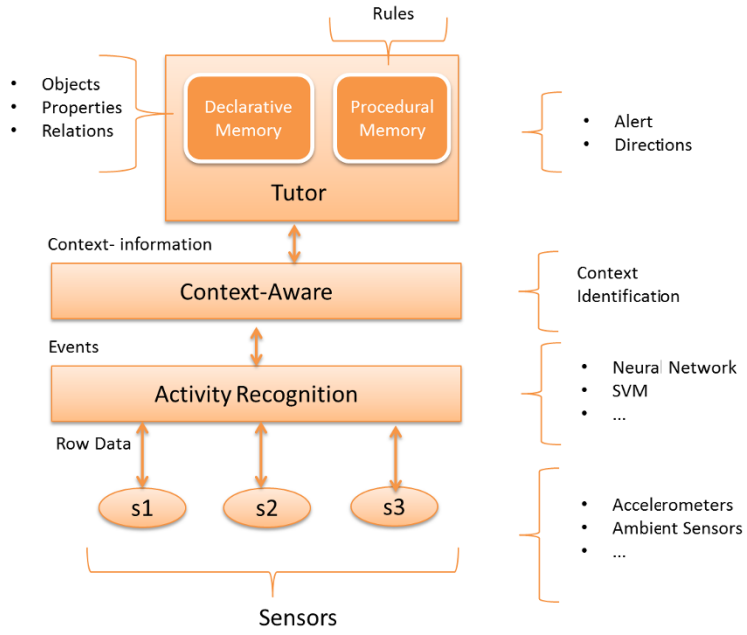


Fig. 4 Architecture Overview

5 Architecture

In the figure 4 we show an overview of the cognitive architecture. It is possible set four main components:
Sensors: they are in charge to provide row data from the environment. We have brought a set of sensors that they range to cinematic sensors like accelerometers from ambient sensors for the detection of parameters like light, sound and humidity. Cameras for the motion detection are take in consideration also.
Activity Recognition: this layer receives in input the row data acquired by the sensors and performs a pre-processing of the data in order to detect single

user's actions. The detail of activity recognition process is low at this level because we set the user's actions sensor-by-sensor. For a more details analysis we combine the single action at the upper layer.
Context- Aware: in this layer we combine the single actions detected at lower layer in order to provide an activity recognition of the users
Tutor: this level is in charge to assist the user to perform activity like cooking.
The proposed architecture monitors the user's activities and tutor him/her during his acting when it detects some incorrect actions, alerts or fixing advices are provided. In order to do that, it is necessary describe concepts, entities and actions of the environment.
In this paper we focus on the description of the tutor layer because it is the novelty we want to bring out.
The tutoring level is in charge to model the knowledge base of the domain of interest. As described before, act-r is composed of two modules: the declarative module and procedure module. The *declarative module* is in charge to describe the knowledge base in terms of entities, properties and relations of the domain of interest.

In order to assist the subject during living activity, it necessary describes the set of entities of interest.
In our scenario, we want to model entities belong at a home environment as a kitchen. Figure 5 shows how modeling physical entities like dish or knife in act-r. Each entities is a *chunck* which defines a concept, everything must be define as chunck either logical concept as electrical state *on,off*. Of course, in this step we have to declare each entities we will act with.

```
<chunk name="dish" type="chunk"/>
 <chunk name="knife" type="chunk"/>
 <chunk name="book" type="chunk"/>
 <chunk name="heater" type="chunk"/>
 <chunk name="inflammable" type="chunk"/>
 <chunk name="metallic" type="chunk"/>
 <chunk name="device" type="chunk"/>
 <chunk name="hasurface" type="chunk"/>
 <chunk name="movable" type="chunk"/>
 <chunk name="food" type="chunk"/>
 <chunk name="oven" type="chunk"/>
 <chunk name="pickup" type="chunk"/>
 <chunk name="placeon" type="chunk"/>

 <chunk name="on" type="chunk"/>
 <chunk name="off" type="chunk"/>
```

Fig. 5 Declaration of the entities

```
<chunk-type name="swicht_off">
 <slot name="object" equals="nil"/>
 <slot name="value" equals="nil"/>
 </chunk-type>

 <chunk-type name="swicht_on">
 <slot name="object" equals="nil"/>
 <slot name="value" equals="nil"/>
 </chunk-type>

 <chunk-type name="property">
  <slot name="object" equals="nil"/>
  <slot name="attribute" equals="nil"/>
  <slot name="value" equals="nil"/>
 </chunk-type>

 <chunk-type name="is-member">
  <slot name="object" equals="nil"/>
  <slot name="category" equals="nil"/>
  <slot name="judgement" equals="nil"/>
 </chunk-type>
```

Fig. 6 Declaration of properties

Declaring the entities is the first step to do in order to model the domain of interest. The second step is to declare the properties of the entities.
Figure 6 shows examples of those properties like *switch-on* which describe the property of a entity to be switched, or *is-member* which describes the property of an entity to be member of a category, etc.

The last step of the declaring process is to state which kind of relations exists between the entities.
Figure 7 shows some examples of relations. In order to coach the subject during his/her living activities it is important model some actions also. Examples are placing an object on a surface or picking an object.

```
<chunk name="p4" type="property">
  <slot name="object" equals="oven"/>
  <slot name="attribute" equals="device"/>
  <slot name="value" equals="true"/>
 </chunk>

 <chunk name="p5" type="action_so">
  <slot name="name" equals="pickup"/>
  <slot name="object" equals="dish"/>
 </chunk>

 <chunk name="p6" type="action_mo">
  <slot name="name" equals="placeon"/>
  <slot name="object" equals="dish"/>
  <slot name="entity" equals="oven"/>
 </chunk>
```

Fig. 7 Declaration of the relations among the entities

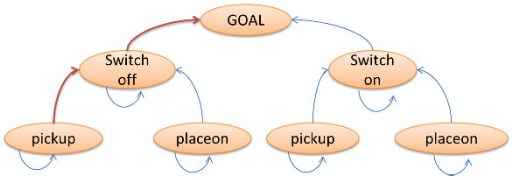


Fig. 8 Set of Actions

The second module, the *procedure module* define the production rules. A production rule is describe by two parts, a condition and an action. The condition of a production state the test of the content of buffers and modules. The action of a production specifies the actions to be done when a rule is matched. Basically, it describes the changes to be made to set of chunks in buffers. For the sake of brevity, we report only two examples of production rules.

Figure 9 shows an interesting example about how the production rule change the value of the chunck in the buffer. In the declare module we state that a dish has the property to be inflammable and a oven is an heater. By using on this two facts, it possible to define the rule that verifies if a dish is placed on oven then it will change the state of the oven switching it from on to off. The usage of the architecture is oriented to coach the subject in his/her daily activities. In other words, we want to submit to the architecture a specific challenges which describe simple activities that the patient could perform. The target of the system is to assist the subject in these challenge by monitoring its activities. As example, let suppose the system has to support the subject to performs a really simple activity like pick up a dish on a oven and switch the oven off. Figure 8 show a possible set of sequence actions which the subject could perform. The system have to tutor him/her to avoid

```
<production name="checkDishOnOven">
 <conditions>
  <match buffer="goal" type="state">
   <slot name="value" equals="=true"/>
  </match>
  <match buffer="goal" type="action_mo">
   <slot name="name" equals="=placeon"/>
   <slot name="object" equals="=dish"/>
   <slot name="entity" equals="oven"/>
  </match>
  <query buffer="retrieval">
   <slot name="state" equals="true"/>
  </query>
 </conditions>
 <actions>
  <modify buffer="goal" >
   <slot name="entity" equals="off"/>
   <slot name="object" equals="oven"/>
   <slot name="name" equals="swicht_off"/>
  </modify>
 </actions>
</production>
```

Fig. 9 Rule production to check the placing of an entity

```
<production name="checkDishisPickuped">
  <conditions>
   <match buffer="goal" type="state">
    <slot name="value" equals="=false"/>
   </match>
   <match buffer="goal" type="action_so">
    <slot name="name" equals="=pickup"/>
    <slot name="object" equals="=dish"/>
   </match>
   <query buffer="retrieval">
    <slot name="state" equals="true"/>
   </query>
  </conditions>
  <actions>
   <modify buffer="goal" >
    <slot name="name" equals="pickup"/>
    <slot name="object" equals="dish"/>
   </modify>
  </actions>
 </production>

 </procedural-memory>
```

Fig. 10 Rule production to check the taking of an entity

to perform incorrect actions or miss the sequence of corrected actions by submit suggests or directions to the subject. By defining the goal and set a challenge to reach the goal, the architecture is able to support the patient during his activities.

6 Conclusions

Dementias today represent an epochal challenge for public health, being the age the greater risk factor associated to their occurrence and being Europe, and Italy in particular, at a time of demographic development which sees the population becoming increasingly older. The e-health represents a real paradigm of innovation, and it is emerging as a new industry alongside pharmaceuticals and medical devices within which, ICT has had and has a key role.

In this paper a new approach to handle a variety of dangerous situations to abnormal behaviors of cognitive impaired patients is presented. The proposed approach is based on a cognitive model and can profitable used to detect dangerous situations and analyze abnormal behaviors. Using this approach, guidance and coaching can be done with patients whose cognition

is impaired. An overview of the cognitive architecture has been described, together with an example of The system have to tutor him/her to avoid to perform incorrect actions or miss the sequence of corrected actions by submit suggests or directions to the subject. Significant efforts in this area is needed to address the complex issues of coaching of cognitive impaired patients. The paper begins to investigate a novel solution to the problem exploring the possibility of utilization of a cognitive architecture able to support the patient during his activities by defining the goal and set a challenge to reach the goal in a non invasive way. A potential limitation of adopting this solutions is to find patients that can participate in a systematic exercise program under the guidance of caregivers. It is not possible, at this stage, to give an outlook on a real evaluation of the system, nevertheless an ongoing work is to develop a prototype that can be used to perform several tests.

Acknowledgements This work is partially supported by the eAsy intelligent service Platform for Healthy Ageing (ALPHA) Project.

References

1. Commission, E.: ehealth and ageing. Technical report (2016)
2. Pirolli, P.: A computational cognitive model of self-efficacy and daily adherence in mhealth. Translational Behavioral Medicine (2016) 1–13
3. Akker, H., Jones, V.M., Hermens, H.J.: Tailoring real-time physical activity coaching systems: A literature survey and model. User Modeling and User-Adapted Interaction **24** (2014) 351–392
4. Buttussi, F., Chittaro, L., Nadalutti, D.: Bringing mobile guides and fitness activities together: A solution based on an embodied virtual trainer. In: Proceedings of the 8th Conference on Human-computer Interaction with Mobile Devices and Services. MobileHCI '06, New York, NY, USA, ACM (2006) 29–36
5. Bielik, P., Tomlein, M., Krátky, P., Mitrík, v., Barla, M., Bieliková, M.: Move2play: An innovative approach to encouraging people to be more physically active. In: Proceedings of the 2Nd ACM SIGHIT International Health Informatics Symposium. IHI '12, New York, NY, USA, ACM (2012) 61–70
6. Qian, H., Kuber, R., Sears, A.: Towards developing perceivable tactile feedback for mobile devices. International Journal of Human-Computer Studies **69** (2011) 705 – 719
7. Coronato, A., Paragliola, G. In: A Safe Kitchen for Cognitive Impaired People. Springer International Publishing, Cham (2013) 17–25
8. Anderson, J.R., Matessa, M., Lebiere, C.: Act-r: A theory of higher level cognition and its relation to visual attention. Hum.-Comput. Interact. **12** (1997) 439–462
9. Anderson, J.R., Bothell, D., Byrne, M.D., Douglass, S., Lebiere, C., Qin, Y.: An integrated theory of the mind. PSYCHOLOGICAL REVIEW **111** (2004) 1036–1060

Wellness & LifeStyle Server: a Platform for Anthropometric and LifeStyle Data Analysis

Giovanna Sannino[1], Alessio Graziani[2],
Giuseppe De Pietro[1], and Roberto Pratola[2]

[1] Institue of High Performance Computing and Networking (ICAR-CNR),
Via Pietro Castellino 111, Naples 80131 Italy
{giovanna.sannino, giuseppe.depietro}@icar.cnr.it
[2] Engineering Ingegneria Informatica S.p.A,
Via San Martino della Battaglia, 56 - 00185 Rome - Italy
{alessio.graziani, roberto.pratola}@eng.it

Abstract. Health self-tracking has become a popular research issue in recent years. This interest is due to a number of partially related causes, such as the increase in the number of people with long-term health conditions, the growing importance of self-care, the spread of miniaturised and easy-to-use measuring devices and smartphone applications and the availability of social networking platforms. With some basic computer skills and a few low-cost gadgets, it is relatively easy to produce and share an amount of personal health information unimaginable only a few years ago. Within the project SmartHealth 2.0 we have realised a wellness platform aimed at supporting the prevention of an individual's unhealthy behaviors and the monitoring of his/her lifestyle habits. The platform consists of a suite of apps specialized for specific areas of wellness, and a remote component that we have called the Wellness & LifeStyle Server (WLS). In this paper we have detailed the design and the development of this innovative platform.

Keywords: Wellness services, anthropometric assessment, activity monitoring, diet monitoring, healthcare

1 Introduction

Nowadays, every healthy person should be able to independently monitor, without the assistance of specialized staff, certain parameters of his/her health by using simple and easy to use tools. The protracted application of such tools would bring enormous benefits in terms of healthcare promotion by helping to increase an individual's awareness of the state of his/her physical health and of the essential components of a healthy lifestyle.

Putting the individual at the center of such a prevention and monitoring project is the principal idea behind the project SmartHealth 2.0 [1], within which we have realised a wellness platform aimed at supporting these specific objectives.

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_32

Through a set of services, we aim to guarantee to each person the possibility of monitoring over time his/her own physical state, his/her way of eating, the type and the amount of physical activity that he/she undertakes, and the environment in which he/she moves, in order to collect specific physiological parameters and to ensure that these data are processed according to a common and compatible logic. The platform consists of a suite of apps specialized for specific areas of wellness, such as [2, 3], and a remote component that we have called the Wellness & LifeStyle Server (WLS).

This proposed wellness model enables each user to install on his/her smartphone the set of apps that he/she deems appropriate to monitor his/her state of health. Each installed app will monitor a particular aspect of the user's lifestyle and, taking advantage of mechanisms of unification of user-identity and of the sharing and processing of data through the WLS, will contribute to the overall lifestyle monitoring. In fact, through the services of the WLS, each app will be able to make use of the logical processing of lifestyle monitoring data, without the user having to implement such data directly into each app, with the consequent advantage that all the suite apps will be able to provide common indications and suggestions.

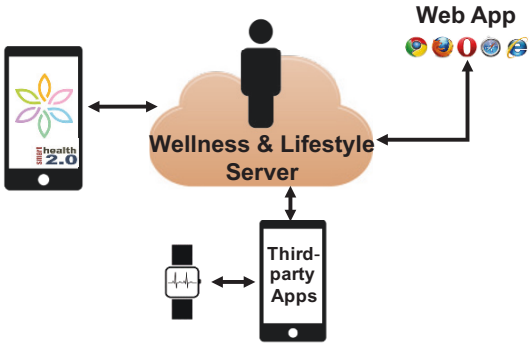


Fig. 1: Wellness Model.

2 Business Processes of the Wellness Server

The business processes that govern the logic of the WLS respond to the characteristics of the macro process shown in Fig. 2, especially regarding the processes that manage the steps of Collection, Analysis and Planning. Meanwhile, the phases of Execution and Monitoring are delegated to external client components, as for example the apps described in [2, 3].

In this type of process the temporal dimension always represents a fundamental element. The data acquired in the monitoring phase are variable by nature and need to be stored as a function of time so that you can rebuild every moment

Fig. 2: Monitoring Cycle of Wellness Server.

of the event history. Any stored data are representative and meaningful values only if they refer to a specific moment in time.

The event of the receipt of the monitoring data is the primary trigger that activates the logic of data processing by the WLS. Alternatively, the Collection, Analysis and Planning phases produce new wellness goals and new assessments useful for the user. In this perspective, the WLS acts as a passive element waiting for any user monitoring data relating to his/her lifestyle.

The WLS differentiates between the processing of anthropometric data and lifestyle data (nutrition, physical activity, sleep, etc.).

1. **LifeStyle data processing**: this is the process which is responsible for analyzing the monitoring data about nutrition, physical activity and sleep. The processing of these data allows the system to ascertain the user's real lifestyle habits and analyze them to identify possible unhealthy behaviors.
2. **Anthropometric data processing**: this is the process that manages the analyzing of the user's physical state and, indirectly, the monitoring of his/her lifestyle. Whenever the client sends an anthropometric measurement, the system produces a user's physical state evaluation helpful both to identify unhealthy situations and to determine patterns and trends through a comparison with previous assessments. Furthermore, on the basis of this assessment and if clear situations of excess fat accumulation have been identified, the system can activate a more elaborated monitoring task by formulating a specific slimming plan.

2.1 LifeStyle Data Processing

This section describes in detail the lifestyle data processing implemented by the WLS. The process has been broken down into two main sub-processes:

– **Receiving the lifestyle monitoring data (the Collection Phase)**: this deals with the validation and saving of lifestyle monitoring data. A monitoring app running on the user's smartphone collects data on his/her habits

and lifestyle. Periodically, in accordance with the app.'s own internal logic, this information is sent to the WLS, and is processed and saved in the users personal repository.

– **Life-style data processing (the Analysis Phase)**: this manages, separately and depending on the type of information, the monitoring data. In particular, the following processing is provided:
  - *Physical Activity Data Processing* allows the user to check the amount and type of activity performed. The amount is compared with weekly targets to verify that the user's behavior is close to the ideal set by the system. Through the analysis of the type of activity, the user can instead determine if he/she is performing an activity compatible with his/her wellness goals. It also performs a calculation of the number of calories burned by this physical activity, this being helpful to check the daily energy balance.
  - *Nutrition Data Processing* allows the user to verify his/her daily nutrition. It provides a check of whether the amount of calories consumed is in line with the user's defined reference caloric requirements. No check is carried out regarding the quality of the food consumed.
  - *Sleep Data Processing* allows the user to verify the quantity and quality of his/her sleep. It tests the adequacy of the daily amount of sleep, and also, by means of suitable indices, provides an indication of its quality.

2.2 Anthropometric Data Processing

This section describes in detail the process of the elaboration of the anthropometric data implemented by the WLS when it receives the data sent by the user. The process has been broken down into its main sub-processes:

1. **Receiving anthropometric data (the Collection Phase)**: this deals with the validation and saving of the anthropometric data sent by the user. Sending anthropometric data is an operation that the user should carry out every 30-40 days. Through this operation, the user allows the WLS to become aware of his/her physical state and to put in place all the logic needed to recognize possible unhealthy situations, indicating the need for certain lifestyle changes. In addition, comparing the new data with the anthropometric data sent previously, the WLS can identify trends and tendencies relating to significant lifestyle habits.
2. **the Anthropometric data processing (Analysis Phase)**: this deals with the production and saving of the anthropometric assessment. Subsequently, it monitors any slimming plan (if there is one active) or checks whether it is necessary to create a new one. Finally, in the case where the user does not require a specific plan, the system simply carries out a comparison between the current and the previous anthropometric rating in order to identify patterns and trends that indicate (indirectly) the presence of any unhealthy behaviors in the user's lifestyle.

3. **Computing the anthropometric assessment (the Analysis Phase)**: this produces the anthropometric assessment starting from the data provided by the user. An anthropometric evaluation should contain the main indices used in the Body Composition Analysis, namely the Body Mass Index (BMI), Percentage Body Fat (BF%), Fat Free Mass index (FFMI), Fat Mass Index (FMI), Fat Mass (FM), Fat Free Mass (FFM), and the Waist to Height Ratio (WHtR). The presence of these indices, combined with the basic anthropometric data (height, weight, waist, etc.), gender and age, allows any third-party software (or a person with adequate expert knowledge) to be able to freely interpret the anthropometric assessment.
4. **Definition of a slimming plan (the Planning phase)**: this is the process responsible for creating a slimming plan. It determines weight goals, duration and daily calorie needs. The plan, in line with the expectations of a style monitoring platform based on the analysis of the physical state of a user, does not provide the definition of any athletic program to perform, nor any particular diet to follow, but stimulates the user to contact a specialist to formulate a meal plan and/or a targeted sports program.
5. **Verifying the trend of a slimming plan (the Analysis-Monitoring Phase)**: this verifies the slimming plan trends by comparing the expected data with the monitored data. The aim is to provide a comparison between the expected health status and the real health status to determine whether the user is on line, or not, with the plan. In any case in which a weight loss in line with that expected by the plan is identified, the verification process verifies if the goal has been reached and in that case successfully ends the plan. Otherwise, if the real health status is not in line with the plan, the system checks if it is possible to successfully complete the plan in the remaining time period. In such a case, it is necessary to inform and motivate the user, because it is clear that a greater commitment is needed than that established at the beginning of the plan. In a case where there seems to be no chance of completing the plan successfully, the process cancels this current plan (which has been unsuccessful), and reformulates a new one by re-running the process of the definition of a slimming plan.

Anthropometric Assessment The anthropometric analysis is the evaluation of the physical and structural characteristics of a person by measuring weight, lengths and body circumferences. In a healthy adult (free of any conditions which may influence the shape of the body) three factors have a determining influence on the modeling of a body: genetics, nutrition and physical activity. Therefore, it is possible to think that the body of a person is the best proxy for their lifestyle (more precisely, the best indicator of the energy balance of a person) and its monitoring allows, even without knowing much of the user's lifestyle, an identification of the physical imbalances caused (most likely) by incorrect behaviors and habits. To put it more simply, through the anthropometric monitoring, the system is able to answer the following questions about the user:

- How am I?: an analysis of the current physical state, with an identification of clear situations of non-normality.
- How should I be?: a definition of values and of an ideal reference range based on age and gender.
- What should I do?: a definition of ideal lifestyle characteristics (food& sport).

The methods and techniques studied to realize the obesity evaluation by the WLS are inspired by the results of an Italian study conducted by the University of Pavia and the Foundation IRCCS Salvatore Maugeri [4]. The research objective of Colombo ed al. was to compare the performance of the main indices used in the medical field in the overweight classification using the same sample of anthropometric data (63 volunteers of either sex aged between 20 and 65 years).

In detail, to calculate the %BF the WLS uses a statistical method based on predictive equations. These equations have been developed by Hodgdon and Beckett [5], and are based on the following anthropometric measurements:

- **neck**: circumference in centimeters of the neck just below the larynx;
- **ab1**: minimum waist circumference in centimeters between the rib cage and the navel (for women);
- **ab2**: waist circumference in centimeters at the navel (for men);
- **Hip**: hip circumference in centimeters measured at the height of the femoral head (for women only).

The following formula allows you to calculate the Body Density in males:

$$\begin{aligned} MaleBD = &-[0.19077 * \log(ab2 - neck)] + \\ &+[0.15456 * \log(height)] + 1.0324 \end{aligned} \tag{1}$$

Instead, the following formula is used for females:

$$\begin{aligned} FemaleBD = &-[0.35004 * \log(ab1 + hip - neck)] + \\ &+[+0.221 * \log(height)] + 1.29579 \end{aligned} \tag{2}$$

After calculating the Body Density of a user, the BF% is calculated by using the formula of Siri [6]:

$$\%BP = 100 + (\frac{4.95}{BD} - 4.5) \tag{3}$$

The FMI and FFMI [7] are two equivalent indices to the BMI index [8], and they are used to calculate the index of lean mass and fat mass (relating the fat mass and lean mass to the height).

The analysis of the FMI and FFMI, unlike the analysis of the BMI only, allows you to understand the actual body composition of an individual, and the study of their changes over time allows you to gain a real awareness of how the body composition of an individual is changing [7, 9].

The following definitions show the conceptual equivalence of the three formulas:

$$BMI = \frac{weight(kg)}{height(mt)^2} \tag{4}$$

$$FMI = \frac{FatMass(kg)}{height(mt)^2} \qquad FFMI = \frac{LeanMass(kg)}{height(mt)^2} \tag{5a, b}$$

Then, we can mathematically write:

$$BMI(kg/m^2) = FFMI(kg/m^2) + FMI(kg/m^2) \tag{6}$$

Next, to calculate the BMI and FMI, it is necessary to know the weight in kg of lean body mass (FFM) and of fat body mass (FM). It is possible to derive these data knowing the weight and the percentage of fat mass of an individual.

$$FM(kg) = \frac{\%BF * weight(kg)}{100} \tag{7}$$

$$FFM = \frac{(100 - \%BF) * weight(kg)}{100} \tag{8}$$

Finally, the WLS bases its assessments on a very simple index calculated and used for the measurement of abdominal fat. This index, known as the WHtR, has become very popular thanks to studies conducted in Europe by Dr. Ashwell and is considered one of the best screening tools in the prevention of cardiovascular diseases and problems related to hypertension [10].

The WHtR is calculated by dividing the waist measurement by the height expressed in cm.

$$WHtR = \frac{waist(cm)}{height(cm)} \tag{9}$$

3 Wellness Server Architecture

The platform, in compliance with the architectural models at multiple levels, has been conceived as an aggregation of three subsystems:

- **WLS Services (the Business Layer)**: this is the subsystem that implements all the business logic and authentication of the wellness platform. The WLS Services subsystem is not directly accessible from the outside, but is usable (in an indirect way) only through the two WLS REST APIs and WLS Web Application subsystems designed for access to the platform.
- **WLS REST API (the Communication Layer)**: through the subsystem, the WLS REST API services offered by WLS Services are accessible to both native apps developed as part of A6.5 project and to any app developed by third parties.
- **WLS Web Application (the Presentation and Communication Layer)** is the web application necessary to access the system. The interaction between the platform and the users should not only be possible via the app (which can also be developed by third parties), but also through the official portal of the platform. The website plays exactly this role. Through

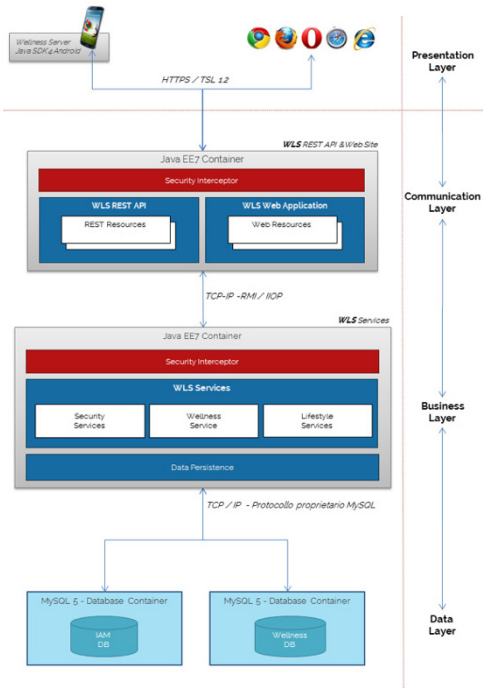


Fig. 3: Software Architecture of the Wellness Server.

the website a user can sign up to the platform, manage his/her data, and monitor his/her physical state and the healthcare activities that he/she has undertaken.

Figure 3 shows the logical architecture of the system, and provides also some important technological information. In particular, it involves the Java technology. In fact, the three sub-systems are developed as Enterprise Applications implemented by using the Java Language and compatible with the specific JEE7.

As shown in figure 3, the system is deployed in two separate containers, one used for the web part (the WLS Web Application and WLS REST APIs) and the other used for the business services (the WLS Services). One of the advantages of having designed the platform by separating it into three distinct subsystems is to make possible different deployment configurations, a feature capable of making the system very flexible and especially adaptable to different operating conditions.

3.1 Wellness Service Architecture

The WLS Services subsystem is the heart of the wellness platform. It implements and provides the remaining components of the platform services for the lifestyle monitoring users.

The main component is the *Measurement Service*, as shown in figure 4 which, by receiving the measurements relating to the wellness activities carried out by the user, activates the business processes described in the previous sections.

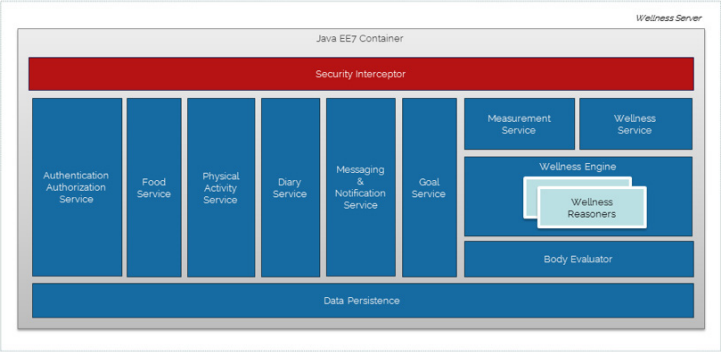


Fig. 4: Software Architecture of the Wellness Service.

In detail, the role and main responsibilities of each component are:

- **the Wellness Service**: this is the component responsible for the services connected with the improvement of the user's physical state. Through this component, you can reconstruct the history of the changes in the physical condition of a person over time.
- **the Measurement Service**: this is the component responsible for services related to the receipt and saving of measurements sent by clients. The engine works with the Wellness and Diary Service.
- **the Body Evaluator**: this is the component responsible for the processing of elementary anthropometric measurements in order to produce an assessment of the physical state of an individual.
- **the Wellness Engine**: this is the component that handles the business processes related to the processing of the measurements sent by clients. It is activated by the Measurement Service and works with the Body Evaluator component, the Diary Service and the Messaging And Notification Service
- **the Food Service**: this is the component responsible for services related to navigation in the food database of the wellness platform (in order that it is possible for the WLS to receive measurements about the power of a user, the client must indicate all foods consumed using those known to the WLS).
- **the Physical Activity Service**: this is the component responsible for services related to navigation in the database of the physical activity.
- **the Diary Service**: this is the component responsible for services related to navigation in the users wellness diary.
- **the Messaging and Notification Service**: this is in charge of the services component for the reporting and and retrieving of messages for the user.

Through this service the user and the app can be informed of the results of the business run as background processes in the WLS Server.
- **the Goal Service**: this is the service responsible for retrieving specific monitoring plans called Goals (monitoring plans are created only if the system identifies certain conditions considered as unhealthy).
- **the Autentication and Authorization Service**: this is the service responsible for the authentication and authorization procedures.
- **the Security Interceptor**: this is the component responsible for intercepting incoming requests and activating the security checks necessary for the execution of the request being made. It collaborates with the Authentication and Authorization Service.

4 Conclusions

In this paper we have detailed the design and development of the WLS, one of the components of the platform realized within the project SmartHealth 2.0, aimed at supporting the prevention of an individual's unhealthy behaviors and the monitoring of his/her lifestyle habits.

Through a set of services, the WLS guarantees to each individual the possibility of monitoring over time his/her own physical state, his/her way of eating, the type and the amount of physical activity that he/she undertakes, and the environment in which he/she moves, in order to collect specific physiological parameters, and to ensure that this data is processed according to a common and compatible logic.

A preliminary experimental phase has demonstrated the validity of the WLS, in terms of its reliability, stability, usability, and facility of use. Additionally, a trial phase of the platform has been carried out in four districts of the Calabria Region (Crotone, Cosenza, San Giovanni in Fiore and Lamezia Terme). We are now involved in the analysis of the data collected during the trial to investigate and demonstrate the real validity of the WLS.

Additionally, we are now investigating the possibility to improve the Autentication and the Authorization Service in order to solve problems related to complex security and privacy policies about wellness data. In fact, the access to this kind of data should be granted or denied according to the right and the need of the healthcare workers to perform a particular job function on specific data sections. To do that, could be used some fine-grained access control policies working on semi-structured electronic wellness & lifestyle records[11–13].

Acknowledgments

The authors gratefully acknowledge support from Engineering S.p.A. and from the project Smart Health 2.0 (PON04A2_C). A special thanks is also due to Dr. Giorgio Amedeo Agrò for his invaluable contribution to this study.

References

1. Smart Health 2.0, SH.2.0. (2014). Available at: http://www.smarthealth2-0.com
2. M. Forastiere, G. De Pietro, and G. Sannino. An mHealth Application for a Personalized Monitoring of One's Own Wellness: Design and Development. In Innovation in Medicine and Healthcare 2016 (pp. 269-278). Springer, 2016.
3. L. Verde, G. De Pietro, P. Veltri, and G. Sannino. An m-health system for the estimation of voice disorders. In Multimedia & Expo Workshops (ICMEW), 2015 IEEE International Conference on (pp. 1-6). IEEE, 2015.
4. O. Colombo, S. Villani, G. Pinelli, C. Trentani, M. Baldi, O. Tomarchio, & A. Tagliabue, *To treat or not to treat: comparison of different criteria used to determine whether weight loss is to be recommended*. Nutrition journal, 7(1), 1, 2008.
5. J. A. Hodgdon , M. B. Beckett (1984). Prediction of percent body fat for US Navy women from body circumferences and height. Naval Health Research Center (CA).
6. W. E. Siri, *Body composition from fluid spaces and density: analysis of methods. 1961*. Nutrition, 9(5), 480-491, 1993.
7. U. G. Kyle, Y. Schutz, Y.M. Dupertuis, & C. Pichard, *Body composition interpretation: contributions of the fat-free mass index and the body fat mass index*. Nutrition, 19(7), 597-604, 2003.
8. Z. Mei, L. M. Grummer-Strawn, A. Pietrobelli, A. Goulding, M.I. Goran, & W.H. Dietz, *Validity of body mass index compared with other body-composition screening indexes for the assessment of body fatness in children and adolescents.* The American journal of clinical nutrition, 75(6), 978-985, 2002.
9. Y. Schutz,U. U. G. Kyle, & C. Pichard, *Fat-free mass index and fat mass index percentiles in Caucasians aged 18-98 y.* International Journal of Obesity & Related Metabolic Disorders, 26(7), 2002.
10. M. Ashwell, P. Gunn, & S. Gibson, *Waisttoheight ratio is a better screening tool than waist circumference and BMI for adult cardiometabolic risk factors: systematic review and metaanalysis*. Obesity reviews, 13(3), 275-286, 2012.
11. F. Amato, G. De Pietro, M. Esposito, and N. Mazzocca, An integrated framework for securing semi-structured health records. Knowledge-Based Systems, 79 (pp. 99-117), 2015.
12. F. Amato, F. Moscato. A model driven approach to data privacy verification in E-Health systems. Transactions on Data Privacy 8.3 (pp. 273-296), 2015.
13. F. Amato, F. Colace, L. Greco, V. Moscato, A. Picariello, Semantic processing of multimedia data for e-government applications, Journal of Visual Languages & Computing, vol. 32 (pp.35-41), 2016.

Semantic Information Retrieval from Patient Summaries

Mario Sicuranza[1], Angelo Esposito[1 2], Mario Ciampi[1]
[1] Institute for High Performance Computing and Networking
of the Italian National Research Council, Naples, Italy
[2] Department of Engineering, University of Naples "Parthenope", Naples, Italy
mario.sicuranza@na.icar.cnr.it, angelo.esposito@na.icar.cnr.it, mario.ciampi@na.icar.cnr.it

Abstract. This paper presents a novel architectural model approach for the extraction of strictly necessary information from Patient Summary documents. These clinical documents, synthesizing the medical history of a patient, contain heterogeneous information, characterized by different confidentiality levels. For this reason, patients should be able to define privacy rules acting on the specific information they intend to share. However, the main system approaches currently used permit a patient to specify privacy policies only at document level. Thus, a patient is obliged to deny the access to the whole document to protect his/her privacy if only one kind of information is assumed particularly sensitive. To face this problem, the proposed architectural model is provided with semantic-enhanced features with the aim of allowing patients to define access rights based on document sections or information nuggets, trying moreover to fill the semantic gap between the queries and the information present in the documents.

Keywords—*Information Extraction; Privacy Protection; Patient Summary; Natural Language Processing.*

1 Introduction

Multiple studies have demonstrated that the application of the Information and Communication Technology (ICT) in healthcare allows improving the quality of care services, reducing medical errors, and decreasing costs [1]. One of the most important feature provided by the use of ICT in this sector is represented by the possibility of enabling health professionals to immediately retrieve clinical information about a patient of interest. This feature is typically offered by suitable technological systems opportunely designed and developed for exchanging digital clinical information among Health Information Systems (HISs) of different healthcare facilities or organizations [2].

Many technological systems for clinical information exchange have been developed worldwide for making heterogeneous HISs interoperable each other. In the European context, the most important results have been achieved in the epSOS project [3], whose aim was the design and implementation of an interoperability framework

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_33

for making national HISs of the EU Member States able to exchange two kind of clinical documents structured according to the HL7 CDA Rel. 2.0 standard [4], which are ePrescription and Patient Summary (PS). More in general, a large number of platforms for HIS management have been implemented at national, regional, and local levels for exchanging also other kind of clinical documents [5].

A PS is a document that gathers important social-health information about a patient, regarding his/her medical history. It consists of several sections reporting different kind of clinical content: in particular, relevant clinical content is contained in the *Allergy* and *Problem* sections, which report respectively the list of allergies and the list of pathologies and/or chronic diseases of the patient. All the information present in a PS is registered by the General Practitioner (GP) of the patient. The consultation of this information is a valuable option for a physician, as it allows faster clinical evaluation of the patient, both in case of emergency or more in general when a health problem occurs.

However, considering the extreme heterogeneity of such clinical content, the sensibility of the data contained in a PS, even depending on the subjectivity of the patient, can extensively vary: from data with low level of confidentiality (e.g. immunization) to data highly sensitive (e.g. seropositivity). This aspect, in particular, may arise several problems related to patient privacy, as national and international specifications have a propensity for the patient centralization in the choice of privacy policies [6]. As a matter of fact, a patient has to be able to specify the access rights to her/his clinical documents to be satisfied by the health professionals interested to consult them. Thus, if a PS contains data that the patient considers highly confidential, he/she could associate to the document very restrictive access policies, in this way denying the access also to other information present in the document that the patient could assume less sensitive [7]. For this reason, the technological systems currently used would have to be evolved with further semantic-enhanced features aiming at permitting a more flexible access to the information present in a document according to the privacy policies expressed by the patient.

This paper presents a novel architectural model approach based on semantic features for retrieving suitable information from PS documents, satisfying both i) the requests made in natural language by health professionals about the social-health state of a patient, and ii) the privacy rules indicated by the patient in exam. According to the proposed architectural approach, a patient may associate different security levels to different document parts by specifying for each document section a granular set of constraints to satisfy for accessing a document, such as the roles of the health professionals enabled, the reasons for which health professionals can consult a document, and so on. This flexibility permits thus to increase the inclination to share the clinical information by the patient. In order to make healthcare professionals able to request only the information necessary to their needs, the architectural model provides enhanced semantic features based on ontological models for building accurate and well-suitable queries on documents.

This paper is organized as follows. Section 2 highlights background and related work. Section 3 presents the proposed architectural model for semantic information retrieval from PS documents. Section 4 illustrates some case studies. Finally, Section 5 concludes the paper with key findings and ideas for future work.

2 Background and Related Work

Information Retrieval (IR) is the process of obtaining information that fulfills a specified information need. In general, IR techniques allow the management, representation, storage, organization, and access for information contained in documents. These techniques typically use Information Extraction (IE) mechanisms [8], which are operations able to automatically extract structured information from unstructured and/or semi-structured machine-readable documents (like XML documents). IE is often used for the definition of search indexes. Then, Natural Language Processing (NLP) is the activity of analyzing and representing linguistic input according to grammatical rules and lexicon of the target. I n contrast to IR, the Principle Of Least Privilege (POLP) is an information security approach with the aim of limiting access to the minimal level that the requesting entity needs.

Many IR approaches have been proposed in literature. Yu-Kai Lin et al. in [9] present a system that allows the extraction and reorganization of the information through the time concept. Several approaches adopt IE techniques to obtain relevant clinical information from the web, in particular from web forums where patients describe their experiences in free text. For example, in [10] a pharmacovigilance system with the ability of obtaining information about the drugs and possible side effects is presented. In [11], the current state of the art about content-based multimedia information retrieval is illustrated, whereas in [12] a latent semantic indexing algorithm developed and applied on The Health Improvement Network (THIN) is described. This algorithm allows a high-performance system to retrieve uncodified information present in clinical documents. Some systems use lexicons for performing semantic matches between words and phrases [13], such as the Health Information Text Extraction (HITEx) tool [14] and the work presented in [15], where the Unified Medical Language System (UMLS) database is used for extracting concepts for the principal diagnosis. In [16], IE and Knowledge Discovery in Database (KDD) approaches are combined by integrating data mining and IE methods. The aim is to provide a new high-quality IE methodology and, at the same time, to improve the performance of the underlying extraction system. Many works have also the scope of solving problems about complexity and ambiguity of medical terminology, such as in [17], which describes a technique to represent medical records and queries by focusing only on medical concepts essential for the information need of a medical search task. An approach based on k-anonymization able to tackle the problem of protecting privacy preserving confidential information obtained from web sites and pages in the retrieval phase is presented in [18].

To the best of our knowledge, approaches for the integration of information extraction techniques with mechanisms for privacy protection without losing relevant patient information have not been investigated yet.

3 Architectural Model

This section presents the sematic-based architectural model approach proposed for the development of systems aiming at retrieving significant information from PS

documents according to the POLP security model. First, the overview of the architectural model is illustrated. Then, the section describes: i) the component for privacy management, as well as the ontological model for the definition and formalization of privacy policies, are described; ii) the component for managing queries, putting an emphasis on the ontological model for their representation; iii) the component for the extraction of minimum information from PS documents.

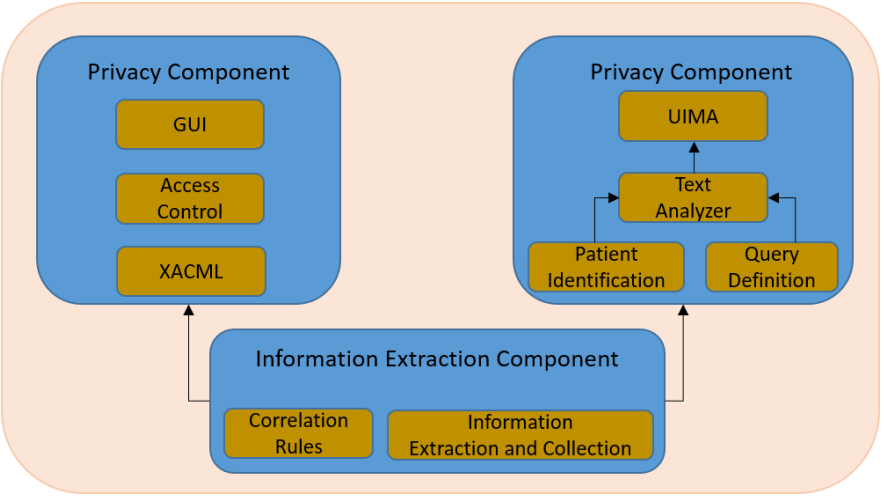


Fig. 1. The proposed architectural model for information retrieval from Patient Summary.

The proposed architectural model allows healthcare professionals to retrieve powerful social-health information from PS documents about patients by meeting in full the privacy policies defined directly by the patients. It combines information extraction techniques with a suitable access control mechanism. In this way, a healthcare professional can quickly access the whole list of patient problems or ask if the patient suffers or has suffered from a specific disease.

The architectural model, shown in Fig. 1, is organized into three modules. The first module is represented by the **Privacy Component (PC)**, which allows patients to define and represent the privacy policies. The second module is the **Query Component (QC)**, which examines the query of the healthcare professionals expressed in natural language and translates it in a "machine-readable" format through NLP techniques. The "machine-readable" query obtained is provided in input to the third module, the **Information Extraction Component (IEC)**, with the aim of permitting the extraction of authorized information of interest from a PS document. The architecture designed is based on a modular approach, so that it can be easily deployed in a transparent way also on Cloud environments.

The **Privacy Component** allows patients to define and manage privacy policies related to clinical documents, providing easy management of the rules policies. In this way, the patient is able to define the access policies in a detailed manner, indicating which information nugget of his/her clinical documents she/he intends to share and which not. For example, a patient could specify, for the Allergy section of a PS:

i) the professional roles that will be able to access and the professional roles that will not be able to access;
ii) if a specific professional role can access the entire section or only a piece of information in the section.

The component is described by a model, named Privacy Ontology Model (POM), which is defined starting from a subset of the access control model described in [19], opportunely extended with additional features. The POM has been formalized by an ontology (that is, an explicit formalized shared specification of a conceptualization [20]). All the concepts and properties have been represented, where the properties are expressed by relationships among the concepts. The ontology defined allows reasoning about knowledge and permits to create a common vocabulary for the information exchange without ambiguity. Fig. 2 shows the UML static structure of the POM expressed through the ontology technology, whereas the formalized definition is provided below.

Definition of the Privacy Model. The POM can be formally described as a tuple POM = (U, R, RO, PU, P, PRU, IR).

The concepts of the ontology are described below**:**

- **User (U):** a set of users U:={u_1, ..u_n}. In the POM, a user is a resource requester (e.g. healthcare professional). It consists of *name*, *surname*, *userid*, and *role*.
- **Role (R):** a set of professional roles R:={r_1,…r_m}. A professional role represents a user's job function. Privacy policies (*Privacy_Rule*) can refer to the *Role*. For this reason, an indirect link has been formalized between *Role* and *Permission* concepts to access a resource (document or nugget).
- **Resource_Object (RO):** a set of resource objects (clinical documents and nuggets) RO:={ro_1, …ro_j}. They represent the content to be protected and to be shared with users like healthcare professionals. The resources are characterized by a *resourceId* and a list of access *purposes*.
- **Purpose (PU):** a set of purposes PU:={pu_1,…,pu_s}. A purpose is the motivation for which a user accesses a given resource object. This concept is related both to the data request (it allows specifying the reasons for the access request) and to the resource object. The patient can choose the set of access purposes for his/her clinical documents. The purpose is characterized by a *name* and a *purposeId.*
- **Permission (P):** a set of permissions P⊆RO×PU. A permission represents an action on a resource object. Each P_i consists of: i) *resource*, which is the resource referenced by the permission); ii) *permissionId*, which is the identifier of the persmission; iii) *for_all,* which is a boolean value indicating if the request is able to access all the values of the object resource or only one of them.
- **Operation (O):** denotes an operation on a resource associated with a permission. It is identified by a *name*.
- **Privacy_Rule (PRU):** a set of privacy rules PRU⊆U×R×P. A PRU_i is characterized by the following properties: i) *is_able* is a property that indicates if the rule permits or denies the access to a resource object; ii) *role* is a property (relationship between **PRU** and **R**) that specifies the list of roles in the rule: this

parameter is optional, and, if it is not indicated, the *user* parameter successively defined is mandatory; iii) *user* is the property (relationship between **U** and **PRU**) that specifies the list of users in the rule: this parameter is optional, and, if it is not indicated, the *role* parameter is mandatory; iv) *permission* is a property (relationship between **PRU** and **P)** that indicates the permission associated with the defined rule. The POM allows a patient to manage his/her resources correctly by defining appropriate privacy rules by the use of the PRU concept.

- **Information_Request (IR):** a set of information requests IR⊆U×R×P×RO×PU. This concept represents the information request to a particular resource. When a user (e.g. a healthcare professional) wants to request for objects managed by the POM, it is necessary to instantiate this concept containing the following properties: i) *resourceId* represents the identifier of the request object; ii) *purpose* represents the reason for requesting the resource, for example *"Emergency"* or *"Treatment"* (or for example *"Cardiac Care"*); iii) *user* represents the requestor who is making the request for an object resource; iv) *role* denotes the professional role of the user; vi) *for_all* indicates that all the values of an object resource can be accessed.

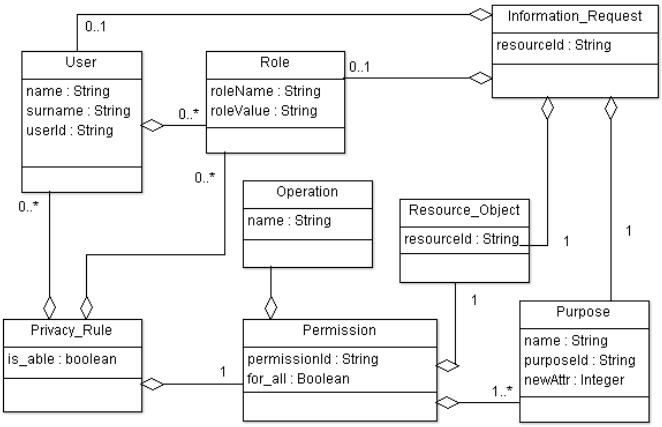


Fig. 2. Graphical representation of the Privacy Ontological Model.

The datatype properties are the following:

- **is_able:** this datatype property allows specifying if the security rule permits (value=*true*) or denies (value=*false*) the access to a resource.
- **for_all*:*** this datatype property allows specifying whether the request is able to access all the information present in a resource object (value=*true*) or is limitated only for a specific value of information. For example, in the first case (value=*true*) the request can ask the list of problems and the answer by the system will be all the list of patient problems, whereas in the second case (value=*false*) the request is able to ask if a patient has a specific problem and the answer of the system will be *yes* or *not value found*.

Privacy policies can be defined by a patient using a suitable GUI (similar to that shown in [19]). Then, the graphical representation of the access policies expressed by the patient can be translated into the standard eXtensible Access Control Markup Language (XACML) language.

The **Query Component** allows the definition and management of well-defined queries on the basis on a model named Query Ontology Model (QOM), starting from a query expressed in natural language.

For the aim of the architectural model proposed in this paper, three different kind of queries have been defined. Each query is composed by two elements: *kind* and *parameter*. A couple *kind-parameter* identifies a specific query and consequently a specific extracting information function for a PS document. The possible types of queries are described below:

1. **Patient Demographics Concept Query:** is related to the generic information of the patient, such as gender or age. In this case, *kind* is valued with *"Info"*, and the possible parameters are: i) *gender*, ii) *age*, and iii) *caregiver*.
2. **Problem Concept Query:** is related to diseases and diagnoses, such as chronic conditions. In this case, *kind* is valued with *"Problem"*, and the possible parameters are: i) *All* or ii) a specific pathology *"x"* (e.g. disease or diagnosis).
3. **Allergy Concept Query:** is related to allergies and intolerances, like allergies for drugs or foods. In this case, *kind* is valued with *"Allergy",* and the possible parameters are: i) *All* or ii) a specific allergen *"y"* (e.g. food or drug).

The parameter *All* indicates that the query requests all the section content. This is useful for example for the queries 2. and 3., which aim to request respectively all the allergies and all the chronic diseases of the patient. The parameters *"x"* or *"y"* may indicate a specific drug or a specific disease. The use of these parameters is useful for example to request if a patient is allergic (or intolerant) to a given drug or for example to ask if a patient suffers from a given disease. The QOM has been formalized by an ontology. The concepts and properties are represented in Fig. 3. The formalization of the POM and QOM allows the merging of the two models to consent appropriate information extraction. A formalized definition of the QOM is described below.

Definition of the Query Model. The QOM can be formally described as a tuple QOM = (TH, AL, PR, PDE, PS, D, FO, D).

The concepts of the ontology are described below:

- **Therapy (TH):** a set of therapies TH:={th1,…,thz}. Each TH_i indicates a drug therapy that was followed by the patient in a previuos period of time, which is associated with his/her PS. It consists of: i) *drug*, which is a LOINC code indicating a specific drug; ii) *start*, which indicates the start day of the therapy; iii) *end*, which indicates the end day of the therapy.
- **Allergy (AL):** a set of allergies AL:={al_1, …, al_u}. An AL_i indicates an allergen opportunely coded. If associated with a PS, it indicates that the patient has an allergic reaction to the allergen. It consists of: i) an *identifier*; ii) the boolean value *is_intolerance* denoting if the allergy is only an intolerance (in this case

value=*true*); iii) the *drug*, which is a code indicating a specific drug; and iv) the *food*, which is a code indicating a specific food.

- **Problem (PR):** a set of problems PR:={PR_1, …,PR_v}. A PR_i indicates a patient problem, for example a patient chronic disease. It consists of an *identifier*, in this case represented by a LOINC code. This concept is specialized in two different concepts: *Disease* and *Diagnosis*.
- **Patient_Demographics (PDE):** a set of patient demographics PDE:={PDE_1, …, PDE_t}. Each PDE_i indicates the demographic information about the patient. It consists of: i) *patientIdentifier*, which is the identifier of the patient; ii) *name* of the patient*; iii) surname* of the patient; iv) *caregiver,* which indicates the name and surname of the caregiver and her/his contact; v) *gender* of the patient; vi) *birthdate* of the patient.
- **Patient_Summary (PS):** a set of Patient Summary documents PS:={PS_1, …. PA_h}. A PS_i is a specific *Resource_Object* consisting of: i) *identifier* of the document; ii) *SectionProblem,* which is a list of patient problems (like *Disease* or *Diagnosis*); iii) *SectionPatientDemographic,* which is a list of demographic information about a patient; iv) *SectionAllergy,* which is a list of allergies or food and drug intolerances of the patient; v) *SectionTherapy,* which is a list of the patient drug therapies.
- **Drug (D):** a set of drugs D:={D_1, …,D_q}. A drug, represented through a code, can be associated with allergies or a clinical treatment of the patient. It consists of *identifierCode* and *name.*
- **Food (FO):** a set of foods FO:={FO_1, …, FO_f}. A food, represented by a code, can be associated with allergies. It consists of *identifierCode* and *name.*
- **Disease (D):** a set of diseases D:= {D_1, …, D_l}. A disease, represented by a code, is a specialization of the *Problem* concept consisting of the *name* of the disease and the value *chronic*, which is represented by a boolean value indicating if the disease is chronic.
- **Diagnosis (DI):** a set of diagnoses DI:= {DI_1, …, DI_g}. A diagnosis, represented by a code, is a specialization of the *Problem* concept consisting of the *name* of the diagnosis and its *date*, which indicates the start and the end date of the problem.

The only datatype property is the following:

- **is_intolerance:** this datatype property allows specifying if the *Allergy* concept refers to an intolerance or an allergy (*true* denotes an intolerance, *false* an allergy).

The QOM allows representing both the content of the PS and the query used by the system for the extraction of the information. The queries expressed by healthcare professionals can be classified into three different types. The classification phase consists of the following actions:

i) textual analysis of the query using the Unstructured Information Management Architecture (UIMA) OASIS standard and Apache framework for classifying the query;
ii) identification of the kind of query;
iii) identification of the parameters;
iv) transformation of the query expressed in natural language into a machine-readable query, characterized by the *kind* and *parameter* elements.

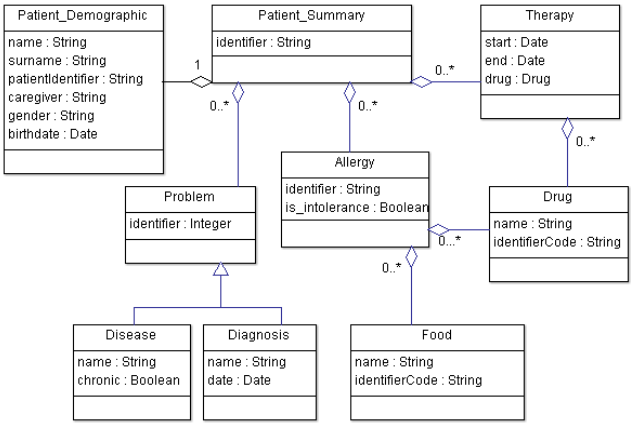


Fig. 3. Graphical representation of the Query Ontological Model.

The **Information Extraction Component** is responsible for retrieving the information from a PS document. This component uses the PC and QC components previously described. The QC allows identifying the clinical information of interest requested by the healthcare professional, whereas the PC determines if the information meets the privacy policies of the patient.

The IEC uses the QC to i) identify the patient subject of the query, and ii) detect the semantic query, composed by the *kind* and *parameter* elements. Once identified the semantic query, the extraction function and the resource object request are identified. The IEC evaluates if the healthcare professional requestor has the right to access the requested information, using the PC. The IEC uses a suitable control algorithm, which allows accessing or not accessing the required information based on privacy policies expressed in XACML language. The algorithm uses the information about the request provided by the *Information_Request* concept and the information associated with the *Permission* and *Resource_Object* concepts. If the requestor has the access rights, then the component uses the extraction function to retrieve the requested and authorized information, otherwise the component denies the access to the requestor. As the de-facto international standard used for structuring PS documents is HL7 CDA Rel. 2 based on the XML language, the extraction function has been defined by using the XPath query language [4]. Fig. 4 shows an example of the extraction function in XPath for the query *"Allergy All"*.

```
DocumentBuilderFactory domFactory = DocumentBuilderFactory.newInstance();
domFactory.setNamespaceAware(true);
DocumentBuilder builder = null;
Node res = null;
XPath xpath = null;
XPathExpression expr = null;
NodeList nodes = null;
builder = domFactory.newDocumentBuilder();
xpath = XPathFactory.newInstance().newXPath();
BodyNamespaceContext namespaceContext = new BodyNamespaceContext();
xpath.setNamespaceContext(namespaceContext);
expr = xpath.compile(
"ClinicalDocument/component/structuredBody/component/section/
    code[@code='48765-2']/parent::*");
res = (Node) expr.evaluate(body, XPathConstants.NODE);
nodes = (NodeList) res;
```

Fig. 4. XPath example query for Information Extraction from Patient Summary.

4 Case Studies

A prototypal system of the proposed architectural model has been implemented in Java language. Two case studies have been considered to show how the proposed architectural approach works for satisfying a request for social-health information. Let us suppose that the healthcare professional *Jon Cole*, with professional role "dentist", needs to know if her patient with identifier *RSSMRA59A01D086E* has an allergy to a particular anaesthetic. Jon in the first case study requests the list of all the allergies of the patient (that is, the entire Allergy section). In the second case study, Jon queries the system in order to know if the patient could have an allergic reaction to a single anaesthetic. Let us suppose that the privacy context expressed by the patient is the one summarized in Table 1.

Table 1. Privacy context in the case studies according to the Privacy Ontological Model.

User	*name*	Jon
	surname	Cole
	userId	user1z1d
	role	role0jj1
Role	*roleName*	role0jj1
	roleValue	Dentist
Resource_Object	*resourceId*	secdoc002 (that is the Allergy section)
	purposes	{Treatment; Emergency}
Permission	*permissionId*	per001
	resource	secdoc002
	for_all	false
Privacy_Rule	*user*	-
	permission	per0001
	is_able	true
	role	role0jj1, roleojjh7

Case Study 1. As shown on the left of Fig. 5, Jon submits the following query to the system: *The patient identifier is RSSMRA59A01D086E. What are the patient's allergies?* The QC allows identifying both the patient and the semantic query *Allergy All* associated with an extraction function and a specific *Resource_Object*. In this case, *Resource_Object* is related to the Allergy section (resource with identifier *secdoc002*) of the patient with identifier RSSMRA59A01D086E. The PC does not allow the access to the section because the *Permission* associated with the dentist role has the attribute *for_all* with a *false* value. If this attribute was *true*, then the system would extract the information and return all the list of patient's allergies.

Case Study 2. As depicted on the right of Fig. 5, Jon submits the following query to the system: *Is the patient RSSMRA59A01D086E allergic to procaine?* The QC allows identifying both the patient and the semantic query *Allergy procaine* associated with an extraction function and a specific *Resource_Object*, which is related to the Allergy section (resource with identifier *secdoc002*) of the patient with identifier RSSMRA59A01D086E. The PC allows accessing the section because the *Permission* associated with the dentist role permits to access a single information (only a specific

allergen) in the Allergy section. The information is extracted and outputted. The patient is allergic to procaine, so the output is: *Yes*, *the patient is allergic to procaine*.

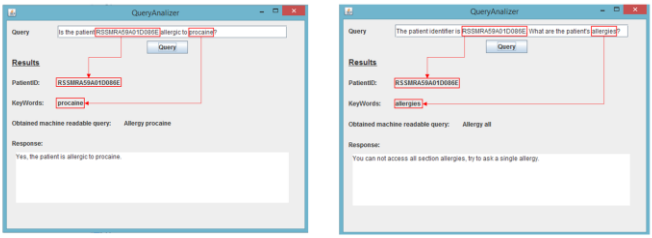


Fig. 5. Results obtained for Case Study 1 (*on the left*) and for Case Study 2 (*on the right*).

5 Conclusion and Future Work

This paper described a novel architectural approach based on semantic-enhanced features for minimum information retrieval from Patient Summary documents in accordance with the privacy policies expressed by a patient. The proposed approach permits, starting from a query expressed in natural language formulated by a healthcare professional, to: i) identify a machine-readable query; ii) verify the patient privacy policies for the specific query; iii) extract the requested information, if possible, and build an output. As future work, it is planned to extend the proposed architectural approach in two ways. On one hand, we intend to consider other kinds of clinical documents (like clinical reports and discharge summaries) where retrieving information, adding the ability to obtain information also from patient's non-textual data, like medical images. On the other hand, it is scheduled the extend the architectural model in order to define appropriate semantic graphs to represent and organize the clinical information extrapolated from documents archived in a hospital information system, supporting thus physicians in both diagnosing diseases and prescribing medications.

References

1. McClellan, M., Rivlin, A.: Improving health while reducing cost growth, what is possible? Health Policy Issue Briefs. Number 1 of 7 (2014)
2. Pluye P., Grad R.M., Dunikowski L.G., Stephenson R.: Impact of clinical information-retrieval technology on physicians: A literature review of quantitative, qualitative and mixed methods studies. International Journal of Medical Informatics, Vol. 74(9) (2005) 745–768
3. epSOS project, online at: http://www.epsos.eu/ (Access date: 25 July 2016)
4. Dolin R.H., Alschuler L., Boyer S., Beebe C., Behlen F.M., Biron P.V., Shabo A.: HL7 Clinical Document Architecture, Release 2. Journal of the American Medical Informatics Association, Volume 13, Issue 1 (2006), 30-39

5. Ciampi M., De Pietro G., Esposito C., Sicuranza M., Mori P., Gebrehiwot A., Donzelli P.: On securing communications among federated health information systems, Vol. 7613 (2012) 235-246
6. General Data Protection Regulation, European Commission.: Regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data, online at: http://ec.europa.eu/justice/data-protection/document/review2012/com_2012_11_en.pdf (Access date: 25 July 2016)
7. Khan M.F.F., Sakamura K.: Context-aware access control for clinical information systems. International Conference on Innovations in Information Technology (2002) 123-128
8. Xu H., Stenner S.P., Doan S., Johnson K.B., Waitman L.R., Denny J.C.: MedEx: A medication information extraction system for clinical narratives, Journal of the American Medical Informatics Association (2013) 19–24
9. Lin Y.K., Chen H., Brown R.A.: MedTime: A temporal information extraction system for clinical narratives. Journal of Biomedical Informatics, ISSN 1532-0464, Vol. 46, Suppl. (2013) S20-S28
10. Liu X., Chen H.: Identifying adverse drug events from patient social media: A case study for diabetes. IEEE Intelligent Systems, Vol. 30, Issue 3 (2015) 44-51
11. Amato F., Greco L., Persia F., Poccia S.R., De Santo A.: Content-based multimedia retrieval. Data Management in Pervasive Systems (2015) 291-310
12. Al-Qahtani M., Amira A., Ramzan N.: "An efficient information retrieval technique for e-health systems", International Conference on Systems, Signals and Image Processing (2015) 257-260
13. Patrick J., Wang Y., Budd P., Recter A., Brandt S., Rogers J., Herkes R., Ryan A., Vazirnezhed B.: "Developing SNOMED CT subsets from clinical notes for intensive care service", Health Care and Informatics Review Online, Vol. 12(3) (2008) 25-30
14. Zeng Q.T., Goryachev S., Weiss S., Sordo M., Murphy S.N., Lazarus E.: "Extracting principal diagnosis, co-morbidity and smoking status for asthma research: Evaluation of a natural language processing system", Medical Informatics and Decision Making (2006)
15. Demner-Fushman D., Chapman W.W., McDonald C.J.: "What natural language processing do for clinical decision support?", Journal of Biomedical Informatics (2009) 760-772
16. Feilmayr C.: Text mining-supported information extraction: An extended methodology for developing information extraction systems, 22nd International Workshop on Database and Expert Systems Applications, IEEE CS, (2011) 217-221
17. Limsopatham N., Macdonald C., Ounis I.: A task-specific query and document representation for medical records search. Advances in Information Retrieval, Lecture Notes in Computer Science, Vol. 7814 (2013) 747-751
18. Abril D., Navarro-Arribas G., Torra V.: Towards privacy preserving information retrieval through semantic microaggregation. IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (2010) 296 – 299
19. Sicuranza M., Esposito A., Ciampi M.: An access control model to minimize the data exchange in the information retrieval. Journal of Ambient Intelligence and Humanized Computing (2015) 741-752
20. Gigante G., Gargiulo F., Ficco M.: A semantic driven approach for requirements verification. Intelligent Distributed Computing VIII Studies in Computational Intelligence, Vol. 570 (2015) 427-436

Adaptive Modular Mapping to Reduce Shared Memory Bank Conflicts on GPUs

Innocenzo Mungiello, Francesco De Rosa

Abstract This paper presents the experimental evaluation of a new data mapping technique for the GPU shared memory, called Adaptive Modular Mapping (AMM). The evaluated technique aims to remap data across the shared memory physical banks, so as to increase parallel accesses, resulting in appreciable gains in terms of performance. Unless previous techniques described in literature, AMM does not increase shared memory size as a side effect of the conflict-avoidance technique. The paper also presents the experimental set-up used for the validation of the proposed memory mapping methodology.

1 Introduction

Exascale scientific and high-performance computing (HPC) are considered essential in many areas [25] and are today embracing a wealth of different applications [4, 3, 2]. During the last years, HPC architectures have been increasingly moving to heterogeneous systems. In fact, it is now common to integrate in the same HPC system manycore CPUs, accelerators like Graphics Processing Units (GPUs), and even Field-Programmable Gate Arrays (FPGAs) devices, which were traditionally used only for implementing highly specialized circuit-level solutions [6, 7, 11, 5, 10]. In this context, GPUs are gaining importance and are now being used for a variety of domains and applications [21]. FPGAs also provide a solution for building customized computing platforms, although they pose non-trivial challenges, mostly involving complexity, high-level design as well as suitable programming models, which are somewhat mitigated by GPUs [31, 16]. In current

Innocenzo Mungiello
University of Naples Federico II and Centro Regionale ICT (CeRICT), Naples, Italy, e-mail: innocenzo.mungiello@unina.it

Francesco De Rosa
University of Naples Federico II, Naples, Italy

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_34

computing technologies, a major component of the power consumption and performance degradation is due to data movement rather than mere processing, which is indeed a widely studied subject in a number of different contexts [26]. This work thus addresses the impact on performance of memory optimization techniques for heterogeneous architectures, particularly GPUs. The emphasis is on the multi-bank organization of the on-chip GPU shared memory, where the whole shared addressing space is partitioned in a cyclic way and is potentially subject to an access conflict problem [14]. In particular, the paper focuses on the experimental evaluation of a new data mapping approach called Adaptive Modular Mapping (AMM) which differs from other works [19, 28, 23, 20] in that it does not involve an increase of the shared memory size, which in some cases can decrease performance as a side effect. The evaluation of the AMM technique highlights a significant gain in terms of performance, as confirmed by an experimental set-up used for performing measures on a physical GPU-based platform.

The paper is structured as follows. Section 2 discusses the background and the motivation of this work. Section 3 recapitulates the approach adopted for data partitioning. Section 4 shows the set-up used for the experimental evaluation. Section 5 concludes the paper with some final remarks.

2 Background

Memory access has traditionally been an important optimization problem for parallel systems [13, 15], and in many classes of systems it may significantly impact perfomance, along with the interconnection subsystem [17, 18, 12]. In particular, memory access *latency* is often an important factor determining the overall system performance. Many solutions have been proposed to face this problem on parallel architectures, e.g. the design of a hierarchical memory system, the introduction of shared scratchpad memories, and the increase of thread-level parallelism to *hide* memory latency.

Most of these features are present on Graphical Processor Units (GPUs) that represent today a major shift in high-performance computing architectures focusing on increasing the execution throughput of parallel applications. In fact, on modern GPU architectures like the NVIDIA Kepler or the newest Maxwell, the memory system is designed in a hierarchical fashion, as shown in Figure 1.

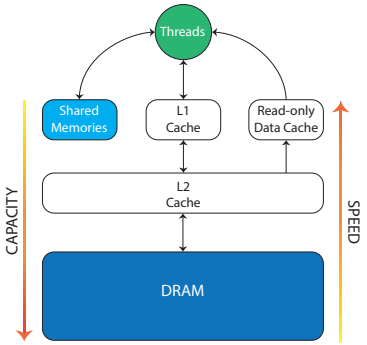


Fig. 1 GPU NVIDIA Kepler Memory Hierarchy.

In this way, parallel processing elements access simultaneously several independent memory banks through complex interconnects. This potentially provides an opportunity for improving the memory bandwidth available to the application. In fact, even if the DRAM latency is too high, it can be hidden by running thousands of threads.

Another way to improve system performance is to use *scratchpad* shared memories. For performance purposes, the on-chip shared memory of a GPU device is normally divided in 32 banks, which can be accessed in a parallel way from all the threads in a *warp*. Data allocated to the shared memory are cyclically distributed over the banks.

Figure 2 shows an example of data mapping on shared memory. In this example the data are 4 byte-word index and, for simplicity, we show only four banks.

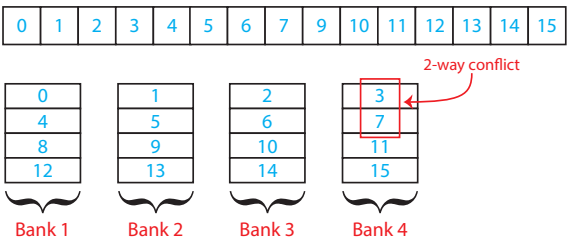


Fig. 2 Shared Memory Data Mapping. The figure shows a possible 2-way bank conflict if a thread would access Bank 4 and fetch value ”3” and another thread would access Bank 4 and fetch value ”7”.

The access mode can affect the performance of a kernel. In fact, allowing all the threads in a warp to fetch data in parallel from this multi-bank memory can lead to great performance improvements, but it requires explicit, bank-aware organization of the data layout. In particular, in the case where two or more threads in the *same* warp try to access *different words* stored in the same bank, the interconnection network is no more able to provide the required data to all the threads in parallel. This situation, called *bank conflict*, is a major problem related to the use of the on-chip shared memory. In particular, if two threads try to access different words stored in the same bank, we have a 2-way bank conflict, as shown in Figure 2. If three threads try to access different words stored in the same bank, a 3-way bank conflict occurs,

and so on. The worst case involves all 32 threads in a warp trying to access different words stored in the same bank, causing a 32-way bank conflict. Whenever a conflict occurs, it is resolved by serializing the accesses in time. As an example, the serialization in a 2-way scenario leads to double latency, decreasing the performance considerably and increasing power consumption, especially if the kernel uses shared memory intensively [27].

2.1 Related Works

Several techniques are presented to solve bank conflicts and reduce memory access latency [8, 29]. The simplest one is *Memory Padding*, presented by NVIDIA in [1, 24, 9]. This technique solves bank conflicts in many cases by simply using an extra empty column of shared memory. While effective and simple, this technique has the disadvantage of wasting shared memory and this can cause problems in certain situations.

A.H. Khan et al. [22] analyze the Matrix Transpose problem and provided a solution very close to AMM techinque for that particular kernel. S. Gao et al. [19] present a framework for automatic bank conflict analysis and optimization. Kim et al. [23] present CuMAPz, a tool to compare the memory performance of a CUDA program and help programmers to optimize them. Sung et al. [28] propose DL, a practical GPU data layout transformation system that solves both problems. Grun et al. [20] present APEX, an approach that extracts, analyzes, and clusters the most active access patterns in the application, and aggressively customizes the memory architecture to match the needs of the application. Z. Wang et al. [30] use a machine-learning approach to transform data layout and improve performance.

3 Data Layout Transformation

Heterogeneous platforms provide some form of customizability that can be exploited to improve performance. GPUs offer a certain degree of freedom enabling the application developer to tailor the bank mapping on the application access pattern. Some applications need access patterns yielding bank conflicts. A naive pattern is shown in Figure 3 where 4 threads access a 4-bank memory in lock-step in such a way that a 25% efficiency is obtained. From the developer point of view a 4x4 matrix accessed with a 4-location strided pattern leads to this behaviour. The simplest way to correct such an access pattern is the Memory Padding technique described in Section 2.1. The resulting bank mapping will be as shown in Figure 3 where, for example, each green cell indicates a concurrent access. In this case the padding technique provides a conflict-free access pattern which improves memory efficiency but also causes a wasted memory problem, as pointed out by the 4 unused red cells, which could under-perform the naive one. In particular, the padding technique can

lead to a performance problem in scenarios exhibiting the scratchpad shared memory as the *limiting factor*, because the number of threads simultaneously eligible to work depends on the memory size. Nvidia CUDA exposes a thread-group based programming model where the whole computation will be completed by several groups of threads simultaneously running. The thread groups are scheduled on the basis of the available resources. Since the padding technique causes more memory to be allocated, a smaller number of threads can be simultaneously eligible to run. These conditions cause decreasing performance although some or all of the bank conflicts are solved. The proposed technique prevents wasting memory by allocating minimal memory, as needed to satisfy the maximum number of threads. The resulting mapping scheme outperforms the padding and the uncorrected ones under the following condition:

cond 1 The scratchpad shared memory is a limiting factor;
cond 2 Memory Padding leads to a smaller number of eligible threads to run simultaneously;

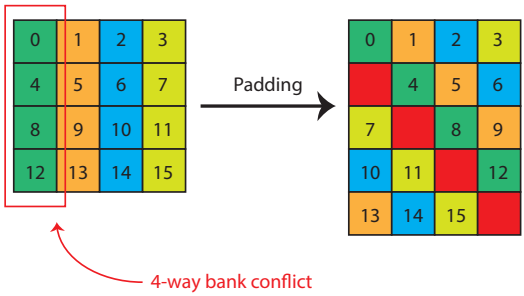


Fig. 3 Padding application. In this case the technique solves all conflicts, but there are empty memory cells highlighted in red.

3.1 Adaptive Modular Mapping

Under the above described conditions, the proposed technique performs a modular remapping of shared memory accesses based on the matrix width, which solves some or all bank conflicts and avoids increased memory size unlike padding techniques. The proposed technique involves the resolution of a linear programming model (Model 4 below) with the aim of extracting new mapping schemes avoiding bank conflicts.

$$\sum_{i=1}^{N_IT} x_{ijk} = 1 \quad \forall j \in \{1..N_TH\} \quad \wedge \quad \forall k \in \{1..N_B\}$$

$$\sum_{j=1}^{N_TH} x_{ijk} = 1 \quad \forall i \in \{1..N_IT\} \quad \wedge \quad \forall k \in \{1..N_B\}$$

$$\sum_{k=1}^{N_B} x_{ijk} = 1 \quad \forall i \in \{1..N_IT\} \quad \wedge \quad \forall j \in \{1..N_TH\}$$

$$\sum_{j=1}^{N_TH} \sum_{k=1}^{N_B} x_{ijk} = N_TH \quad \forall i \in \{1..N_IT\}$$

$$minimize \left(z = \sum_{j=1}^{N_TH} \sum_{k=1}^{N_B} \sum_{k=1}^{N_IT} x_{ijk} \right)$$

Model 4 N_IT stands for number of iterations, N_TH stands for number of threads and N_B stands for number of banks. x_{ijk} means thread j accesses bank k at i-iteration. This is an NP-Hard problem and the simplest feasible solution is selected.

The results of the Adaptive Modular Mapping technique is shown in Figure 5. Suppose a N-bank memory system and a total of M threads with a M-way conflict.

```
__shared__ int shmem[M][N];

int index = threadIdx.x;

for (int i = 0; i < N; i++)
    shmem[index][i]=some_value;
```

The Adaptive Modular Mapping is highlighted in the following code.

```
__shared__ int shmem[M][N];

int index = threadIdx.x;

for (int i = 0; i < N; i++)
    shmem[index][(index+i)%N]=some_value;
```

where:

- % stands for the modulo reduction operator;
- shmem with the CUDA `__shared__` keyword declares a memory shared among all the threads of a block.

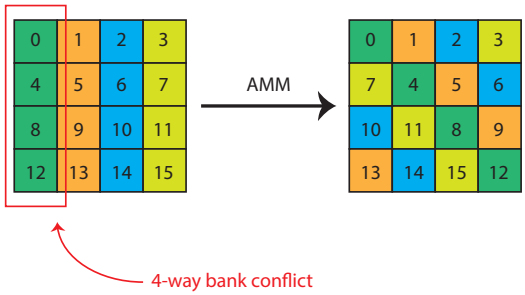


Fig. 5 Adaptive Modular Mapping application. In this case the technique solves all conflicts

As shown above this technique requires a very little effort from the developer point of view, while leading to minor memory requirements and consequently better exploiting parallel computing resources. In particular, the % operator can be realized efficiently by an *AND* operation when dexter is a power-of-two operand.

4 Experimental Evaluation

This section presents the set-up used to carry out the experimental evaluation of the above optimization technique and collect performance data from a physical platform. A suite of kernels has been properly selected to set a scenario where conditions **cond 1** and **cond 2** described in Section 3 are met.

4.1 System Overview

The system includes a host PC running a WMware Virtual Machine with Ubuntu 14.04, with 4 cores, 8 GB of RAM ,and the JetPack 2.1 installed on it. In this environment, NVIDIA Nsight Eclipse Edition is used to write CUDA code and, then, to compile and remotely run it on a Jetson TK1 development board. Memory and Processor frequency of the Jetson TK1 are fixed to 792MHz and 804MHz respectively in order to avoid measurement errors.

4.2 Results

In this section experimental results are discussed. By using the system described in Section 4.1 it was possible to evaluate the impact of the AMM technique on the performance of the code. Five kernels are used to test the AMM technique. They are summarized in Table 1.

Kernel	Suite	Description
Transpose	CUDA SDK	All accesses in global memory are coalesced, there are bank conflicts, and arithmetic operations are low.
Convolution Col	CUDA SDK	All accesses in global memory are coalesced, there are bank conflicts, and arithmetic operations are low.
Lud_Perimeter	RODINIA	All accesses in global memory are coalesced, there are bank conflicts, and arithmetic operations are low.
Foo 1	CUSTOM	Ad-hoc implemented kernel to reproduce coalesced and uncoalesced global memory access pattern on user choice.
Foo 2	CUSTOM	Ad-hoc implemented kernel to reproduce intensive arithmetic operations.

Table 1 Kernels used to test the AMM technique

According to the different characteristics of the kernels, the results can be divided in three classes listed below:

1. All global memory accesses are coalesced, so the memory latency to hide is low. In this scenario the *padding* technique increases the naive kernels performance despite fewer threads are simultaneously running. The AMM technique outperforms both, as shown in Figure 6.

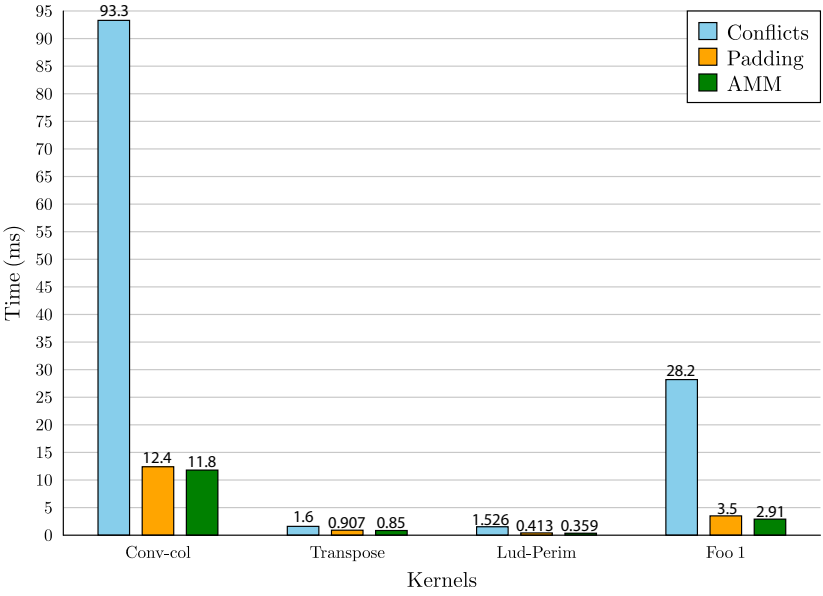


Fig. 6 Case 1 Results. The AMM technique decreases the execution time of all kernels.

2. Global memory accesses are un-coalesced, so the memory latency to hide is high. In this scenario the *padding* technique decreases the naive kernel perfor-

mance because fewer threads are simultaneously running. The AMM technique outperforms both, as shown in Figure 7.

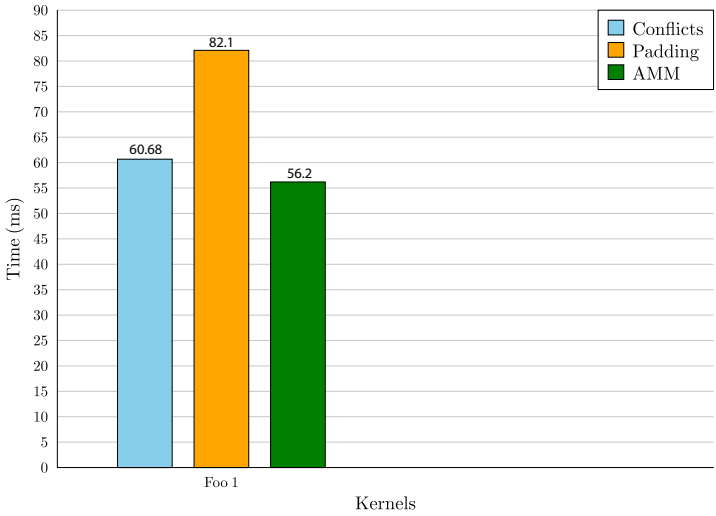


Fig. 7 Case 2 Results. The AMM technique decreases the execution time of the kernel.

3. All global memory accesses are coalesced, so the memory latency to hide is low, but there is a high number of arithmetic instructions. In this scenario, the *padding* technique decreases the naive kernel performance because fewer threads are simultaneously running. The AMM technique in this case equals the naive kernel performance as shown in Figure 8

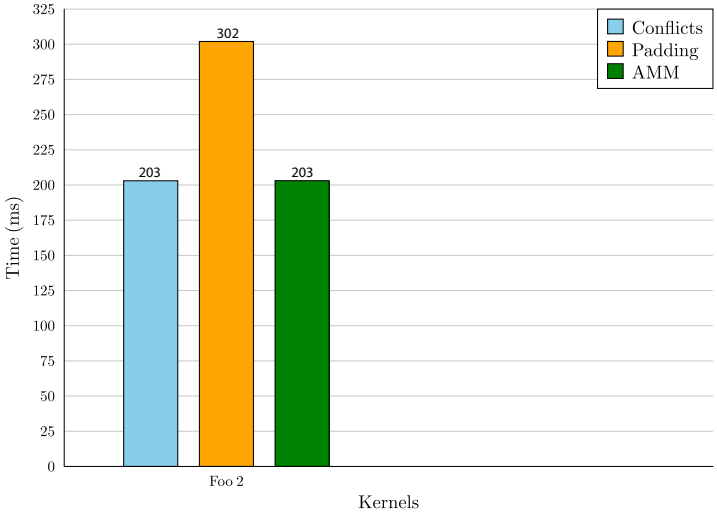


Fig. 8 Case 3 Results. In a compute intensive kernel, AMM technique equals naive performance.

5 Conclusions

GPUs have become extremely important for today HPC as they provide an effective answer for many increasingly demanding applications. This work proposed an optimization technique for GPU on-chip memory. The paper presented both the formal model underpinning the technique and the experimental evaluation on a number of kernels. The results pointed out the significant incidence that such techniques may have on execution time.

Acknowledgments. This work is supported by the European Commission in the framework of the H2020-FETHPC-2014 project n. 671668 - MANGO: exploring Manycore Architectures for Next-GeneratiOn HPC systems.

References

1. CUDA C Programming Guide
2. Amato, F., Fasolino, A., Mazzeo, A., Moscato, V., Picariello, A., Romano, S., Tramontana, P.: Ensuring semantic interoperability for e-health applications. In: Proceedings of the International Conference on Complex, Intelligent and Software Intensive Systems, CISIS 2011, pp. 315–320 (2011)
3. Amato, F., Mazzeo, A., Penta, A., Picariello, A.: Building RDF ontologies from semi-structured legal documents. pp. 997–1002 (2008)
4. Amato, F., Moscato, F.: A model driven approach to data privacy verification in e-health systems. Transactions on Data Privacy **8**(3), 273–296 (2015)

5. Barbareschi, M.: Implementing hardware decision tree prediction: a scalable approach. In: 2016 30th International Conference on Advanced Information Networking and Applications Workshops (WAINA), pp. 87–92. IEEE (2016)
6. Barbareschi, M., Battista, E., Mazzocca, N., Venkatesan, S.: A hardware accelerator for data classification within the sensing infrastructure. In: Information Reuse and Integration (IRI), 2014 IEEE 15th International Conference on, pp. 400–405. IEEE (2014)
7. Barbareschi, M., De Benedictis, A., Mazzeo, A., Vespoli, A.: Providing mobile traffic analysis as-a-service: Design of a service-based infrastructure to offer high-accuracy traffic classifiers based on hardware accelerators. Journal of Digital Information Management **13**(4), 257 (2015)
8. Che, S., Sheaffer, J.W., Skadron, K.: Dymaxion: optimizing memory access patterns for heterogeneous systems. In: Proceedings of 2011 international conference for high performance computing, networking, storage and analysis, p. 13. ACM (2011)
9. Cheng, J., Grossman, M., McKercher, T.: Professional Cuda C Programming. John Wiley & Sons (2014)
10. Cilardo, A.: Efficient bit-parallel $GF(2^m)$ multiplier for a large class of irreducible pentanomials. IEEE Transactions on Computers **58**(7), 1001–1008 (2009)
11. Cilardo, A.: Exploring the potential of threshold logic for cryptography-related operations. IEEE Transactions on Computers **60**(4), 452–462 (2011)
12. Cilardo, A., Fusella, E., Gallo, L., Mazzeo, A.: Exploiting concurrency for the automated synthesis of MPSoC interconnects. ACM Transactions on Embedded Computing Systems **14**(3) (2015)
13. Cilardo, A., Gallo, L.: Improving multibank memory access parallelism with lattice-based partitioning. ACM Transactions on Architecture and Code Optimization **11**(4) (2014)
14. Darte, A., Dion, M., Robert, Y.: A characterization of one-to-one modular mappings. Parallel Processing Letters **6**(01), 145–157 (1996)
15. Darte, A., Schreiber, R., Villard, G.: Lattice-based memory allocation. IEEE Transactions on Computers **54**(10), 1242–1257 (2005)
16. Escobar, F.A., Chang, X., Valderrama, C.: Suitability analysis of fpgas for heterogeneous platforms in hpc. IEEE Transactions on Parallel and Distributed Systems **27**(2), 600–612 (2016)
17. Fusella, E., Cilardo, A.: H^2ONoC: A hybrid optical-electronic NoC based on hybrid topology. IEEE Transactions on Very Large Scale Integration (VLSI) Systems (2016)
18. Fusella, E., Cilardo, A.: Minimizing power loss in optical networks-on-chip through application-specific mapping. Microprocessors and Microsystems (2016)
19. Gao, S., Peterson, G.D.: Optimizing cuda shared memory usage
20. Grun, P., Dutt, N., Nicolau, A.: Apex: access pattern based memory architecture exploration. In: Proceedings of the 14th international symposium on Systems synthesis, pp. 25–32. ACM (2001)
21. Hallmans, D., Åsberg, M., Nolte, T.: Towards using the graphics processing unit (gpu) for embedded systems. In: Proceedings of 2012 IEEE 17th International Conference on Emerging Technologies & Factory Automation (ETFA 2012), pp. 1–4. IEEE (2012)
22. Khan, A., Al-Mouhamed, M., Fatayar, A., Almousa, A., Baqais, A., Assayony, M.: Padding free bank conflict resolution for cuda-based matrix transpose algorithm. In: Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 2014 15th IEEE/ACIS International Conference on, pp. 1–6. IEEE (2014)
23. Kim, Y., Shrivastava, A.: Cumapz: a tool to analyze memory access patterns in cuda. In: Proceedings of the 48th Design Automation Conference, pp. 128–133. ACM (2011)
24. Kirk, D.B., Wen-mei, W.H.: Programming massively parallel processors: a hands-on approach. Newnes (2012)
25. Luebke, D.: Cuda: Scalable parallel programming for high-performance scientific computing. In: 2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, pp. 836–838. IEEE (2008)
26. Lustig, D., Martonosi, M.: Reducing gpu offload latency via fine-grained cpu-gpu synchronization. In: HPCA, vol. 13, pp. 354–365 (2013)

27. Mungiello, I.: Experimental evaluation of memory optimizations on an embedded gpu platform. In: 2015 10th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC), pp. 169–174. IEEE (2015)
28. Sung, I.J., Liu, G.D., Hwu, W.M.W.: Dl: A data layout transformation system for heterogeneous computing. In: Innovative Parallel Computing (InPar), 2012, pp. 1–11. IEEE (2012)
29. Ueng, S.Z., Lathara, M., Baghsorkhi, S.S., Wen-mei, W.H.: Cuda-lite: Reducing gpu programming complexity. In: International Workshop on Languages and Compilers for Parallel Computing, pp. 1–15. Springer (2008)
30. Wang, Z., Grewe, D., Oboyle, M.F.: Automatic and portable mapping of data parallel programs to opencl for gpu-based heterogeneous systems. ACM Transactions on Architecture and Code Optimization (TACO) **11**(4), 42 (2015)
31. Zhang, Z., Fan, Y., Jiang, W., Han, G., Yang, C., Cong, J.: High-level synthesis: From algorithm to digital circuit (2008)

A Review on Data Cleaning Technology for RFID Network

He XU[1,2,a], Jie DING[1,2,b], Peng LI[1,2,c], Wei LI[1,d]
[1] School of Computer Science & Technology/School of software, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
[2] Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks, Nanjing, 210003, China
{[a]xuhe,[c]lipeng}@njupt.edu.cn,{ 1248395233[b], 577935562[d]}@qq.com

Abstract. Radio Frequency Identification (RFID) technology is a kind of automatic recognition of communication technology. With the expansion of the RFID technology application areas, it needs to constantly improve the reliability, correctness and completeness of the RFID stream data. So data cleaning is significant to the RFID system. In recent years, many experts and scholars proposed a large number of cleaning algorithms according to RFID data flow characteristics. This paper gives a review on RFID data cleaning technology and some typical data cleaning algorithms.

1 Introduction

Radio Frequency Identification (RFID) technology is a kind of automatic recognition of communication technology[1]. With the detection, identification and monitoring of the electromagnetic signals, the related data can be read and written, so that information transmission is realized without contact between recognition system and specific target. Therefore, the target can be recognized automatically. RFID technology have the features of reading over a long distance, high storage capacity and other characteristics, which is widely used in the Internet of Things in the supply chain for object tracking and tracing, and have high-profile application prospect. For example, by integrating the RFID and wireless sensors, it is possible to communication together between RFID tag and other equipment. RFID technology not only can help a company to improve the efficiency of information management significantly, also can interconnect among enterprises, sales and manufacturing enterprises, so as to obtain more accurate control and feedback, and finally realize the optimization of the whole supply chain.

With the expansion of the RFID technology application areas, the demand for reliability of business data is increasingly important. In order to reach the level of satisfying the needs of the upper application, data cleaning is essential and directly affects the correctness and completeness of the business data, so it needs to filter and handle RFID data[2].

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_35

2 The characteristics of RFID data flow

The RFID data flow has the following characteristics[3]:

Streaming: each tag always continuously produces data in great quantities, which is continuous and intensive captured by the reader in a period.

Batch: multiple tag data is always captured by certain or more readers at the same time.

Semantic: tag data represents the position and status information of the observed object in a certain observation time.

Unreliability: RFID data is obtained by RFID reader and a variety of interactive mode's electronic tag are not reliable. The types and characteristics of wrong data are not identical, and the external causes(such as environmental factors) resulting in errors are different. All those factors cause the unreliable RFID data. There are some common error types of unreliability: False Negative, False Positive and Redundancy.

False Negative: radio frequency signal is highly vulnerable to environmental disturbance. Particularly, when a large number of tags suddenly come into the read range of reader. The collision and interference of signal among RFID readers will cause miss reading some tags, and this missing read phenomenon is very common. Reader cannot read all tags data within the scope of its read range and write without error.

False Positive: reader catches the tags outside the scope of its read range and write accidentally.

Redundancy: when a reader stay in the same place and can read within a period, the tags continue to send information to the reader, and it will produce a large number of repetitive data, where these extra data lead to some problems such as low efficiency of communication, energy loss and time delay. Generally, the redundancy is divided into tag redundancy and reader redundancy.

3 RFID data cleaning process

In order to improve the reliability, accuracy and integrity of RFID data, the RFID system needs self-contained cleaning module to handle huge amounts of tag data, to make reader read tag data consistent with the original data as far as possible. Usually, the error data types in RFID system are False Negative, False Positive and data redundancy. The general RFID cleaning framework is shown in Figure 1:

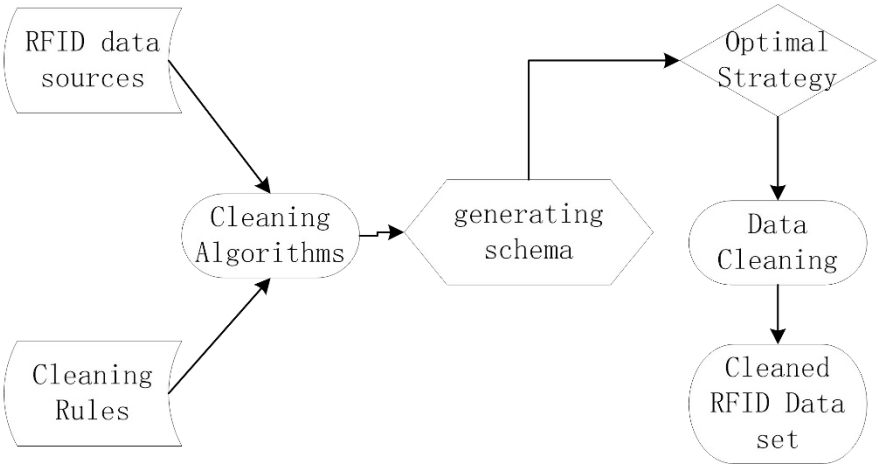


Figure 1. Generic framework of RFID data cleaning

The basic steps of data cleaning is based on the analysis of the causes of data error, process and form, then classification, using specific cleaning strategy for specific error to achieve optimized data quality and meet the requirements of the upper application. Through analysis, extract and optimize the data, it can maximize the implementation of data cleaning and make the data suitable for the upper application as much as possible.

4 Common cleaning algorithms of RFID data

4.1 The data cleaning algorithm based on fixed time window

In the original data cleaning system, the mechanism of using a slide time window technology to smooth filtering data flow is put forward by Y.Bai[4] from UCLA University and used for washing original tag data flow of RFID.

In most of the cleaning model of the RFID data, standard thought of data cleaning is equivalent to "smoothing filter based on time". The concept of "smooth window" is that refers to a window which has a certain length and sliding over time. In the whole life cycle of the system implementation, a fixed value represents the size of the time window which will continue to move forward as time goes on.

This method is simple and fast, but it don't adapt to the requirements of the dynamic environment. The window size is very difficult to set, if too small which can cause missing read (False Negative), too large which can lead to repetitive read (False Positive), thus result in the problem of data flow delay. So the ideal size of the sliding window must be set to ensure tags can be read completely and tags can be captured dynamically. Generally, although this method has shortcomings, it still has strong practicability because of its simple implementation.

4.2 Online extensible data flow cleaning framework

R.Jeffrey et al.[5] contrapose the characteristics of RFID data flow to introduce a data cleaning model based on the structure of pipeline's data cleaning model ESP through

introducing the definition of space and time granularity, which is also known as extensible data flow cleaning model. According to the characteristics of the original data using different cleaning steps, based on the time and space characteristics of RFID data, ESP uses different cleaning steps and makes them form a pipe to process the raw data. ESP can be used by all kinds of practical applications through a simple configuration. This algorithm is suitable for data leaking and repeated reading.

ESP can clean a variety of RFID data effectively, from a single reader to multiple readers, from a single space to multiple space, from a single type to multiple types. It has extensive data cleaning scope. Moreover, this model is very flexible, and can flexible select processing phase according to the different requirements. But it is very difficult to set the ideal time and space granularity parameter in ESP. Only to pick up the optimal particle size can ensure the quality of cleaning and minimize the average error rate of data cleaning. Thus, what is the most needed to deal with is to adjust self-adaptive time and space granularity.

4.3 RFID data cleaning algorithm of adaptive window size

M.Garofalakis et al.[6] has introduced the data cleaning strategy based on time correlation, which is based on probability model. Through dynamic changing the window's size, it is mainly used to solve the problem of leaking of read data. This is the first definition of self-adaptive smooth filtering of RFID data cleaning method - "SMURF".

This method puts RFID data flow as a random event in probability statistics, by using the method of probability theory to fill the data that had read missed. It has the advantages of adaptive deciding window's size according to the size of tag reading rate, and reducing the false negative and false positive which caused by unreasonable selection of window size. But for dynamic tag data, for example, when moving tags leave the reader's reading range quickly, it can make reading rate suddenly decreased. Meanwhile, the method of SMURF uses large Windows smooth, which can lead to more false positive(repeating read). And the shortcomings of sliding window are that can not avoid false negative and false positive, because SMURF is also based on sliding window.

4.4 The RFID data cleaning algorithm based on dynamic Bayesian network

H.Gonzalez et al.[7] put forward the RFID data cleaning algorithm based on Dynamic Bayesian network(DBN), through accessing data cleaning results' accuracy and balance among the cost that is needed to pay to realize the cleaning cost optimization. The fewest resources are used to clean the most original tag data under the premise of ensuring the accuracy of the cleaning's results.

DBN uses an index called implicit model, the actual value of which is the noise value, to determine the real location, and then predicts the tag data by observing historical data, finally calculates a probability value as a standard to measure tags whether exist or not.

Compared with sliding Windows, ESP and SMURF, the advantages of DBN is that do not need to record the latest tag data, because the relationship of observed value and predicted value are combined, and the weight of new data is given higher. While its shortcomings is that can not guarantee the dynamic update because

predicted values and the observed values are gained from historical data, the results of cleaning dynamic tags are not very ideal.

4.5 RFID data cleaning algorithm based on particle filter

RFID-Particle Filter Cleaning(RPFC)[8] is a kind of new tracing cleaning algorithm based on Particle filtering to realize Bayesian filter, using abstract Particle to simulate movements of tags under real situation, and get the mobile tags' location. Then original data will be converted to smooth uncertainty data flow in regard to the state of the tag location. Non-deterministic cleaning method which puts forward in RPFC starts from the two parts: semantic cleaning and misreading cleaning. Semantic cleaning which is based on uncertainly RFID data semantic uses means of adopting weighted particle swarm to track, simulate, and show changes of moving tags' location information. The effective probability value are used to express uncertainty of RFID data semantic. Misreading of cleaning consists two kinds, which are reader's false negative and false positive. And different methods are needed to start uncertainly cleaning respectively.

RPFC using RFID tag data' characteristic of uncertainly semantic to fill the false negative data from adjacent readers which is exist adjacent relation of single tag reader's mobile path (even if certain reader read data for a long time but it has not a high success rate, the missed tag data can also be filled by other readers), and that will not lead to false positive. It's very effective to solve the problem of single reader's false negative when limit error in a small range.

5 Redundancy problems

In order to ensure the tag data can be identified successfully, redundancy is inevitable. There are also many scholars who put forward several solutions, which mainly are divided into the two types of data redundancy and reader redundancy[9], to improve the accuracy of the RFID system.

5.1 Data redundancy

1.Based on Bloom Filters approximation

Metwally et al.[10] put forward using the Bloom Filters(BF) to detect redundancy, but BF do not have automatically delete function, that is, BF will fail due to fill up with data flow produced continuously for a long time. In order to solve this problem, Deng [11] put forward using Stable Bloom Filter (SBF) to process the RFID data flow dynamically, remove old data and set up the unit number corresponding to the maximum number of input data. Wang[12] proposed a cleaning algorithm to solve the distributed redundant data flow and used BF separately for distributed data flow, and achieved all redundant data filtered by sharing BF, which is not hard to find that will cause the waste of bandwidth resources, which introduced a method of cleaning RFID data redundancy based on space-time bloom filter. However, these methods will lead to the mistaken delete redundant data, which does not meet the requirements of data integrity and accuracy.

2.Sorted-Neighborhood Method(SNM)

SNM neighbor sorting algorithm[13] is a classic algorithm of data cleaning to the problem of data redundancy. SNM algorithm detects repetitive data records by comparing the sliding window, and only compares M records of windows each time. This method has improved the speed of comparison, which can improve the efficiency of matching effectively. But at the same time, the precision of repetitive records which are detected by SNM algorithm are limited by sorting of keywords, while keywords' quality directly affect the result of match. When selecting improper keywords, it is possible to make two data which are repetitive recorded far apart and never even in the same sliding window at the same time, so that means it will miss a lot of repetitive records on account of which cannot be identified. However, the selection of sliding window of size M is difficult. When M is big, comparison's times increased, which will result in a lot of unnecessary comparison. When M is small, which will lead to match missed. So these weaknesses determine the SNM algorithm need continuous optimization.

3.Improvement of SNM algorithm[14]

Full-text index technology is a kind of special functional indexes based on tags and face to the full text, which can provide full information, using meaningful words or phrase in the original as retrieval content, pointing to relevant pages or links and do not need information's indexing which can complete the retrieval.

The full-text index technology combined with the traditional algorithm of SNM forms a new algorithm of repetitive redundant data cleaning. In the process of sorting, full-text index technology can effectively make repetitive records appear in the same sliding window, reduce the error of repetitive record detection, improve the efficiency of detection, and make up for the deficiency of the SNM algorithm. When matching two data' similarity, different weight is set according to the importance of different fields, then determined whether two records are repetitive records by compared with similarity. It can reduce the number of unnecessary comparison when looking for repetitive records and does not affect the efficiency, so as to eliminate repetitive redundancy data more effectively.

5.2 Reader's redundancy

1.RRE algorithm

Purdue University and MOTOROLA laboratory put forward a method named RRE[15], which is used to solve the reader's redundant. The basic idea of this method is listed as follows: reader sends its own identification number and identified tag number to all tags recognized to write. Tag stores the largest number of reader identification number. Stored maximum tag number of reader will lock its tags, and the operation is repeated. All tags are checked that identified by each readers, which reader that has not locked any tags is redundant. This method record readers' information by means of writable tags between the reader and the tag communication layer, thus completing the reader and tag one-to-one correspondence, so each tag can only be identified by the only reader. It avoid the problem of data redundancy which is caused by the multiple readers reading the same tag, which reduced new share and

the amount of processing data. Experiments show that, compared with the greedy algorithm, this method is superior in the respect of eliminating reader's redundancy. RRE algorithm has the defect of its outstanding cleaning effect only in the case of reader is static and it does not work in majority of realistic situation.

2.LEO algorithm

Layered Elimination Optimization (LEO) algorithm[16] focuses on the situation that RRE algorithm can not remove dynamic redundancy reader accurately and puts forward the corresponding improvement, which write the lock information of tag by the reader being read first, that is, the reader will possess this tag if it is a deliver when the default tag data received information from the reader for the first time, and locking the owner of each tag at the same time. The reader which had locked tag is an effective work reader, and the probability of repetitive read tags' times are reduced, then redundant reader can be eliminated effectively.

The effect of LEO cleaning is much better than RRE algorithm, but its result of redundant data' cleaning depends on the order of reader reading data, and it may affect the overall efficiency of the algorithm if being read in order.

3.I-RRE

Literature [17] referred Improvable Redundant Reader Elimination (I-RRE) algorithm for the RFID system requirements of the algorithm which are very similar to RRE algorithm. In view of the shortcomings of the reader is evenly distributed, RRE algorithm can not identify the redundant reader correctly, and LEO algorithm rely on the reader's order. I-RRE algorithm has added a variable C-holder to record and identify the number of reader which are the second largest, according to the tags' C-holder value within the reader range to determine the priority, which can achieve the purpose of identifying the redundant reader .

Although I-RRE algorithm improves the insufficient of the RRE algorithm and LEO algorithm, when the distribution of RFID system's reader is very dense, each tag can be read by three or more readers, and the number of reader's tags in radio-frequency region are equal. Then, this algorithm will not be able to perform correctly.

4.TRRE

In the process of running RRE algorithm, when the reader updates the number of tags within the scope of reading, tags are needed to be repeatedly written in reader's identifier, which can lead to increase RRE algorithm execution time, especially tag needs more storage space when its complexity of writing operation time is larger. In view of these shortcomings of RRE algorithm, the TRRE (Two-Step Redundant Reader Elimination) algorithm[18] had been proposed. Suppose it does not need to exchange information among readers, and RFID system had solved the problem of tag conflicts before TRRE algorithm is used, then just send reader's ID information to all tags within radio-frequency region, in this way, it only needs to write reader's ID information replacing the operation that reader's ID information and identification tags' gross wrote simultaneously by tags that had been read by multiple readers. This can reduce the complexity of the write operation.

However, TRRE algorithm and LEO algorithm, both are affected by the reader's reading sequence. In addition, due to the network instability that will cause reader's reading sequence is random, the reliability of the TRRE algorithm needs to be improved.

5.MRRE

Assume that the reader can communicate with all tags in radio frequency range, and RFID middleware's tag data information which is transmitted by reader can be received and saved by RFID middleware. Literature [19] presents a cleaning algorithm MRRE based on middleware.

Relative to the above four kinds of algorithms, MRRE algorithm has obvious advantages, such as it does not need reader to write information for the tag, also does not need to know the system topology structure, in addition, it does not need reader to read and write tags frequently as well. These advantages can reduce the burden of RFID system, to some extent, it can eliminate or reduce the effects of RFID system performance, reduce the probability of the system mistakenly identify redundant readers, and improve the performance of algorithm. What's more, MRRE algorithm guarantees that it will identify individual which contains the largest number of readers at first as effective work reader when multiple reader's reading range exist cross, to minimize the amount of tags which is in cross area of reader identification range, by this way to reduce conflict rates of RFID read-write to minimum. MRRE algorithm is better than other algorithms in terms of system deployment's rationality. The algorithm by using middleware which has rich resources to store tag information that are uploaded by reader, are also applied in a situation where system can not use tags to store information.

The insufficiency of MRRE algorithm shows, in the RFID system, where readers are distributed densely and multiple reader's radio frequency range occur cross, that will result in reader conflict phenomenon. If every reader has recognized the same number of tags, the algorithm will not obtain optimal results.

A series of algorithms, such as RRE, LEO, I-RRE, TRRE, MRRE, and MXREO(RRE+LEO)[20], are used to remove redundant readers. Their insufficient mainly displays in two points: One is reader conflict appeared in the RFID system where reader are distributed densely, and when every reader has read the same amount of tags, these algorithms cannot effectively distinguish redundant readers; The other is that most algorithms depend on the reader's recognition sequence, while it is involved in algorithm's efficiency which needs to be improved. From these points of view perfect algorithms are needed to improve the recognition's effectiveness of redundant reader.

6 Conclusions

In recent years, many experts and scholars proposed a large number of cleaning algorithms according to RFID data flow characteristics. This paper gives a review on RFID data cleaning technology and some typical data cleaning algorithms. With the development of RFID technology, cleaning strategy should be improved and explored

continuously from several aspects because the demand of the data' quality is increasingly higher. At the same time, with the constant development of network technology, security has become an issue which can not be ignored. How to ensure the security and privacy of obtaining data in the dense RFID reader environment are needed to continuously explore in future work.

Acknowledgments. This work is financially supported by the National Natural Science Foundation of P. R. China (No.61373017, No.61572260, No.61572261, No.61672296, No.61602261), the Natural Science Foundation of Jiangsu Province (No.BK20140886, No.BK20140888), Scientific & Technological Support Project of Jiangsu Province (No. BE2015702, BE2016185, No. BE2016777), Natural Science Key Fund for Colleges and Universities in Jiangsu Province (No.12KJA520002), China Postdoctoral Science Foundation (No.2014M551636, No.2014M561696), Jiangsu Planned Projects for Postdoctoral Research Funds (No.1302090B, No.1401005B), Natural Science Foundation of the Jiangsu Higher Education Institutions of China(Grant No. 14KJB520030), Jiangsu Postgraduate Scientific Research and Innovation Projects (SJLX15_0381, SJLX16_0326), Project of Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks(WSNLBZY201509), and NUPTSF(Grant No. NY214060, No. NY214061).

References

1. Zoltan K. Using Rfid And Gis Technologies For Advanced Luggage Tracking[J]. SEA-Practical Application of Science, 2015,2(8): 229-234.
2. Managing and mining sensor data[M]. Springer Science & Business Media, 2013.
3. Xie L, Yin Y, Vasilakos A V, et al. Managing RFID data: challenges, opportunities and solutions[J]. IEEE Communications Surveys & Tutorials, 2014, 16(3): 1294-1311.
4. Bai Y, Wang F, Liu P. Efficiently Filtering RFID Data Streams[C]. CleanDB, 2006:1-8.
5. Jeffery R, Alonso G, Franklin M J, et al. A pipelined framework for online cleaning of sensor data streams[J]. Computer Science, 2005:1-12.
6. JefferyR, Garofalakis M, Franklin M J. Adaptive cleaning for RFID data streams[C]. Proceedings of the 32nd international conference on Very large data bases(VLDB Endowment), 2006: 163-174.
7. Gonzalez H, Han J, Shen X. Cost-conscious cleaning of massive RFID data sets[C]. IEEE 23rd International Conference on Data Engineering. IEEE, 2007: 1268-1272.
8. Want T, Zhang J. Research on RFID data cleaning method based on particle filter[J].Electronic Technology & Software Engineering, 2014(1):214-215.
9. Chen Li.Research on Data Cleaning Methods in RFID Middleware [D].China: Wuhan University of Technology, 2013.
10. Metwally A, Agrawal D, El Abbadi A. Duplicate detection in click streams[C]. Proceedings of the 14th international conference on World Wide Web. ACM, 2005: 12-21.
11. Deng F, Rafiei D. Approximately detecting duplicates for streaming data using stable bloom filters[C]. Proceedings of the 2006 ACM SIGMOD international conference on Management of data. ACM, 2006: 25-36.
12. Wang YL, Wang C, Jiang XH,et al. RFID duplicate removing algorithm based on temporal-spatial Bloom Filter[J].Journal of Nanjing University of Science and Technology, 2015,39(3):253-259.
13. Zhang JZ, Fang Z, Xiong YJ, et al. Optimization algorithm for cleaning data based on

SNM[J].Journal of Central South University (Science and Technology).2010,41(6):2240-2245.
14. Xu X,,Wang J,Yu Y.Research on full text index[J]. Computer Engineering.2002(2):101-103.
15. Carbunar B, Ramanathan M K, Koyuturk M, et al. Redundant-reader elimination in RFID systems[C]. IEEE SECON,2005:176-184.
16. Hsu C H, Chen Y M, Kang H J. Performance-effective and low-complexity redundant reader detection in wireless RFID networks[J]. EURASIP Journal on Wireless Communications and Networking, 2008(1): 1-9.
17. Jiang Y. Improvable-Redundant-Reader Elimination in RFID system[J].Computer Engineering and Applications. 2011,47(5):101-103.
18. Chen J,Yang Z.A Fast and Efficient Algorithm for Redundant Reader Elimination in RFID Application System[J].Computer Knowledge and Technology, 2009,5(16):4285-4288
19. Lv S,Yu S.A Middleware-Based Algorithm for Redundant Reader Elimination in RFID Systems[J]. ACTA ELECTRONICA SINICA.2012,40(5):965-970
20. Chen C. Research on Clean Technology of RFID Sensing Data of Manu-Facturing in IOT[D]. Guangdong University of Technology, 2014.

An Efficient RFID Reader Network Planning Strategy Based on P2P Computing

He XU[1,2,a], Weiwei SHEN[1,2,b], Peng LI[1,2,c], Cong QIAN[1,2,d]
[1] School of Computer Science & Technology/School of software, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
[2] Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks, Nanjing, 210003, China
{[a]xuhe,[b]1214042932,[c]lipeng}@njupt.edu.cn,[d]1227388900@qq.com

Abstract. Radio Frequency identification (RFID) technology has been used in many areas, which needs to deploy many readers in the RFID network. Extra readers in the network will consume more power, resulting in a waste of RFID network resources. Therefore, the elimination of redundant readers is important to optimize an RFID network deployment. RFID reader redundant deployment optimization can ensure that users use the minimum number of readers to cover a specific range of all tags. Peer-to-peer (P2P) computing or networking is a distributed application architecture that partitions tasks or workloads between peers. P2P based RFID network is a novel distributed systems. In this paper, an efficient redundant reader elimination strategy is discussed, which is based on the P2P-based RFID reader network.

Key Words. RFID, Network Planning, P2P, Internet of Things

1 Introduction

Radio-frequency identification (RFID)[1-3] is one method for automatic identification and data capture, which uses electromagnetic fields to automatically identify and track tags attached to objects. RFID are used in many areas, for example, an RFID tag attached to an automobile during production can be used to track its progress through the assembly line; RFID-tagged pharmaceuticals can be tracked through warehouses; and implanting RFID microchips in livestock and pets allows positive identification of animals. RFID technology has been widely used in supply chain, electronic toll collection system or passive key-less entry, location services, medical industry and other fields.

Peer-to-peer (P2P) computing or networking is a distributed application architecture that partitions tasks or workloads between peers. P2P systems had been used in many Internet applications[4-8], and the P2P architecture was popularized by the file sharing

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_36

system. Due to the distributed RFID information system need to real-time update the information, integration of P2P and RFID technology has gained a great interest by researchers[9-14]. The advantage is that peers can communicate each other in P2P network.

RFID system usually includes the following parts: RFID tag, RFID reader and back-end application and database system. In order to accurately monitor a working area, it needs dense deployments of RFID readers and tags in RFID-based application system. In fact, when multiple readers in the same working environment communicate through a common wireless channel, a signal from a reader may produce frequency interference on other readers. This frequency interference occurs when a reader sends a communication signal to read the tag, and the communication signal is interfering with other reader reading the tag. Even if the reader does not overlap the interrogation zone it will interfere with the operation of other readers. The backscatter signal from the tag is so weak that it is vulnerable to interference. Thus, frequency interference of the coverage area can cause inaccurate reading and also lengthening the reading interval. Consequently, for the large-scale deployment of the reader in RFID system, the impact caused by interference of readers in RFID coverage area should be analyzed[15,16]. Extra readers in the network will consume more power, resulting in a waste of reader resources. Therefore, the elimination of redundant readers is important to optimize an RFID network deployment. RFID reader redundant deployment optimization can ensure that users use the minimum number of readers to cover a specific range of all tags. Although redundant readers improve the tracking accuracy, they increase the energy consumption and additional write overhead[17]. When multiple readers share the same perception of the area and the communication channel, a signal from a reader can interfere with other readers[18]. In addition, the redundant data generated by the redundant readers take up more computing resources to achieve the purpose of data gathering. Therefore, effective algorithm to detect and minimize redundant readers is essential to developing RFID network.

Over the past few years, researchers have done extensive research on RFID collision, among which the most important issue is the signal collision. Signal collision is divided into two categories: tag signal collision and reader signal collision[19-25]. Another way to solve the anti-collision problems is to reduce the number of redundancy readers in RFID network. This approach not only increases the effectiveness of RFID systems, but also saves power consumed by redundancy readers. Technology of eliminating redundant readers has been widely used in large warehouses, production lines and large retailer areas. Some research has been done related to this issue [26,27]. RRE algorithm[26] and LEO algorithm[27] is the most representative algorithms of the method. RRE algorithm is a greedy algorithm. the main idea is to record "tag number" that the maximum number of tags a reader can cover. The reader having the maximum number of tags will become the holder of the corresponding tag. Repeat the above steps until all the tags have corresponding readers in the network. In the end, if there is no tag assigned to a reader, it will be eliminated as a redundant reader. Reference[27] pointed out that RRE algorithm is difficult to eliminate redundant readers in some certain RFID network, in order to deal with it, the authors proposed LEO algorithm using a layered approach. The main idea of the LEO algorithm is “first read first own” and all readers send requests to the tags within their coverage's to obtain records of tags. The first reader that sends a quest to the tag will become the holder of the tag. If the tag

already has other reader's ID, the ID can not be changed. Finally, eliminate readers which have no tags.

In the remainder of this paper, we first discuss the RFID data features (section 2). In section 3, we describe related definitions that are used in proposed strategy. The section 4 presents the proposed strategy based on P2P-RFID network in detail. Finally, the section 5 concludes our paper and future work is given.

2 RFID Data Features

Not every successful reading of a tag (an observation) is useful for business purposes. A large amount of RFID data may be generated that is not useful for managing inventory or other applications. Data Cleaning refers to a program which can detect and correct the identifiable error in data files[28]. RFID data have the following features[29]: Temporarily, dynamic and association; Semantic richness; Inaccuracy and heterogeneity; Flow, bulk and mass.

In earlier RFID application, the RFID data were directly sent to the application, then the raw data were processed by applications. The current trend is to provide RFID application with RFID middleware platform, while middleware shields the reader hardware and the upper system, shares the processing of data, and preprocess the raw data (Data Cleaning).

Taking from the angle of readers, inaccuracies of RFID raw data can result in the following errors[30] : (1)Positive error: tag data that should not be recognized is read for some reasons (such as noise) by readers;(2)Negative error: tag data should be the recognized is not read by readers;(3)Read redundancy: tag data is read by multiple times.

3 Related Definitions

There are some classical data cleaning algorithms, such as SMURF model which combines statistical knowledge based on a sliding window cleaning the tag data. Online pipe cleaning frame (ESP) does data cleaning after five stages processing of point, smoothing, merge, decision and blurring. In this paper, RFID reader redundancy problems have been discussed. Reader repeatedly reads the tag data and the reader's identification range overlapping causes a lot of redundant data in RFID system, thus seriously affecting the identification accuracy and efficiency. In the real environment the readers and tags are often very dense so the redundant data read simultaneously by tags and multiple readers cannot be ignored. In the following paragraph, we will describe the methods of redundant reader elimination in detail. The following definitions are given for redundancy readers and the elimination of redundant readers.

Suppose RFID(R, T) denotes an RFID system, where R represents a set of readers, and T represents a set of tags. Suppose Coverage (R, T) represents tag covering domain of the RFID system.

Definition 1 : R^1 is a set of redundant readers set.

Definition 2 : R^2 is a process set of looking for the R^1, which connections is $R^1 \geqslant R^2$.

Definition 3 : Let the read range radius of a reader is M, if the distance L between two readers L<2M, the two readers are called neighboring readers.

Definition 4 : Suppose R_{iN} is a set of neighboring readers of R_i, the neighboring coverage density(NC_i) of R_i is $NC_i = \sum_{j=1}^{n} Coverage\ (R_{ij})$.

Definition 5 : Suppose R_{iN} is a set of neighboring readers of R_i, the weight of reader R_i is $Wi = \frac{Coverage\ (\mathrm{Ri})}{NCi}$.

Figure 1 is an example for reader redundancy in RFID system. Reader R_1 and R_2 cover the same tag T_1, so they are neighboring readers. And the like, R_2 and R_3 are neighboring readers, and R_3 and R_4 are neighboring readers.

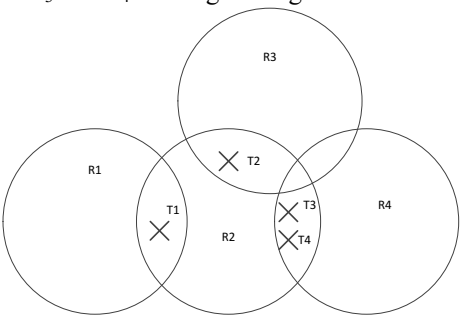


Figure 1. Example for Reader redundancy in RFID system

4 Proposed Strategy

The redundancy elimination algorithm for the reader proposed in this paper is based on the P2P-RFID network system, where one reader can have the ability to communicate with each other among readers, therefore, the following is introduction of P2P-RFID reader network. P2P network is a distributed network of application layer on the IP network, participants in the network i.e. Peers share a portion of the hardware resources (such as processing power, storage capacity, network connectivity, etc.) they have. P2P-RFID reader network system[31] is a new way of P2P-based RFID reader network systems. The P2P-RFID reader network takes advantage of P2P technology, which can work as client and server mode. Embed ZigBee (ZigBee is a short-range, low-power wireless communication technology based on the IEEE802.15.4 standard) technology in the reader, compose P2P network between them, and they can communicate with each other. P2P-RFID reader network architecture proposed in literature[31] is shown in Figure 2.

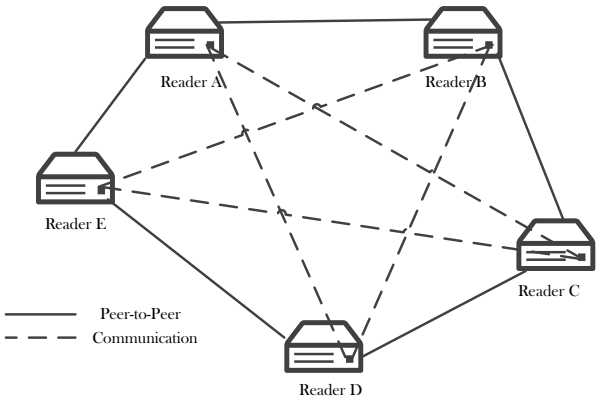


Figure 4

Figure 2. P2P-based RFID reader network system

LEO is applied for high-density wireless RFID network as a distributed optimization technology of wireless RFID network[27], which can eliminate most of the redundant readers, and greatly reduce the "write-to-tag" operation, but LEO relies on RFID reader query sequence. The result is different when query sequence is different. NCD algorithm improves accuracy relative to other distributed algorithms, but the calculation is more complicated, so it does not apply to RFID network in high-density environments.

NCD algorithm is divided into four steps:

(1) Each reader inquiries near readers and tags to calculate the tag covering domain, and writes the reader ID in all the coverage lists of tags it covers.

(2) According to the coverage list and according to the Definition 5, calculate the weight of the reader.

(3) All readers will write the command containing their reader ID and weight into tags they cover. Each tag chooses a holder which has the largest weight in its neighboring areas.

(4) Reader inquiries its covered tags and read the tag holder. A reader having no tags will be eliminated as a redundant reader.

This paper combines the advantages of LEO and NCD algorithm, and proposes a redundant RFID reader algorithm in an environment of high-density readers. First, we make the following assumptions:

(1) The covering domain tag of the reader is known;

(2) P2P-RFID reader network as mentioned in literature[31] are use, by using P2P communication to know whether two readers are neighboring readers.

Query-Response process of reader Ri (i=1…n, the algorithm will be executed in parallel for n times, n is the total number of reader, total number of tag T is m) is listed as the following step:

(1) Reader broadcasts an "inquiry message "(Do you have a holder?) to tags in coverage domain.

(2) Tags in coverage domain of Ri reply to Ri.

(3) If tag's holder = "NULL", reader Ri will write its ReaderID to the tag, else Ri will ignore this reply.

(4) If there is no NULL in all the replies of tags in coverage domain of reader Ri, then Ri will be eliminated as a redundant reader.

(5) Reader Ri broadcast a "query message" to query neighboring reader Rj (j=1…n, n is the number of neighboring readers of Ri).

(6) Reader Ri's neighboring reader Rj replies the tag value.

The above is the entire query response process of the reader and the neighboring reader and tag. After the end of the process, Ri calculates Neighboring Coverage Density Di and Weight Wi according to definition 4 and definition 5. If the neighboring coverage density Dj of reader Rj is 0 and the number of the coverage domain tag Tj is not 0, then reader Rj must not be a redundant reader. List the weight of Ri and Tag Ti, Ti chooses the reader that has the largest weight to be its "holder", finally the readers which are not chosen by any tag are redundant readers.

5 Conclusions

In this paper, an efficient redundant reader elimination strategy is discussed, which is based on the P2P-based RFID reader network. The P2P-RFID reader network takes advantage of P2P technology, which can work as client or server mode, and can communicate with each other. In the P2P-RFID network, through using P2P communication to identity whether the two readers are the neighboring readers. Because LEO relies on RFID reader query sequence, and NCD algorithm improves accuracy relative to distributed algorithms, but the calculation of NCD algorithm is more complicated. This paper makes use of the advantages of LEO and NCD algorithm, and proposes the redundant reader elimination strategy in an environment of high-density readers, which can eliminate redundant readers. The future work includes deployment and refinement of the proposed strategy in realistic environments.

Acknowledgments. This work is financially supported by the National Natural Science Foundation of P. R. China (No.61373017, No.61572260, No.61572261, No.61672296, No.61602261), the Natural Science Foundation of Jiangsu Province (No.BK20140886, No.BK20140888), Scientific & Technological Support Project of Jiangsu Province (No. BE2015702, BE2016185, No. BE2016777), Natural Science Key Fund for Colleges and Universities in Jiangsu Province (No.12KJA520002), China Postdoctoral Science Foundation (No.2014M551636, No.2014M561696), Jiangsu Planned Projects for Postdoctoral Research Funds (No.1302090B, No.1401005B), Natural Science Foundation of the Jiangsu Higher Education Institutions of China(Grant No. 14KJB520030), Jiangsu Postgraduate Scientific Research and Innovation Projects (SJLX15_0381, SJLX16_0326), Project of Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks(WSNLBZY201509), and NUPTSF(Grant No. NY214060, No. NY214061).

References

1. Karmakar, Nemai Chandra. Handbook of smart antennas for RFID systems[M]. New Jersey: Wiley, 2010.
2. Wu Y, Sheng Q Z, Zeadally S. RFID: opportunities and challenges[M]. Next-Generation Wireless Technologies. Springer London, 2013: 105-129.
3. Sakurai S. Prediction of sales volume based on the RFID data collected from apparel shops[J]. International Journal of Space-Based and Situated Computing, 2011, 1(2-3): 174-182.
4. Takeda A, Oide T, Takahashi A. Simple dynamic load balancing mechanism for structured P2P network and its evaluation[J]. International Journal of Grid and Utility Computing, 2012, 3(2-3): 126-135.
5. Okaie Y, Nakano T. A game theoretic framework for peer-to-peer market economy[J]. International Journal of Grid and Utility Computing, 2011, 2(3): 183-195.
6. Sharifkhani F, Pakravan M R. A review of new advances in resource discovery approaches in unstructured P2P networks[C]. Advances in Computing, Communications and Informatics (ICACCI), 2013 International Conference on. IEEE, 2013: 828-833.
7. Eftychiou A, Vrusias B, Antonopoulos N. A dynamically semantic platform for efficient information retrieval in P2P networks[J]. International Journal of Grid and Utility Computing, 2012, 3(4): 271-283.
8. Takaoka M, Uchida M, Ohnishi K, et al. A generalised diffusion-based file replication scheme for load balancing in P2P file-sharing networks[J]. International Journal of Grid and Utility Computing, 2012, 3(4): 242-252.
9. Sula A, Spaho E, Matsuo K, et al. A new system for supporting children with autism spectrum disorder based on IoT and P2P technology[J]. International Journal of Space-Based and Situated Computing, 2014, 4(1): 55-64.
10. Nandakumar R, Chintalapudi K K, Padmanabhan V, et al. Dhwani: secure peer-to-peer acoustic NFC[C]. ACM SIGCOMM Computer Communication Review. ACM, 2013, 43(4): 63-74.
11. Wang C S, Chen C L, Deng D J. P2P-based mobile navigation system with location service[J]. Peer-to-Peer Networking and Applications, 2015, 8(1): 22-31.
12. Hui F C P, Chan H C B. Position Estimation Methods for an RFID-based Positioning System with a P2P Network Architecture[C]. European Wireless 2014; 20th European Wireless Conference; Proceedings of. VDE, 2014: 1-6
13. Hui C P. RFID-based location tracking system using a peer-to-peer network architecture[D]. The Hong Kong Polytechnic University, 2015.
14. Urien P, Piramuthu S. LLCPS and SISO: A TLS-based framework with RFID for NFC P2P retail transaction processing[C]. 2013 IEEE International Conference on RFID. IEEE, 2013: 152-159.
15. D. W. Engels. The reader collision problem[R]. White Paper MIT-AUTOID-WH-007, MIT Auto ID Center, November 2001.
16. D.-Y. Kim, B.-J. Jang, H.-G. Yoon, J.-S. Park, and J.-G. Yook. Effects of reader interference on the RFID interrogation range[C]. In the Proceedings of the 37th European Microwave Conference, pages 728–731, October 2007.

17. N. lrfan and M. C. E Yagoub. Efficient Approach for Redundant Reader Elimination in Large-Scale RF10 Networks[C]. 2010 First International Conference on Integrated Intelligent Computing, pp. l 02-107,5-7 Aug. 2010.
18. Leong K S, Ng M L, Cole P H. The reader collision problem in RFID systems[C]. 2005 IEEE International Symposium on Microwave, Antenna, Propagation and EMC Technologies for Wireless Communications. IEEE, 2005, 1: 658-661.
19. Su-Ryun Lee, Sung-Don Joo, and Chae-Woo Lee. An enhanced dynamic framed slotted ALOHA algorithm for RFID tag identification[C]. The Second Annual International Conference on Networking and Services, 2005, Page(s): 166–172
20. Jae-Ryong Cha and Jae-Hyun Kim. Dynamic framed slotted ALOHA algorithms using fast tag estimation method for RFID system[C]. Consumer Communications and Networking Conference, 2006, Page(s): 768–772
21. Xu Huang and Son Le. Efficient Dynamic Framed Slotted ALOHA for RFID Passive Tags[C]. The 9th International Conference on Advanced Communication Technology, 2007, Page(s):94–97
22. Leian Liu and Shengli Lai. ALOHA-Based AntiCollision Algorithms Used in RFID System[C]. International Conference on Wireless Communications, Networking and Mobile Computing, 2006, Page(s): 1–4.
23. J. Waldrop, D.W. Engels, and S.E. Sarma. Colorwave: an anticollision algorithm for the reader collision problem[C]. IEEE International Conference on Communications, 2003, Page(s): 1206-1210.
24. Ziming Guo and Binjie Hu. A Dynamic Bit Arbitration Anti-Collision Algorithm for RFID System[C]. IEEE International Workshop on Anticounterfeiting, Security, Identification, 2007, Page(s): 457–460.
25. Leian Liu, Zhenhua Xie, Jingtian Xi, and Shengli Lai. An improved anti-collision algorithm in RFID system[C]. 2nd International Conference on Mobile Technology, Applications and Systems, 2005.
26. B. Carbunar, M. K. Ramanathan, M. Koyuturk, C. Hoffmann, and A. Grama. Redundant-reader elimination in RFID systmes[C]. Second Annual IEEE Communications Society Conference on Sensor and Ad Hoc Communications and Networks (SECON), pages 176–184, September 2005.
27. C.-H. Hsu, Y.-M. Chen, and C.-T. Yang. A layered optimization approach for redundant reader elimination in wireless rfid networks[C]. In IEEE Asia-Pacific Services Computing Conference, pages 138–145, 2007.
28. Z. Guo, A.Zhou. Data quality and data cleaning research[J]. Journal of software, 2002, 13(11):2076-2082.
29. Y. Gu, G.Yu, T.Zhang. RFID complex event processing technology[J]. Computer science and exploration, 2007(03).
30. Juels A, Pappu R. Squealing Euros: Privacy protection in RFID-enabled banknotes [C]. Financial cryptography. Springer Berlin Heidelberg, 2003:103-121.
31. Xu H, Wang S, Wang R. A Novel RFID Reader System Framework Based On Peer-to-Peer Network[J]. International Journal of Advancements in Computing Technology, 2011, 3(3): 104-110.

Energy Optimization Algorithm of Wireless Sensor Networks based on LEACH-B

Peng Li[1,2], Wanyuan Jiang[1], He Xu[1,2], Wei Liu[1]
1College of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
2Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks, Jiangsu Province, Nanjing 210003, China
Email: lipeng@njupt.edu.cn

Abstract. Against to the problem that energy consumption unevenly in the wireless sensor network (WSN), we propose an optimized protocol LEACH-B on the basis of LEACH. In this paper, clustering and cluster head election method is improved. Firstly, the great convergence and global optimization capability of BIRCH algorithm can reasonably divide the entire network area into a plurality of sub-regions. Then, electing the cluster head taking the residual energy factor into account, in the sub-region. The experiment result shows that compared with the classic routing protocol LEACH, the LEACH-B can effectively balance the network load and improve the utilization of energy, so as to extend the network life cycle.

Keywords: Wireless Sensor Networks,LEACH, BIRCH, energy optimization

1 Introduction

With the rapid development of communication networks, wireless sensor network technology as a core component of networking technology, is a new and challenging technology. Wireless sensor network is more and more widely used in many areas such as the military, education, environmental protection, medical, industrial and so on. With the development of micro-system technology, pattern recognition, and other related technologies, its applications areas will be more widely[1].

Researchers place many sensor nodes with information acquisition, information process function in the wireless sensor network area to be detected which form a multi-hop, self-organizing network, through the communicate way of wireless data transmission. The random placement of a large number of sensor nodes in unattended place or extremely harsh environments may lead to nodes distribute unevenly in the detection area and appear the phenomenon that nodes densely or sparse distribute in some regions even appear the isolated node. However the isolated nodes can not be ignored, once disregard may result in information collect blind spot, and the paralysis of network seriously. Due to limited hardware resources and battery power, sensors

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_37

nodes can not replace the battery when in use and each node has communications and routing function, the energy problem of node has been one of the key issues in wireless sensor networks[2]. Therefore, how to ensure network quality and minimize the energy consumption of each node in the network to improve network life cycle have become one of the key technologies in wireless sensor network research.

In many research projects of wireless sensor network, network protocol is an essential part.In this paper, we study the low energy adaptive clustering hierarchy, through the analysis of LEACH[3-5](Low Energy Adaptive Clustering Hierarchy) protocol on the working principle, network model, energy consumption, advantages and disadvantages, and the general improvement direction. Because of some deficiencies of the protocol, we propose an improved protocol LEACH-B, which optimizes the energy consumption, extends the life cycle and improves the equilibrium of the network load.

Through the study of LEACH protocol, we analyze the shortcomings of the three stages cluster head election, clusters forming and data transmission in LEACH protocol, and propose a routing algorithm call LEACH-B based on Birch clustering algorithm. The main idea of LEACH-B is to divide the network area into several subregions through the BIRCH algorithm with good convergence and global optimization. Then selecting a cluster head node taking the residual energy of each node in the cluster into consideration. Finally,the data transmission is stable. Theory analysis and simulation results show that the algorithm can optimize network energy consumption and prolong the network life cycle.

2 Related Work

Currently, the research of the optimization of energy consumption in wireless sensor networks mainly focus on routing protocols. Wireless sensor network routing protocol are divided into two parts: the flat routing protocols and clustering routing protocols[6]. The disadvantage of flat routing protocols is that in order to maintain greater routing table which requires a large expenditure and storage space. This is not suitable for large-scale networks. To a certain extent, clustering routing protocol can eliminate these deficiencies. LEACH protocol, one of the clustering routing algorithm in WSN, is the first network protocol which delivers the collected data from cluster head node to sink node directly. Compared with the flat routing protocols, LEACH can conserve the energy and extend the network life cycle to a certain degree.

From the communication model of LEACH protocol in the literature[7], we can reach a conclusion that the energy consumption of the nodes in receiving data and sending data process only have related with the size of data and the distance between nodes. Thus merging large amounts of data and reducing data transmission distance will be the focus of energy economize research.

Currently, researchers have carried out extensive research on wireless sensor network protocol. They put forward a number of new methods to improve the efficiency of clustering routing algorithm, to save energy, such as LEACH-POSC[8], LEACH-K[9], LEACH-R[10].

LEACH-PSOC protocol is an improvement algorithm which is based on particle swarm algorithm and improves LEACH clustering. Its main role is to divide the entire network area into plurality sub-regions through the particle swarm algorithm with good convergence and global optimization and elect the cluster head by taking the residual energy of nodes into consideration. The algorithm can balance network load and extend the network life cycle. However LEACH-PSOC is not suit for the wireless sensor network where nodes distributed unevenly, and not allow the node to be a cluster alone. In addition, the remote node must choose a cluster to join in. Information transfer from the cluster head node to base station may result in more energy consumption.

LEACH-K makes improvements on LEACH based on K-means algorithm. LEACH-K make the improvements on the aspect that the cluster distribute unevenly, the election of cluster head doesn't take residual energy into account and data transmission with the single-hop routing. The cluster head number K has important impact on the result which must be considered to improve. Also LEACH-K is not sensitive to isolated nodes, thus more isolated nodes will lead to node utilization decreased, and consume more energy in the transmission of information.

LEACH-R mainly makes improvements in the cluster head election stage. The clustering and cluster head election of LEACH-R is the same as LEACH, which uses the optimized threshold T(n) to select the cluster head and takes cluster head election reappointment regime. After a round of successful cluster head election, the following rounds will not replace the cluster head, which means it can save energy consumption in the clustering stage. Although it has low energy consumption in the clustering stage, once a cluster head energy exhausted, all of the nodes must be interrupt transfer unconditionally and restart the new round. This approach may lead to incomplete information transmission.

In short, we need to consider not only the clustering for saving energy but also the remote nodes' clustering problem. To prolong the life cycle of the network as much as possible and to ensure the security and integrity of information transmission, we need save energy in the cluster stage, and make improvements in the cluster head election and stable transfer phase. Based on the above analysis, the improved energy optimization algorithm protocol LEACH-B based on BIRCH algorithm is proposed.

3 LEACH protocol analysis

The whole running processes of LEACH protocol is call a round which is divided into cluster structure formation and stable data transmission two stages.

1)Cluster structure formation stages: all nodes are divided by the cluster, each node is randomly selected to be the cluster head.

2)The election process: each node generates a random value between number 0 and 1, if the value is less than T (n), then the node is elected as cluster head. Formula T (n) is shown as follows:

$$T(n)=\begin{cases}0 & otherwise \\ \dfrac{p}{1-p*(r \bmod \frac{1}{p})} & n\in G\end{cases} \quad (1)$$

In the above formulation: p is the ration that cluster heads number takes percentage of the total nodes number; r is the current election rounds; G is the collection of the common nodes in the remaining 1/ p rounds.

After the cluster head node is successfully elected, the cluster head nodes use CSMA-MAC protocol broadcast ADV messages (ADV message includes cluster head messages, cluster head ID and other information) to announce that they are selected to be the cluster head. Non-cluster head nodes receive messages from the head nodes and send the join request to the cluster head nodes according to the strength of the received message signal. The cluster head nodes assign a corresponding TDMA slot tables for each member node, making the node sends data at different time periods to avoid information congestion.

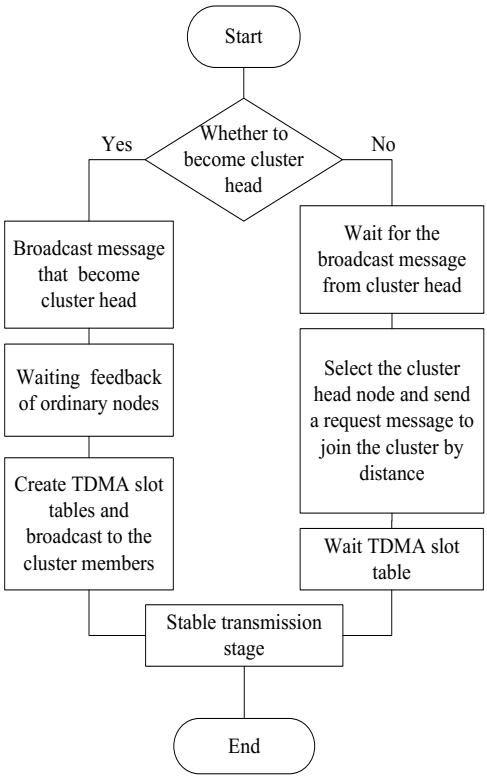


Figure.1. Cluster establishment phase

In the stable data transfer phase, inner nodes send monitoring data to the cluster head nodes on the time slot assigned by head nodes. After receiving the information sent from inner nodes, the cluster head nodes send the information to the sink data after data fusion.

The basic idea of LEACH is to choose cluster head nodes randomly and cyclically, so that each sensor node can share the power load of the whole network to reduce energy consumption and improve survival time of the overall network, LEACH protocol structure is shown in Figure 2.

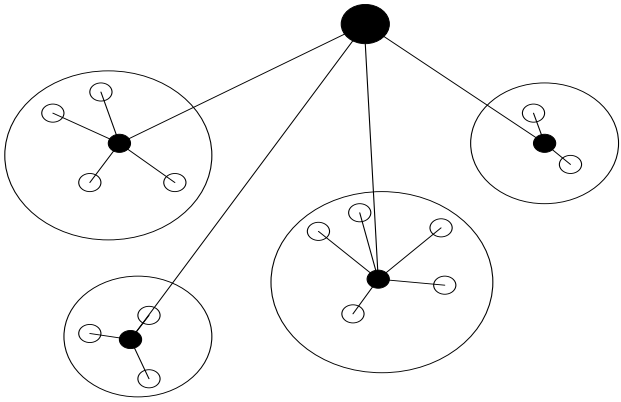


Figure.2. LEACH protocol structure

The advantage of LEACH is that it ensures all nodes have the same opportunity to be selected as the cluster head node, that the network tasks can be assign to each sensor node and each node loads has the balanced energy load. Meanwhile the cluster head nodes deal with the data gathered from common nodes and forward it to the sink node, which greatly reduces the amount of redundant data, improves the utilization of network energy, effectively extending the life cycle of WSN.

However, each round the cluster should be reconstruction, it increases the network energy consumption. Since that LEACH assumes that all nodes can communicate with the sink node directly, and each node has a different MAC compute power, so the agreement is not suitable for large-scale WSN. The agreement also doesn't mention the issue that how cluster head nodes distribute uniformly in the network. Therefore, the phenomenon that all selected head nodes distribute in a certain area in the network is likely to be appear, which makes no head nodes around some common nodes resulting in uneven distribution of energy consumption. LEACH assumes that all the nodes have the same initial energy at the begin and every cluster head node consume approximately the same energy. Therefore.the agreement is not suit for applying to energy imbalance network. The protocol doesn't consider the residual energy of cluster head node, if the cluster head node with low residual energy has been selected, it might not be able to complete the transfer process. Its cluster head generation algorithm can not guarantee cluster head nodes uniform distribution in the network. When the two nodes have a long distance, the energy consumption produces by a single hop router more than the way in multi-hop. Therefore, if the we want to use protocol in the actual network, it still need to be adjusted.

4 LEACH Protocol Optimization

4.1 BIRCH Algorithm

BIRCH clustering algorithm is one of the hierarchical clustering algorithm. Each node is regarded as a cluster in the strategy, these clusters are then combined for the growing cluster until all nodes in a cluster, or a termination condition is satisfied. BIRCH clustering algorithm uses the clustering feature value CF of data points to build a CF tree. Clustering can be quickly performed and the similar degree of the class can be estimated by CF value.

Specific:CF is a triad(N,LS SS), N:the number of the vector in the cluster; LS:the linearity of the vector; SS:the squares of the vector. When given N data points of d dimensional $\{X_1, X_2, ..., X_n\}$, the formulation is defined as follows:

$$\text{Cluster center: } \vec{X}_0 = \frac{\sum_{i=1}^{N} \vec{X}_i}{N} . \tag{2}$$

$$\text{Cluster radius: } R = \left(\frac{\sum_{i=1}^{N} (\vec{X}_i - \vec{X}_0)^2}{N}\right)^{1/2} [12] . \tag{3}$$

The value CF has excellent linear quality, for the merge of two clusters CF = CF1+CF2= (N,LS,SS) = (N1+N2 , LS1+LS2 , SS1+SS2). With the CF value, a CF tree can be constructed as **Figure.3.** :

The implementation process of BIRCH algorithm:

1)Calculate the CF value of the new data point;
2)Start inserting the new CF value from the root;
3)If it not the leaf node, selecting the sub-tree node which has the value close to the CF value to do insertion; If it is the leaf node, selecting the cluster which has the value close to the CF value to do insertion;
4)Go back and check;
5)After insertion, if the cluster diameter is greater than the threshold B, then do not insert and consider the new node as a new cluster; Also if the number of the nodes in the cluster greater than the threshold T, then split the cluster; Especially if the new division is the root node that we must build the new root node and the tree height plus one.

Note: the split way of the cluster in the tree is
1)Assume that the split cluster is U and its child nodes sets is x1, x2, x3....xn, 2)Look for two nodes xi and xj which have the longest distance.
3) Divide the set into two parts, the nodes have a close distance with xi, assign to a group, other nodes have a close distance with xj assign to another group.

4)So set U split into U1 cluster which includes node xi and U2 cluster includes node xj.

So the CF tree is complete, then the nodes continue to clustering.

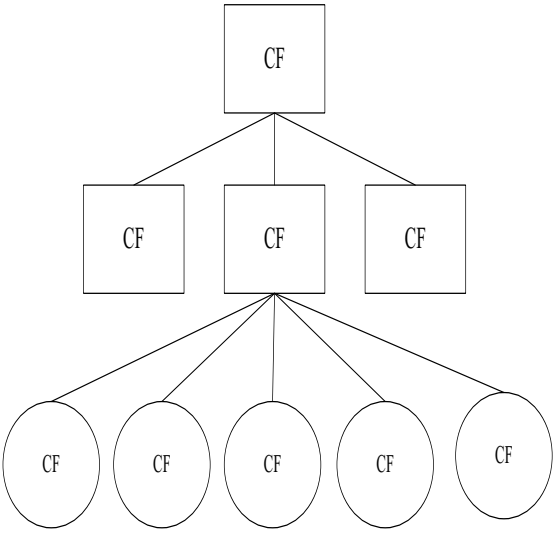


Figure.3. BIRCH algorithm structure diagram

4.2 Improved algorithm of LEACH protocol based on BIRCH

In order to simplify the network model structure, we assume that WSN has the following characteristics: sensor nodes uniformly distributed in the monitoring area, periodic collection of data; sensor nodes and the base station are not moving; each node has a unique identification (ID); communications link is symmetry.

Step1 network clustering stage:

Against to the cluster of LEACH protocol where cluster head node broadcast information and sub-nodes choose a cluster head which has the strength signal to send a join request may lead to unevenly distribution of cluster heads,we make the following improvements:

Assuming that the number of nodes in WSN is n. At the begin, all the sensor nodes have to send a message containing the geographic location to the base station. Since the position of base station is fixed so we can build a two-dimensional coordinate system, in which the position of base station is at (0,0) point and we can obtain the location information of other nodes which make up the initial entire network.

The distance function in BIRCH algorithm is the Euclidean distance. Suppose there are two nodes A and B in the network that the coordinates of point A is (x0, y0) and point B is (x1, y1) so the Euclidean distance $d = \sqrt{(x_0 - x_1)^2 + (y_0 - y_1)^2}$.

Specific process is as follows:

1)Calculate a new data point CF value which is seen as a separate cluster;

2) Other nodes join the cluster according to the distance;

3)When a node adds to the cluster, calculating the radius of the cluster according to the equation (2) and (4,2).

4)Compare the value with cluster radius threshold B, if the value is greater than the threshold don't perform insertion and see the node as a new cluster.

5)Similarly, if the sub-node number is great than the threshold threshold T due to the insertion which make the number of node increase, the cluster start to split according to the way in BIRCH algorithm.

6)Then repeat the above process until all nodes in the region are completed the clustering stage.

Through the improved method above, which using cluster radius threshold T and sub-nodes threshold B to limit the cluster size and allowing individual and remote node to be a cluster, so that we can get more reasonable cluster.

Step2 cluster head establish stage:

Against to the selection of cluster head node which is determined by the value T(n) and random number generates by nodes, and the value T (n) does not consider the residual energy and the distance of each node to the base station node, which may cause the selected cluster head node energy is too low and lead to failure of the entire cluster. Therefore we propose the following improved method.

After completing the clusters establishment phase, we can get k clusters. In each cluster, nodes perceive their remaining energy and inform other nodes . By contrast the remaining energy of nodes in the cluster, the maximum residual energy node elected as the cluster head node, so k cluster head nodes in the whole network. Selecting the maximum residual energy node to be cluster head per round, can guarantee the security of data transmission and prevent the phenomenon that information transmission interruptions.

The splitting way is shown in the figure 4.

1）Assume that the split cluster is U and its child nodes sets is x1, x2, x3....xn,

2)Look for two nodes xi and xj which have the longest distance.

4) Divide the set into two parts, the nodes have a close distance with xi, assign to a group, other nodes have a close distance with xj assign to another group.

4)So set U split into U1 cluster which includes node xi and U2 cluster includes node xj.

5 Simulation and Analysis

We use the matlab tool to simulate and verify the validity of protocol LEACH-B. we assume that WSN is consisted of 100 sensor nodes and nodes are randomly distributed in the 100*100m area, for detecting temperature and wet degree of the area in order to prevent fires. The remote base station locates in coordinates (50,175). Supposing that base station is deployed outside of the region which has the supplementary energy source, nodes are same and static randomly deployed in a region, the initial energy of nodes is same, and the spread channel of the transmitting and receiving is symmetrical.

This article uses the communication model in the document[13]. Let the distance between two nodes is d, the wireless coverage radius is r.The energy consumption of transfer l bit is as formulation(4):

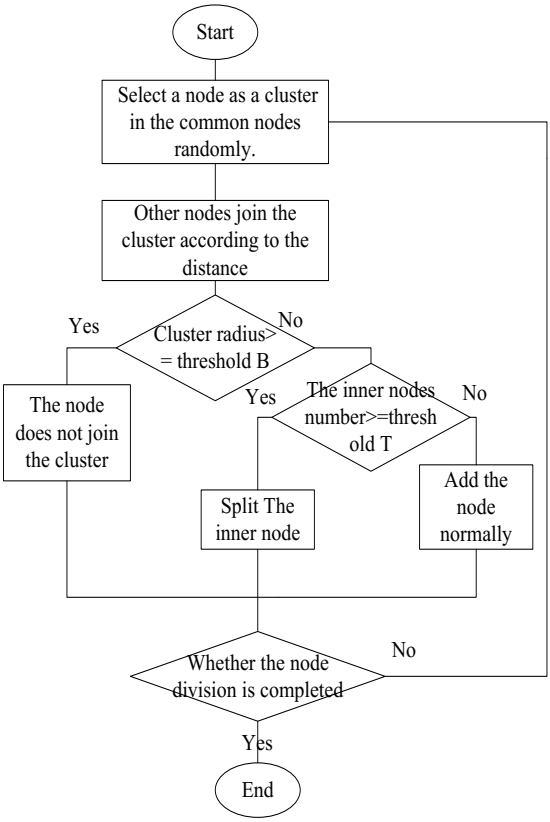


Figure.4. Cluster divide stages of LEACH-B

$$E_{TX}(ld) = \begin{cases} lE_{elec} + l\varepsilon_{fs} d^2, d < r \\ lE_{elec} + l\varepsilon_{mp} d^4, d \geq r \end{cases} . \quad (4)$$

The formulation that a node consumes energy when receive l bit data as follow:

$$E_{RX}(l) = lE_{elec} . \quad (5)$$

Supposing that the initial energy of node is 2J, and E_{elec}=50nJ/bit, ε_{fs}=10pJ /bit. m^{-2}, ε_{mp}=0.0013pJ/bit· m^{-4} 。

Table 1. Simulation parameters table

parameter name	values
network area	（100m*100m）
base station coordinate	(0m,0m)
the number of nodes	100
initial energy	2J
transmission and reception circuit consumes	50nJ/bit
amplifiers factor(space)	10pJ/bit·m-2
amplifiers factor 2(multipath propagation)	0.0013pJ/bit·m-4
data fusion energy	5nJ/bit·signal-1
packet size	4000bit

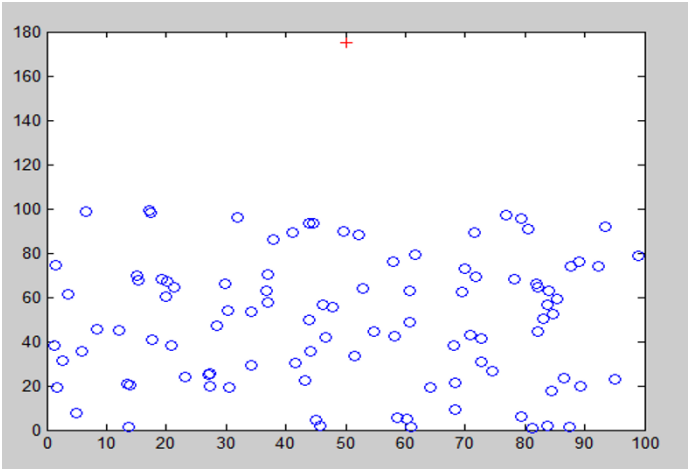


Figure.5. Initial node scene

As we can see from the figure5, nodes are randomly distributed in the area of 100 * 100, the base station node locates at coordinates (50,175) place. The basic idea of LEACH is the concept of “round”, each “round” including cluster stage and stable transmission stage.

Table 2 shows the relationship between the cluster radius threshold B, sub-nodes threshold T and the “round” of LEACH-B.

As shown in Table 2, the paper conducts several experiments that can be found when T is about 20~30m, the experimental data obtained relatively good. The value B represents the cluster radius which is must smaller than d(87.7m), in this way the spread is in free space, consuming less energy than the multipath propagation. Within this range, after many experiments, we find that the best cluster radius is between 20-30m.

Table 2. The round(hundred) of all nodes death

T B	10	20	30	40	50	60	70	80
10	50	82	83	75	73	70	65	62
20	62	87	86	80	75	72	70	68
30	80	89	89	83	82	80	78	76
40	78	82	82	80	81	80	78	74
50	78	82	81	79	80	79	75	72
60	76	80	79	76	75	78	72	69
70	74	70	72	72	72	70	69	65
80	63	68	67	70	6	65	63	62

To sum up,when value T and B both between 20-30m, the LEACH-B algorithm can extend the network life cycle to the maximum extend, and reduce energy consumption. Running about 9000 rounds almost all nodes died.

In this paper we use the two indicators number of death nodes and remaining energy of the nodes to illustrate the performance of LEACH-B. With this two ways, it can fully reflect the quality of the network. When a node energy is less than 0, we consider the node in death status, when the survival nodes account for less than 20% of the total network node that we consider the network in death status.

5.1 Compare of the number of dead nodes

We can see from figure 6 that the first node in death status appears at about 2600 round in LEACH protocol, when it runs to 5000 rounds almost all the nodes are in death status. Compared with LEACH protocol, the first death appears earlier in about 1200 rounds in LEACH-B, this is because a node randomly selection at the beginning of the clustering stage leads to broadcast messages to other 99 common nodes which inevitably lead to waste a lot of energy on the node clustering. At the same time it can be seen that when T=30 B=30, the network life cycle value achieves the best, until about 9000 rounds, almost all the nodes are in death status.

Although the first death status appearing time becomes early, but the network life cycle of LEACH-B has been extend about 70% compared with LEACH. To some extent, it can reduce network energy consumption and prolong the network life cycle effectively.

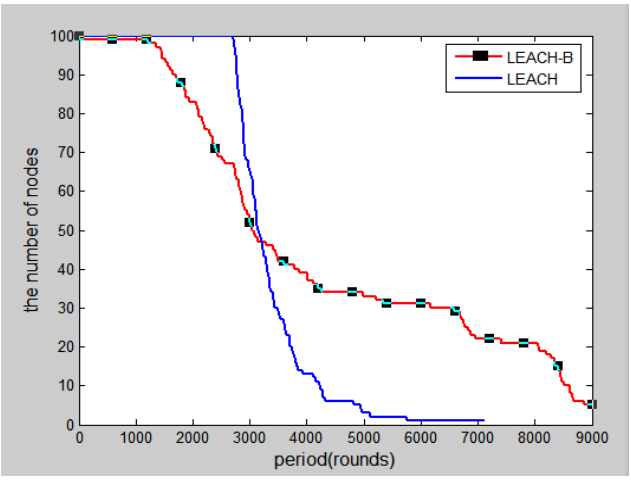


Figure.6. The number of nodes and the number of rounds of LEACH and LEACH-B(T=30 B=30)

5.2 Comparison of system residual energy

From figure 7 we can see that the system remaining energy is the total energy minus the consumption energy which sensor nodes using receiving and sending data. The degree of energy declining is relatively large before the 2000 rounds, after 2000 rounds the degree of energy declining become slowly, and the remaining energy of LEACH-B is always higher than LEACH protocol. It can be seen, when the program starts to run, the energy consumption of two algorithms have little difference, but as the experiment progresses, the degree of energy declining of LEACH-B is slowly and with the increasing of running rounds this advantage is more obvious.

Compare with the LEACH protocol, LEACH-B can effectively extend the life cycle and save energy consumption.

Conclusion

Basing on the energy consumption model of WSN, this paper proposes a new algorithm called LEACH-B which against to the issue the unevenly distribution of cluster head nodes in the LEACH protocol. The simulation shows that the algorithm can improve the network energy utilization efficiency, balance the network load and extend the network life. At the same time, LEACH-B improves the working efficiency of WSN and the data volume of base station receives making a better robustness and communication efficiency of the WSN. In the future, the focus of the research is the data fusion issue on cluster head node which can further improve the algorithm performance, reduce the energy consumption and extend the network life.

Acknowledgments. The subject is sponsored by the National Natural Science Foundation of P. R. China (No.61572260, No.61572261, No.61672296, No.61602261), the Natural Science Foundation of Jiangsu Province (No.BK20140886, No.BK20140888), Scientific & Technological Support Project of

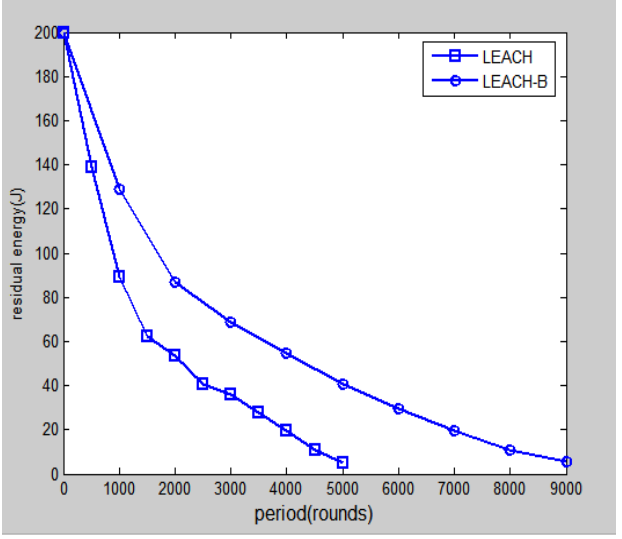


Figure.7. The remaining energy of network and the number of rounds of LEACH and LEACH-B(T=30 B=30)

Jiangsu Province (No. BE2015702, No. BE2016777, BE2016185), Natural Science Key Fund for Colleges and Universities in Jiangsu Province (No.12KJA520002), China Postdoctoral Science Foundation (No.2014M551636, No.2014M561696), Jiangsu Planned Projects for Postdoctoral Research Funds (No.1302090B, No.1401005B), Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks Foundation (No.WSNLBZY201508).

References

1. Loscri, V., Marano,S., Morabito, G.: A Two Levels Hierarchy for Low-Energy Adaptative Clustering Hierarchy(TL-LEACH). Proceedings "VTC2005", Dallas (USA), (2005) 1809-1813
2. Nan Meng, Jiahong Wang, Eiichiro Kodama, Toyoo Takata: Reducing data leakage possibility resulted from eavesdropping in wireless sensor network. Vol. 3. International Journal of Space-Based and Situated Computing (2013) 55-65
3. Majid Bahrepour, Nirvana Meratnia, Mannes Poel, Zahra Taghikhaki, Paul J.M. Havinga: Use of wireless sensor networks for distributed event detection in disaster management applications. Vol. 2. International Journal of Space-Based and Situated Computing (2012) 58-69
4. Jun Xia, Rui Yun, Kai Yu, Fei Yin, Haifeng Wang, Zhiyong Bu: A coordinated mechanism for multimode user equipment accessing wireless sensor network. Vol. 5. International Journal of Grid and Utility Computing (2014) 1-10
5. Konstantin Mikhaylov, Jouni Tervonen: Energy-efficient routing in wireless sensor networks using power-source type identification. Vol. 2. International Journal of Space-Based and Situated Computing (2012) 253-266
6. Li Shancang, Zhang Kewang: Principles and applications of wireless sensor networks. China Machine Press, Beijing (2008) 58~59.
7. Jing Zhang, Ting Yang, Chengli Zhao: Energy-efficient and self-adaptive routing algorithm based on event-driven in wireless sensor network. Vol. 7. International Journal of Grid and Utility Computing (2016) 41-49

8. Chen Xiaojuan, Wang Zhuo, Wu Jie: An improved WSN LEACH routing algorithm. Vol. 26. Sensor Technology (2013) 116-121
9. Chen Xiaoya, Liu Dundong: Improvement on LEACH protocol base on K-means cluster. Xiamen University (2014)
10. Yong-zhen Li, Ai-li Zhang, Yu-zhu Liang: Improvement of Leach Protocol for Wireless Sensor Networks. Vol.12. Third International Conference on Instrumentation (2013) 322-326
11. Olutayo Boyinbode, Hanh Le, Makoto Takizawa.:A survey on clustering algorithms for wireless sensor networks. Vol. 1. International Journal of Space-Based and Situated Computing (2011) 130-136
12. An Wenjuan, Hu Zhengming: The research and improvements of Real-time IP traffic identification technology based on clustering algorithm. Beijing University of Posts and Telecommunications (2013)
13. Mishra T, Panda A.R.: Energy Efficient Coverage and Connectivity with Varying Energy Level in WSN. Vol.12. Computational Intelligence and Networks (CINE), 2015 International Conference on (2015) 86-91

On the Security of a Cloud-Based Revocable IBPRE Scheme for Data Sharing

Jindan Zhang, Baocang Wang

Abstract In ESORICS2014, Liang et al. proposed an efficient cloud-based revocable identity-based proxy re-encryption scheme for public clouds data sharing, aimed at both supporting user revocation and delegation of decryption rights. The main strategy is to let the cloud periodic re-encrypt ciphertexts under the current time period to the next time period. If the user is revoked in the forth coming time period, he cannot decrypt the ciphertexts by using the expired private key anymore. Compared with traditional revocation technique by using PKG, this method has the advantage of computation and communication efficiency. However, in this paper we show an attack which allow the revoked user can decrypt the ciphertexts under the future time period, if the revoked users colludes with the proxy. Although cloud-based revocable identity based proxy re-encryption is a great idea for public cloud storage sharing, it needs further research before this scheme can be practically adapted.

1 Introduction

Nowadays it is a very popular method to store personal or business files on the cloud for its very cheap cost and instant accessing everywhere/everytime. Cloud service providers can manage the hardware and software for cloud storage very flexible and efficient in a cost-effective way. However there are some issues need to be solved before cloud storage can be adapted more widely. Roughly speaking, there are two main kinds of obstacles for cloud storage.

Jindan Zhang
State Key Laboratory of Integrated Service Networks, Xidian University, Xi'an 710071, China, e-mail: 69957106@qq.com

Baocang Wang
State Key Laboratory of Integrated Service Networks, Xidian University, Xi'an 710071, China, e-mail: bcwang79@aliyun.com

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_38

The first one is the bandwidth, although there are rapid development on the communication bandwidth these years such as we witness the growing from 2G, 3G to 4G, several years ago the personal average bandwidth is about 30kb while now it is about 20Mb, the growth rate is about 1:200. However this growth rate for cloud storage is still a great problem. Please note the data generation and growing speed is more sharply. Several years ago average a person has 100MB datum but now this amount is 1TB, the growth rate is about 1:10000. Although we shall use 5G technique in the near future, the bandwidth shall be still a problem compared with the burst growing of datum.

The second one is the security and usability, most users prefer to upload their datum to the cloud only if their files can be sure to be protected well. Most users require the controllability of their uploaded files, or more better, the visibility of their files. However now most cloud storage services can not do this well. In the past when users store their file in their own computers or storage equipments, they can be sure their files being secure for only themselves control on the equipments, there are some physical boundaries on the file's security. However if we upload the files to the cloud, the physical boundaries are not exist anymore, every one's files are virtually organized in the cloud servers, and the attacker can break in the servers to get other person's files or launch the side channel attack on other person's files. If the uploaded files can not be used anywhere/anytime/anyplace, then cloud storage will be no more attractive, thus usability is also an important concern for cloud storage service. Furthermore, how to share files with others is another important usage for cloud storage, if the data owners can not easily and securely share their files with others, lots of effective business corporation among many parties can not be implemented, which will hurdle cloud computation's widely adapted.

To solve these challenges, especially the security problem in cloud storage, we need develop some novel techniques such as PDP, POR, verifiable computation etc. Here we focus on one basic requirement for cloud storage, how to share files securely and flexibly. Proxy re-encryption is a primitive used to share files securely by using cryptographic techniques [1] [2]. In proxy re-encryption [3] [4], a file encrypted under delegator's public key can be re-encrypted by the delegatee's public key by the proxy, without the proxy knowing the underlying plaintexts or secret keys. For the cloud has very powerful computation ability and is semi-trusted, thus it can be the proxy for re-encryption. Thus until now there are many work on the construction of proxy re-encryption [18] [5] [7, 8] [12] [13] [14–17] and how to use proxy re-encryption to share the files with others [9–11], these works are very impressive. In 2001, Boneh and Franklin proposed the first practical identity based encryption scheme, which can reduce the burden caused by public certificates via public key infrastructure. By utilizing the benefits of identity based encryption and proxy re-encryption, a new primitive called identity based proxy re-encryption [6] has been proposed and used for sharing cloud data securely. Recently, Liang et al. proposed a cloud-based revocable identity based proxy re-encryption scheme, which can not only support ciphertext re-encryption, but also support revocation of cloud users, which is a basic requirement for cloud data sharing. However in this paper, we show their scheme can not resist the collusion attack, that is, if the revoked users collude

with the proxy, they can derive some secret information (master private key) which can decrypt any ciphertext in any period. Thus although we think cloud-based revocable identity based proxy re-encryption is a great idea for public cloud storage sharing, it needs further research before this scheme can be practically adapted.

2 Definition and Security Model

2.1 Definition

Definition 1. A Cloud-Based Revocable Identity-Based Proxy Re-Encryption (CR-IB-PRE) scheme consists of the following algorithms. Below we let $\mathscr{I}, \mathscr{T}, \mathscr{M}$ be identity space, time space and message space, respectively.

1. Setup: the setup algorithm intakes a security parameter k and a maximal number of users N, and outputs the public parameters mpk, the master secret key msk, the initial state st and an empty revocation list RL. For simplicity, we assume the following algorithms include mpk implicitly.
2. KeyGen: the private key generation algorithm intakes msk, and a users identity $id \in I$, and outputs a private key sk_{id} for the user id and an updated state st.
3. TokenUp: the token update algorithm intakes msk, an identity id, a token update time period $T_i \in \mathscr{T}$, the current revocation list RL and st, and outputs a token τ_i, where $i \in [1, poly(1^k)]$.
4. DeKeyGen: the decryption key generation algorithm intakes sk_{id}, τ_i, and outputs a decryption key $sk_{id|i}$ for the user id under the time period T_i or $\perp$, if id has been revoked, where $i \in [1, poly(1^k)]$.
5. ReKeyGen: the re-encryption key generation algorithm intakes $sk_{id|i}$, msk, T_i and T_i', and generates the re-encryption key as follows, where $1 \leq i \leq i'$.
   a. ReKeyToken: The re-encryption key token generation algorithm intakes msk, T_i, $T_{i'}$, outputs a re-encryption key token $\psi_{i \to i'}$.
   b. ReKey: the re-encryption key algorithm intakes $sk_{id|i}$ and $\psi_{i \to i'}$, outputs a re-encryption key $rk_{id \to id'}$ which can be used to transform a ciphertext under (id, T_i) to another ciphertext under $(id, T_{i'})$.
6. Enc: the encryption algorithm intakes id, T_i, and a message $m \in \mathscr{M}$ and outputs an original ciphertext C under (id, T_i) which can be further re-encrypted.
7. ReEnc: the re-encryption algorithm intakes $rk_{id|i \to i'}$, and a ciphertext C under (id, T_i), and outputs either a re-encrypted ciphertext C under $(id, T_{i'})$ or a symbol $\perp$ indicating C is invalid, where $1 \leq i \leq i'$.
8. Dec:the decryption algorithm intakes $sk_{id|i}$ and outputs either a message m or a symbol $\perp$ indicating C is invalid.
9. Revoke: the revocation algorithm intakes an identity to be revoked id, a revocation time period T_i, the current revocation list RL, and a state st, and outputs an updated RL.

Correctness: For any (mpk, msk) output by Setup, any time period $T_i \in \mathscr{T}$, any message $m \in \mathscr{M}$, and all possible states st and revocation list RL (where $i \in [1, poly(1^n)]$), if sk_{id} is output by $KeyGen(msk, id), \tau \leftarrow TokenUp(msk, id, T_i, RL, st), sk_{id|i} \leftarrow DeKeyGen(sk_{id}, \tau_i), rk_{id|i\to j} \leftarrow ReKeyGen(sk_{id|i}, msk, T_i, T_j)$ (note for simplicity we set $j = i+1$ here), we have

- if id is not revoked by T_1: $Dec(sk_{id|1}, Enc(id, T_1, m)) = m$.
- if id is not revoked by T_i:

$$Dec(sk_{id|i}, ReEnc(rk_{id|i-1\to i}, \cdots, ReEnc(rk_{id|1\to 2}, Enc(id, T_1, m)))\cdots) = m$$

3 Review of Construction

In our construction we let state st be an unspecified data structure DS, and it depends on which structure we use, e.g., st can be a binary tree. By a tuple (RL, st) we mean a revocation list and its corresponding data structure

1. Setup$(1^k, N)$. The setup algorithm runs $(q, g, p, \mathbb{G}, \mathbb{G}_T) \leftarrow \mathscr{G}(1^k)$, where q is the order of group $\mathbb{G}$, it chooses $\alpha, \beta \in_R \mathbb{Z}_q^*$, group elements $g_2, g_3, v_1, v_2 \in_R \mathbb{G}$, a random n-length set $U = \{u_j | 0 \le j \le n\}$ and a target collision resistant (TCR) hash function $TCR_1 : \mathbb{G} \to \mathbb{Z}_q^*$, where $u_j \in_R \mathbb{G}$. The public parameter is $mpk = (g, g_1, g_2, g_3, v_1, v_2, U, TCR_1)$, the master secret key is $msk = (g_2^{alpha}, g_3^{\beta})$, $RL = \emptyset$ and $st = DB$, where $g_1 = g^{\alpha}$.
2. KeyGen(msk, id). PKG chooses $r_{id} \in_R \mathbb{Z}_q^*$, sets the partial private key sk_{id} as $sk_{id_1} = g_3^{\beta}(u_o \prod_{j\in\mathscr{V}_{id}} u_j)^{r_{id}}$, $sk_{id_2} = g^{r_{id}}$, where V_{id} is the set of all j for which the j-th bit (of id) is equal to 1.
3. TokenUp(msk, id, T_i, RL, st). PKG will check RL first so as to see whether id is revoked or not. If it is revoked, output $\perp$; else proceed. Choose $r_{T_i} \in_R \mathbb{Z}_q^*$, and set the token τ_i as $\tau_{i,1} = (g_2^{\alpha}/g_3^{\beta}) \cdot (v_1 \cdot v_2^{T_i})^{r_{T_i}}$, $\tau_{i,2} = g^{r_{T_i}}$, where i is the index for the time period.
4. DeKeyGen(sk_{id}, τ_i). A user id runs the algorithm as follows.
   a. Choose $\tilde{r} \in_R \mathbb{Z}_q^*$, and randomize the token as $\tau_{i,1} = \tau_{i,1} \cdot (v_1 \cdot v_2^{T_i})^{\tilde{r}}$, $\tau_{i,2} = \tau_{i,2} \cdot g^{\tilde{r}}$.
   b. Choose $r_1, r_2 \in_R \mathbb{Z}_q^*$, and set the updated secret key $sk_{id|i}$ for identity id and time period T_i as

$$\begin{aligned} sk_{id|i,1} &= sk_{id_1} \cdot \tau_{i,1} \cdot (u_0 \prod_{j\in\mathscr{V}_{id}} u_j)^{r_1} \cdot (v_1 \cdot v_2^{T_i})^{r_2} \\ &= g_2^{\alpha}(u_0 \prod_{j\in\mathscr{V}_{id}} u_j)^{\hat{r_1}} \cdot (v_1 \cdot v_2^{T_i})^{\hat{r_2}} \\ sk_{id|i,2} &= sk_{id_1} \cdot g^{r_1} = g^{\hat{r_1}}, sk_{id|i,3} = \tau_{i,2} \cdot g^{r_2} = g^{\hat{r_2}}, \end{aligned}$$

where $\hat{r_1} = r_{id} + r_1$, $\hat{r_2} = r_{T_i} + \tilde{r} + r_2$. Note the user will share $r_1, r_2, \tilde{r}$ with the *PKG* (suppose it is fully trusted) such that the *PKG* can store $(id|i, \hat{r_1}, \hat{r_2})$ in a list $List^{sk_{id|i}}$ for further use.

5. ReKeyGen($sk_{id|i}, msk, T_i, T_i'$). The re-encryption key $rk_{id|i\to i'}$ is generated as follows.
   a. ReKeyToken(msk, T_i, T_i'). If a user *id* holding $sk_{id|i}$ is allowed to update his key to another time period $T_{i'}$, *PKG* generates the re-encryption key token $\varphi_{i\to i'}$ as $\varphi_{i\to i'}^{(1)} = \frac{(v_1 \cdot v_2^{T_{i'}})^{TCR_1(\xi)}}{(v_1 \cdot v_2^{T_i})^{\hat{r_2}}}$, $\varphi_{i\to i'}^{(2)} = (\hat{C_0}, \hat{C_1}, \hat{C_2}, \hat{C_3}) \leftarrow Enc(id, T_{i'}, \xi)$, where $\xi \in_R \mathbb{G}_T$, $\hat{r_2}$ is recovered from $(id|i', \hat{r_1}, \hat{r_2})$ which is stored the $List^{sk_{id|i}}$.
   b. ReKey($sk_{id|i}, \varphi_{i\to i'}$). After receiving $\varphi_{i\to i'}$ from PKG, the user *id* generates the re-encryption key as follows.
      i. Choose $\rho \in_R \mathbb{Z}_q^*$, and set $rk_1 = sk_{id|i,1} \cdot \varphi_{i\to i'}^{(1)} \cdot (u_0 \prod_{j\in\mathcal{V}_{id}} u_j)^{\rho}$, $rk_2 = sk_{id|i,1} \cdot g^{\rho}$, and $rk_3 = \varphi_{i\to i'}^{(2)}$.
      ii. Output the re-encryption key $rk_{id|i\to i'} = (rk_1, rk_2, rk_3)$.
6. Enc(id, T_i, m). Given an identity *id*, a time period T_i, and a message $m \in \mathbb{G}_T$, the encryption algorithm chooses $t \in_R \mathbb{Z}_q^*$, and sets the original ciphertext C as $C_0 = m \cdot e(g_1, g_2)^t, C_1 = g^t, C_2 = (u_0 \prod_{j\in\mathcal{V}_{id}} u_j)^t, C_3 = (v_1 \cdot v_2^{T_i})^t$. We assume that the identity *id* and the time period T_i are implicitly included in the ciphertext.
7. ReEnc($rk_{id|i\to i'}, C$). Parse the ciphertext C under (id, T_i) as (C_0, C_1, C_2, C_3), and the re-encryption key $rk_{id|i\to i'}$ as (rk_1, rk_2, rk_3). The re-encryption algorithm computes $C_4 = \frac{e(C_1, rk_1)}{e(C_2, rk_2)} = e(g^t, g^{\alpha} \cdot (v_1 \cdot v_2^{T_{i'}})^{TCR_1(\xi)})$, and next sets the re-encrypted ciphertext C under $(id, T_{i'})$ as (C_0, C_1, C_4, rk_3). Note if C under $(id, T_{i'})$ needs to be further re-encrypted to the time period $T_{i''}$, then the proxy parses rk_3 as $(\hat{C_0}, \hat{C_1}, \hat{C_2}, \hat{C_3})$. Given a re-encryption key $rk_{id|i'\to i''} = (rk_1', rk_2', rk_3')$, the proxy computes $C_4' = \frac{e(\hat{C_1}, rk_1')}{e(\hat{C_2}, rk_2')}$, and sets the ciphertext C under $(id, T_{i''})$ as $(C_0, C_1, C_4, \hat{C_0}, \hat{C_1}, \hat{C_4'}, rk_3')$.
8. Dec($sk_{id|i}, C$). Given a ciphertext C under (id, T_i), the decryption algorithm works as follows.
   a. For the original ciphertext $C = (C_0, C_1, C_2, C_3)$, the decryptor computes

$$\frac{e(C_1, sk_{id|i,1})}{e(C_2, sk_{id|i,2}) e(C_3, sk_{id|i,3})} = e(g_1, g_2)^t$$

   and outputs the message

$$\frac{C_0}{e(g_1, g_2)^t} = \frac{m \cdot e(g_1, g_2)^t}{e(g_1, g_2)^t} = m$$

   b. For the the re-encrypted ciphertext C:

i. If the re-encrypted ciphertext is re-encrypted only once, i.e. $C = (C_0, C_1, C_4, rk_3 = (\hat{C}_0, \hat{C}_1, \hat{C}_2, \hat{C}_3))$, then the decryptor computes

$$\frac{\hat{C}_0 e(\hat{C}_2, sk_{id|i,2}) e(\hat{C}_3, sk_{id|i,3})}{e(\hat{C}_1, sk_{id|i,1})} = \xi$$

Accordingly, the decryptor can finally compute

$$C_0 \cdot \frac{e(C_1, (v_1 v_2^{T_i})^{TCR_1(\xi)})}{C_4} = m$$

ii. If the ciphertext under id is re-encrypted l times from time period T_1 to T_{l+1}, we denote the re-encrypted ciphertext as $C^{(l+1)} = (C_0^{(1)}, C_1^{(1)}, C_4^{(1)}, \cdots, C_0^{(l)}, C_1^{(l)}, C_4^{(l)}, rk_3^{(l+1)})$, where $C_0^{(1)}$ and $C_1^{(1)}$ are the components of original ciphertext under (id, T_1), and $rk_3^{(i+1)} = (C_0^{(i+l)}, C_1^{(i+l)}, C_2^{(i+l)}, C_3^{(i+l)})$ is the ciphertext under $(id, T_{i+1})(i \in [1, l])$. We recover the message m as follows. First set

$$\frac{C_0^{(l+1)} e(C_2{}^{(l+1)}, sk_{id|l+1,2}) e(C_3{}^{(l+1)}, sk_{id|l+1,3})}{e(C_1{}^{(l+1)}, sk_{id|l+1,1})} = \xi^{(l)}$$

from $i = 1$ to 2 set

$$\frac{C_0^{(i)} e(C_1^{(i)}, (v_1 v_2^{T_{i+1}})^{TCR_1(\xi^{(i)})})}{C_4^{(i)}} = \xi^{(i-1)}$$

finally compute

$$\frac{C_0^{(1)} e(C_1^{(1)}, (v_1 v_2^{T_2})^{TCR_1(\xi^{(1)})})}{C_4^{(1)}} = m$$

9. Revoke(id, T_i, RL, st). Update the revocation list by $RL \leftarrow RL \cup \{id, T_i\}$ and return the updated revocation list.

4 Our Attack

Our attack is based on the following observation: Assume user A is revoked at time period T_i', and he maliciously at time period $T_i (i \leq i')$ colludes with the cloud which holds the re-encryption key, they can derive a powerful secret key which can decrypt the ciphertext for any time period T_j for $j \geq i'$. Considering this scenario: data owner distributes his data contents with users $A, B, C, \cdots, Z$, and these users have paid fee for this content sharing service. Note here user A can be revoked by the data owner

for his payment is not continued any more. A try to corrupt with the cloud by pay a much less fee to the cloud, but the cloud want to hide this corrupt for the data owner, thus it honestly do the re-encryption process but leak the re-encryption keys to user A. By using this re-encryption key, A can derive the powerful secret key which is enough to decrypt the future data contents of data owner, while data owner can not blame the cloud for leaking the re-encryption keys, for these re-encryption keys maybe leaked by other ways, such as leaking in the outsourcing process to the cloud. Concretely the attack can be described as following:

1. According to the DeKeyGen algorithm, User A at time period T_i is with the secret key

$$\begin{aligned} sk_{id|i,1} &= sk_{id_1} \cdot \tau_{i,1} \cdot (u_0 \prod_{j \in \mathscr{V}_{id}} u_j)^{r_1} \cdot (v_1 \cdot v_2^{T_i})^{r_2} \\ &= g_2^{\alpha} (u_0 \prod_{j \in \mathscr{V}_{id}} u_j)^{\hat{r_1}} \cdot (v_1 \cdot v_2^{T_i})^{\hat{r_2}} \\ sk_{id|i,2} &= sk_{id_1} \cdot g^{r_1} = g^{\hat{r_1}}, sk_{id|i,3} = \tau_{i,2} \cdot g^{r_2} = g^{\hat{r_2}}, \end{aligned}$$

2. The cloud holds the re-encryption key $rk_{id|i-1 \to i}$ where $i-1 \leq i$, which is

$$rk_1 = sk_{id-1|i-1,1} \cdot \varphi_{i-1 \to i}^{(1)} \cdot (u_0 \prod_{j \in \mathscr{V}_{id}} u_j)^{\rho}, rk_2 = sk_{id|i-1,1} \cdot g^{\rho}$$

$$rk_3 = \varphi_{i \to i'}^{(2)}$$

where

$$\varphi_{i-1 \to i}^{(1)} = \frac{(v_1 \cdot v_2^{T_i})^{TCR_1(\xi)}}{(v_1 \cdot v_2^{T_{i-1}})^{\hat{r_2}}}$$

$$\varphi_{i-1 \to i}^{(2)} = (\hat{C}_0, \hat{C}_1, \hat{C}_2, \hat{C}_3) \leftarrow Enc(id, T_i, \xi)$$

The cloud sends the re-encryption key to user A.

3. User A then can decrypt the $\varphi_{i-1 \to i}^{(2)}$ and get ξ, with this ξ, he can get $(v_1 \cdot v_2^{T_{i-1}})^{\hat{r_2}}$, and thus can get

$$\begin{aligned} Fsk_{id|i-1,1} &= g_2^{\alpha} (u_0 \prod_{j \in \mathscr{V}_{id}} u_j)^{\hat{r_1}+\rho} \\ Fsk_{id|i-1,2} &= g^{\hat{r_1}+\rho} \end{aligned}$$

4. It is easily to see $Fsk_{id|i-1,1}, Fsk_{id|i-1,2}$ are powerful secret key for id, which can decrypt any period ciphertext for id. For example, for ciphertext (C_0, C_1, C_4, rk_3), user A first decrypt rk_3 (with only (C_0, C_1, C_2) to get ξ, and then finally get

$$C_0 \cdot \frac{e(C_1, (v_1 v_2^{T_i})^{TCR_1(\xi)})}{C_4} = m$$

5. Thus user A can decrypt period ciphertext for id, whether he has revoked or not.

5 Conclusion

In this paper, we show an attack to a recently proposed cloud-based revocable identity based proxy re-encryption scheme by colluding the proxy with the revoked user. Our research results show that, although using cloud for revoking users with identity based proxy re-encryption is a great idea to reduce the complicated burden for private key generation centers if using traditional method, it needs carefully design and more consideration for being adapted widely.

Acknowledgements This work was supported by the National Natural Science Foundation of China (no. 61572390), the 111 Project (No. B08038), the second author is the corresponding author.

References

1. G. Ateniese, K. Fu, M. Green and S. Hohenberger. Improved Proxy Re-Encryption Schemes with Applications to Secure Distributed Storage. In *NDSS(2005), pages 29–43.*
2. G. Ateniese, K. Fu, M. Green and S. Hohenberger. Improved proxy re-encryption schemes with applications to secure distributed storage. In *ACM Transation Information System Security 9 (2006), no. 1, pages 1–30.*
3. M. Blaze, G. Bleumer and M. Strauss. Divertible Protocols and Atomic Proxy Cryptography. In *Advances in Cryptology - Eurocrypt'98,* LNCS 1403, pp. 127–144. Springer–Verlag, 1998.
4. R. Canetti and S. Hohenberger, Chosen Ciphertext Secure Proxy Re-encryption. In *ACM CCS 2007*, pp. 185–194.2007.
5. R. Deng, J. Weng, S. Liu and K. Chen. Chosen Ciphertext Secure Proxy Re-encryption without Pairing. In *CANS'08,* LNCS 5339, pp.1-17, Springer-Verlag, 2008.
6. M. Green and G. Ateniese. Identity-based proxy re-encryption. In *ACNS 2007*, volume 4521 of *LNCS*, pages 288–306, 2007.
7. B. Libert and D. Vergnaud. Unidirectional Chosen-Ciphertext Secure Proxy Re-encryption. In *PKC 2008,* LNCS 4939, pages 360–379, Springer–Verlag, 2008.
8. B. Libert and D. Vergnaud. Unidirectional Chosen-Ciphertext Secure Proxy Re-Encryption. In *IEEE Transactions on Information Theory*, vol. 57, No. 3, pages 1786–1802, 2011.
9. K. Liang, W. Susilo, J. K. Liu. Privacy-Preserving Ciphertext Multi-Sharing Control for Big Data Storage. In *IEEE Transactions on Information Forensics and Security*, Vol. 10, No. 8, pages 1578–1589, 2015.
10. K. Liang, M. H. Au, J. K. Liu, W. Susilo, D. S. Wong, G. Yang, T. Phuong, Q. Xie. A DFA-Based Functional Proxy Re-Encryption Scheme for Secure Public Cloud Data Sharing. In *IEEE Transactions on Information Forensics and Security*, Vol. 9, No. 10, pages 1667–1680, 2014.
11. K. Liang, W. Susilo. Searchable Attribute-Based Mechanism With Efficient Data Sharing for Secure Cloud Storage. In *IEEE Transactions on Information Forensics and Security*, Vol. 10, No. 9, pages 1981–1992, 2015.
12. J. Shao and Z. Cao. CCA-secure proxy re-encryption without pairing. In *PKC 2009,* LNCS 5443, pages. 357–376, Springer–Verlag, 2009.
13. J. Shao, Z. Cao, P. Lin. Generic construction for CCA-secure unidirectional proxy re-encryption. In Security and Communication Networks, no. 2, pages 1-16, 2009.
14. J. Weng, R. H. Deng, C. Chu, X. Ding, and J. Lai. Conditional proxy re-encryption secure against chosen-ciphertext attack. In *ACM ASIACCS 2009*, pages 322–332, 2009.

15. J. Weng, Y. Yang, Q. Tang, R. Deng, and F. Bao. Efficient conditional proxy re-encryption with chosen-ciphertext security. In *ISC 2009*, volume 5735 of *LNCS*, pages 151–166, 2009.
16. J. Weng, Y. Zhao, G. Hanaoka. On the Security of a Bidirectional Proxy Re-encryption Scheme from PKC 2010. In *PKC 2011*, pages 284-295, 2011.
17. J. Weng, M. Chen, Y. Yang, R. Deng, K. Chen and F. Bao CCA-secure unidirectional proxy re-encryption in the adaptive corruption model without random oracles. In *Science China Information Sciences*, 53, : 593-606, 2010.
18. S. Chow, J. Weng, Y. Yang, R. Deng Efficient unidirectional proxy re-encryption. In *AFRICACRYPT 2010*, volume 6055 of *LNCS*, pages 316–332, 2010.

Semantic Integration and Correlation of Digital Evidences in Forensic Investigations

Flora Amato, Giovanni Cozzolino, Nicola Mazzocca

Abstract The growing adoption of digital and computer-related technologies has led to the increase of computer crimes occurrences. Today we are exposed to a great number of threats that involve IT technologies, such as terrorism or cyber-scams through Internet, child pornography, espionage and tampering of computer equipment. Digital investigations play a leading role to guarantee a legal use of digital technologies, as well to search for evidences left by the author of a crime. Digital investigations often combine a wide span of skills and require the adoption of different tools and techniques in order to capture, store and analyze evidences from a Computer Crime Scene. The lack of integration and interoperability between these tools protract the analysis process and pauperize presentation results for non-technical parties (from legal, academic or business areas). This work presents a method, based on Semantic Web technologies, to integrate, correlate and query different sources of forensic data with the goal of more valuable reconstruction of a case.

1 Introduction

Digital Forensics practices are facing an increasing number of challenges. IT systems are exposed to a constant threat of various and increasingly sophisticated types of attacks. The constant evolution of the technological landscape with the introduc-

Flora Amato
DIETI - Università degli Studi di Napoli "Federico II", via Claudio 21, Naples, e-mail: flora.amato@unina.it

Giovanni Cozzolino
DIETI - Università degli Studi di Napoli "Federico II", via Claudio 21, Naples e-mail: giovanni.cozzolino@unina.it

Nicola Mazzocca
DIETI - Università degli Studi di Napoli "Federico II", via Claudio 21, Naples e-mail: nicola.mazzocca@unina.it

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_39

tion of new technologies as well the large volumes of data, complicate the work of Digital Forensic experts, facing an almost unmanageable amount of information despite the exiguity of time to accomplish their tasks.

In [1] the author outlined five major challenge areas for digital forensics: 1. The complexity problem, arising from data being acquired at the lowest format with increasing volume and heterogeneity, which calls for sophisticated data reduction techniques prior to analysis. 2. The diversity problem resulting from a lack of standard techniques to examine and analyze the increasing numbers and types of sources. 3. The consistency and correlation problem resulting from the fact that existing tools only support the acquisition process, but doesn't otherwise assist the examiner in other phases of investigation. 4. The volume problem, resulting from increased storage capacities and the number of devices that store information. 5. The generation of a unified timeline of events from multiple sources that present different time zone references, timestamp interpretations, clock skew/drift, etc.

Additional problems may arise in the case of multiple parties being involved in the analysis process due to different levels of expertise or communication problems that may arise from a lack of an agreed set of terms and definitions. A side-problem is focused on the presentation: the results of a DF investigation have to be delivered to a court's jury or an organization's decision making board in a more understandable manner avoiding technical or intricate terminology. During the processing of evidences, of course, the investigator should not affect their admissibility or probative value.

Moreover, DF cases require the adoption of practices and tools usually restricted to specialized subset of activities such as file carving, log analysis tool, etc. These tools produce outputs in different formats and become quickly outdated by newly introduced technologies or data formats. A digital investigator, during the analysis process, needs to process and integrate data from heterogeneous sources such as hard disks, memory, mobile devices, security logs[2, 3], network traffic and so on, in order to find traces and reconstruct the events.

The authors in [4] also confirmed that the volume of data to be processed, combined with the heterogeneity of sources that they cam from and the level of expertise necessary to ensure the court admissibility of the resultant evidence, represent one of the most valuable challenge for digital forensics.

The motivation of this work is to propose a method to automate and enrich the analysis and reasoning process of a digital investigation with the adoption of practices and technologies of Semantic Web. The addition of semantically expressed assertion to the artifacts, produced during the acquisition, examination and analysis phases, should improve the overall result of the presentation process, enabling more powerful searches and more accurate refinement and reporting of traces[5, 6, 7].

The reminder of the paper is structured as follows: Section 1.1 presents the theoretical background in the area of Digital Investigations discussing the basic principles of digital evidence and presenting the challenges that the field faces; in Section 2 there is a short presentation of related work that has also merged partially or fully these two areas; Section 1.2 discusses the motivation of the adoption of a semantic-based approach, providing a high-level view of how these two disciplines can be

combined and with what advantages; Section 3 presents the proposed methodology, discussing the overall structure of it and the method that deal with the integration, correlation and querying of the source data; in Section 4 we describe the system architecture that could implement the proposed methodology; finally, in Section 5 we present some conclusions of the work.

1.1 Digital Investigations and Digital Evidences

Digital evidence is the central point of any digital investigation or forensics process. A multitude of different definitions of digital evidence can be found in the literature such as:

- any data stored or transmitted using a computer that support or refute a theory of how an offence occurred or that address critical elements of the offence such as intent or alibi [8]
- information of probative value stored or transmitted in digital form. [9]
- any digital data that contain reliable information that supports or refutes a hypothesis about an incident [10]

Thus, its rigorous and authenticated handling is of extremely importance. Digital evidence can be identified and interpreted from different abstraction layers, such as from a physical point of view (e.g. electrical charges in memory transistors), or from a logical perspective (e.g. files, folders, etc.). The interpretation of the evidence and its relevance through the context of the current investigative case is what adds value to it.

In [11] author has identified three basic properties of digital evidence, namely fidelity, volatility and latency. Fidelity is a property that refers to Chain of Custody forensic principle, and involves the adoption of techniques that grant the verification of the integrity of evidences through the process. Volatility is a property of an evidence related to the nature of the support on which they hare stored (registers, memory, disk, etc.): this property affects considerably the practice of acquisition of evidences and further processing of them. The focus of current work is on the third property, latency: a digital encoding needs additional information on how to be contextualized and interpreted.

A Digital Forensic Investigation is the set of principles and procedures followed during the lifecycle of digital evidence, from acquisition to analysis and presentation. Forensic tools are playing a major role in modern digital investigation practice, since evidences came from a multitude of different sources such as media devices (e.g. hard disks, USB flash drives), network sources (IDS Firewall logs, packet or flow captures), live memory and mobile devices. The distributed nature of digital evidence, in conjunction with the specialized and evidence-oriented design of forensic tools, calls for advanced methods of tool interoperability and correlation of evidence.

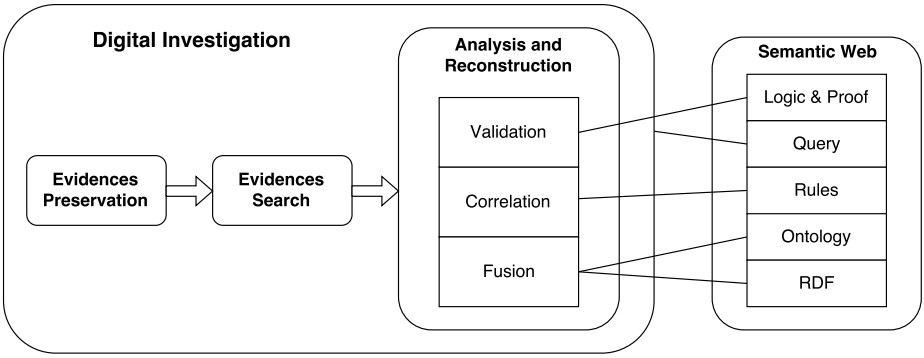


Fig. 1 Relation between Digital Investigation phases and Semantic Web technologies.

1.2 Motivations

The proposed approach for integrating semantic web technologies in digital investigation processes and tools can improve, through semantics vocabularies, the automation of parts of the analysis with respect to evidence discovery and correlation. The main potential advantages of such an approach regard:

- Information Integration: the RDF data model promotes easy integration of heterogeneous data due to its schema independence and standardized statement representation of subject-predicate-object form.
- Classification and Inference: the RDF/OWL combination can infer class membership and typing information from data based on the ontological specified definitions.A semantic reasoner can use definitions and restrictions of the ontology, in order to infer dynamically class membership of the instances.
- Extensibility and Flexibility: RDF/OWL provide backward and forward compatibility, with respect to the data model of forensic tools input and output. OWL provides the flexibility that different ontologies can be defined according to the stakeholder's interests and scope, but still be able to integrate through ontology mapping processes.
- Search: Documents are enriched with semantic markup that is later used over or along traditional keywords for document indexing purposes. A query can be extended through a reasoning engine and generate more meaningful data.

Semantically-based approach can provide an automated framework to formalize and describe in a more expressive way a representation of existing domain knowledge. This framework covers all the main phases of a digital investigation: System Preservation, Evidence Searching and Event Reconstruction. In particular, during Event Reconstruction phase, evidence that have been collected during the first phase and extracted or filtered during the second phase, are analyzed, integrated and correlated with the goal of providing a reconstruction of any performed past actions. Figure 1 shows the relations between the phases of a digital investigation process and the Semantic Web technologies adopted.

2 Related Work

Although correlation has been previously described as one of the most important goal of the analysis phase, current support for correlation analysis is poor among most forensic tools. In [12] authors have proposed a framework for supporting automated integration and correlation of data objects as contained in memory, network traces, disk images, log files and so on. Correlation techniques suffer of domain-depending problem, while digital investigators have to deal with a growing variety of different domains and thus event correlation is becoming even more difficult. Another challenge is the extensibility of such proposed solutions that usually lack of contextual information.

Multiple approaches adopting Semantic Web technologies have been proposed in literature. DFXML (Digital Forensics XML)[13] is an XML-based approach, in which author intends to provide a standardized XML vocabulary for representing metadata of evidences, providing various intuitive forensics-related XML tags such as `<source>`, `<imagefile>`, `<acquisition_date>`, `<volume>`, etc. To promote the DFXML format, author provide also a set of tools (e.g Fiwalk[1]) that can either produce or consume DFXML for various purposes. Another XML-based approach is described in [14], where authors proposed the DEX (Digital Evidence Exchange) format. DEX can also wrap the output from various tool in an XML-based representation but with an additional capability of encoding the specific instructions given and the sequential order of the tools used. The tool defines a set of metadata that describe what was the sequence of actions leading to the specific extracted artifacts. Although these projects promote a standardized format and tool interoperability, they are missing to convey the semantic content of what they represent[15].

Different approaches are RDF-based, that, despite the ability to express arbitrary metadata, have found limited usage in the area of digital investigations. The most famous adoption of RDF is in the AFF4 forensic format project[16, 17]. The Advanced Forensic Format (AFF) is file format and container to store digital evidence. The container includes both a copy of the original data as well as arbitrary metadata. Metadata can include system-related information such as device-related data or user-specified ones. RDF makes an important improvement due to the support of creating attributes with standard or custom types instead of only string types. To improve the chain of custody, in [18] author introduces into AFF4 domain-specific RDF concepts, such as evidence access information (e.g. date and geographical location of access), examiner information, as well as artifact-related information that were produced during the investigation process (e.g. name of artifact, date, action).

FORE (Forensics for Rich Events) architecture[19] is composed of a forensic ontology, a generic event log parser and a custom-defined rule language, FR3. The forensic ontology is based on two main concepts, the Entity for representing tangible objects and the Event for representing the entities' state changes over time. In

[1] Fiwalk is a tool that produces DFXML describing files in a disk image. It provides an abstraction layer over a Sleuth Kit distribution thus allowing new features such as reporting of differences between two disk images based on their DFXML representation.

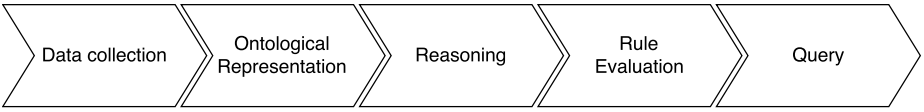


Fig. 2 Phases of the proposed methodology.

order to represent causal linkage, an ObjectProperty linking individuals of the Event class is defined. These semantic representations of the events are then added to the knowledge base[20]. Rules expressed in the FR3 language are evaluated against the knowledge base in order to add new causality links between events. By taking advantage of the OWL capabilities of referencing external ontologies, the authors are able to express correlation rules that combined concepts and events from disparate domains[21]. Similar to the former approach, in [22] authors demonstrate examples of using a blank ontology for encoding results retrieved from forensic tools with the ability to encode various forensically relevant types of data and their relationships. The authors use the ontology query language (SQWRL) for extracting additional information of probative value.

3 Methodology

This section describes the proposed methodology for digital evidence integration and correlation. The basic work-flow of the methodology is presented in Figure 2.

The Data Collection phase include all the forensic acquisition techniques. The goal of this phase is to generate the appropriate input for the subsequent phases. In this phase the examiner has to be compliant with forensic principles during the acquisition of involved data sources (e.g. hard disks, network packet captures, OS and application logs, memory contents, mobile devices etc.) in order to respect the chain of custody. This is important for the subsequent admissibility of the generated results. Due to computational complexity issues, a preprocessing and data reduction step can be also applied using common techniques such as KFF (Known File Filter), in order to reduce the amount of input data for the subsequent phases.

The Ontological Representation phase expects to parse the input data, with appropriate software or even hardware tools, and to represent them in a common and explicit format using the RDF data model and respective ontologies. In this phase can be considered, if existent, shared domain ontologies, or can be specified new ad-hoc ontologies. Parsing tools have to be flexible enough to operate with shared ontologies as well as custom made ontologies specified by the investigator. The task of this step should be the transformation of the input data into a set of triplet constituting the RDF data model. The asserted axioms can refer other ontology files that describe classes and/or properties. The final output can be stored using the various ontology formats such as RDF/XML or RDF/OWL.

The Reasoning step involves the use of an OWL-based reasoner upon the previously generated ontological representation of data, in order to infer additional axioms based on the instances' relations. Different types of new axioms can be inferred, enriching the instances with information regarding the definition of their class, properties or subclasses. Thanks to property restrictions or subclass hierarchy, the reasoner can classify instances as members of specific class, so a resource can belong to multiple classes in parallel. Furthermore a resource can dynamically be classified with higher precision compared to concepts defined in the original data. Moreover, the inference engine analyze property relations in order to infer new property axioms that are attached to the individuals.

In the Rule Evaluation step, SWRL Rules are evaluated by a rule engine and the newly asserted axioms are inserted back to the ontology. SWRL Rules can assert new additional axioms that cannot be inferred due to the DL-safe expressivity limitations of OWL. This phase can be realized also with the support of external Rule Engines, and the newly inferred axioms should be translated again back to the RDF data model and integrated in the existing ontology.

Finally, queries can be expressed in the SPARQL language and submitted to a SPARQL endpoint hosting the sets of asserted and inferred axioms. Both OWL-inference as well as SWRL evaluation relations in the form of predicates can be established between resources of the same or different datasets.

4 System Architecture

In this section, will be resented a reference implementation of the methodology with a brief discussion of the tools and techniques used. The method is quite flexible the system presented in this section is just one of the different ways that the methodology can be implemented, also thanks to the continuous evolution of Semantic web technologies.

System architecture has been designed to support a wide range of data sources such as hard disk forensic images, network packet captures, firewall logs, and so on. A schematic overview of the architecture is presented in Figure 3 and its components are further discussed below.

Evidence Manager

The evidence manager is tasked to load binary content of evidence files, given their locations and names, for further processing. The Evidence Manager is also responsible for identifying the type of given source and for verifying the integrity of the given files computing hash values of image files and comparing them with values optionally provided with original files.

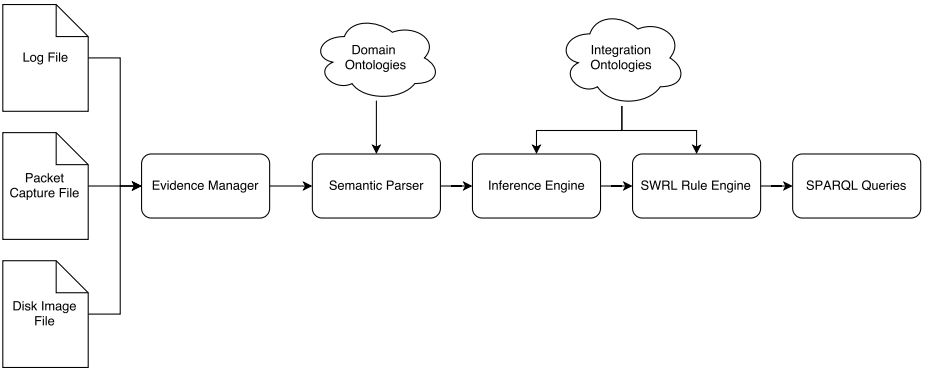


Fig. 3 System architecture for proposed methodology.

Semantic Parser

Semantic parser module accepts the path and name of an evidence file and a collection of strings which could be sets of IP addresses, MD5 hash values, timestamps etc. The module returns an object that represents an OWL ontology along with the OWL Axioms and OWL Annotations it consists of. The parser also load the ontologies for the different formats of data or tool outputs.

Inference Engine

The inference engine imports all the referenced domain ontologies that the semantic parsers have used and is now able to perform automated reasoning according to the OWL/OWL2 specifications. Reasoners can specify the granularity of which types of OWL axioms the examiner wants to be inferred. One type is the generation of class assertion axioms such as in the case of class hierarchies where an individual that is a member of a subclass can be inferred to be a member of the parent class as well. Another useful type is the generation of inverse object properties axioms where two individuals are establishing reverse links based on the asserted object property and the defined inverse object property of the latter that can improve the performance of query execution later on.

An additional ontology can be created to import all the referenced domain ontologies as well as their indirect references. In this ontology can also be defined new classes, additional restrictions, new object properties etc. The inference engine is then able to infer new axioms based not only on the domain ontologies but on the investigator custom-defined one. This allows for a level of flexibility since the domain ontologies may not be easily modifiable but the OWL specification allows, even more promotes, the creation of supplementary ontologies that may reuse entities defined in other ontologies.

SWRL Rule Engine

SWRL Rule Engine module uses SWRL rules in order to establish relations between individuals belonging in different ontologies but representing similar concepts like IP addresses as well as correlate different individuals. SWRL rules play a major role in the automated integration and correlation parts of the methodology. The SWRL rules can be kept in a separate text file which promotes the decoupling of the actual rules from the implementation of the methodology as well as enables sharing and reuse of the rules in multiple cases as well. In the case that the call to the reasoning engine may need to take advantage of axioms that can only be inferred by the SWRL rules, then the call order may need to be reversed. Another option can be that a call to the reasoning engine is done twice, one before the rule engine and the other after it.

SPARQL Queries

The final module is responsible for accepting SPARQL queries from the user and evaluating them against a SPARQL query engine. The set of RDF triples that have been either asserted during the semantic parsing of the source data or inferred by the reasoning or the rule engine are loaded in-memory and SPARQL queries can be evaluated against it. The queries can once more be stored in separate files thus promoting reuse and decoupling from the implemented system. The Jena framework is providing the ARQ query engine that supports the SPARQL RDF query language.

5 Conclusions

In this work we propose the adoption of the Semantic Web framework as an enabling technology that can alleviate some of the most challenging problems that the field of digital forensics faces. The main objective was to take advantage of key features of Semantic Web in the area of heterogeneous data integration, support for automation and improved analytical capabilities for the investigator in the form of an expressive and flexible querying layer. We propose a reusable and extensible methodology based on the semantic representation, integration and correlation of digital evidence describing both a generic conceptual framework bridging the two fields as well as describing an architecture that implements it. The outcome of the present work is that the Semantic Web technologies can improve the process of a digital investigations due to their abilities related to automation and data integration. An ontological representation of the collected data can alleviate the problem of heterogeneous forensic tools and the capabilities of an inference and rule engine can improve analytical skills by allowing easy and fast ways to integrate and correlate forensic evidences.

References

1. S. Raghavan, "Digital forensic research: current state of the art," *CSI Transactions on ICT*, vol. 1, no. 1, pp. 91–114, 2013.
2. A. Cilardo, A. Mazzeo, L. Romano, and G. Saggese, "An FPGA-based key-store for improving the dependability of security services," 2005, pp. 389–396.
3. A. Cilardo, "Exploring the potential of threshold logic for cryptography-related operations," *IEEE Transactions on Computers*, vol. 60, no. 4, pp. 452–462, 2011.
4. D. Lillis, B. Becker, T. O'Sullivan, and M. Scanlon, "Current Challenges and Future Research Areas for Digital Forensic Investigation," 05 2016.
5. X. Wang, D. Huang, I. Akturk, M. Balman, G. Allen, and T. Kosar, "Semantic enabled metadata management in petashare," *International Journal of Grid and Utility Computing*, vol. 1, no. 4, pp. 275–286, 2009.
6. A. Eftychiou, B. Vrusias, and N. Antonopoulos, "A dynamically semantic platform for efficient information retrieval in p2p networks," *International Journal of Grid and Utility Computing*, vol. 3, no. 4, pp. 271–283, 2012.
7. C. Lai and C. Moulin, "Semantic indexing modelling of resources within a distributed system," *International Journal of Grid and Utility Computing*, vol. 4, no. 1, pp. 21–39, 2013.
8. E. Casey, *Digital Evidence and Computer Crime: Forensic Science, Computers, and the Internet*, 3rd ed. Academic Press, 2011.
9. S. W. G. on Digital Evidence (SWGDE) International Organization on Digital Evidence (IOCE), "Digital evidence: Standards and principles," https://archives.fbi.gov/archives/about-us/lab/forensic-science-communications/fsc/april2000/swgde.htm#Definitions.
10. B. Carrier and E. H. Spafford, "An event-based digital forensic investigation framework," in *Digital forensic research workshop*, 2004, pp. 11–13.
11. B. L. Schatz, "Digital evidence : representation and assurance," Ph.D. dissertation, Queensland University of Technology, 2007. [Online]. Available: http://eprints.qut.edu.au/16507/
12. A. Case, A. Cristina, L. Marziale, G. G. Richard, and V. Roussev, "Face: Automated digital evidence discovery and correlation," *Digit. Investig.*, vol. 5, pp. S65–S75, Sep. 2008.
13. S. L. Garfinkel, "Automating disk forensic processing with sleuthkit, xml and python," in *Systematic Approaches to Digital Forensic Engineering, 2009. SADFE'09. Fourth International IEEE Workshop on*. IEEE, 2009, pp. 73–84.
14. B. N. Levine and M. Liberatore, "Dex: Digital evidence provenance supporting reproducibility and comparison," *digital investigation*, vol. 6, pp. S48–S56, 2009.
15. F. Amato, G. De Pietro, M. Esposito, and N. Mazzocca, "An integrated framework for securing semi-structured health records," *Knowledge-Based Systems*, vol. 79, pp. 99–117, 2015.
16. S. L. Garfinkel, "Aff: A new format for storing hard drive images," *Commun. ACM*, vol. 49, no. 2, pp. 85–87, Feb. 2006.
17. M. Cohen and B. Schatz, "Hash based disk imaging using {AFF4}," *Digital Investigation*, vol. 7, Supplement, pp. S121 – S128, 2010, the Proceedings of the Tenth Annual {DFRWS} Conference.
18. G. Giova, "Improving chain of custody in forensic investigation of electronic digital systems," *International Journal of Computer Science and Network Security*, vol. 11, no. 1, pp. 1–9, 2011.
19. B. Schatz, G. M. Mohay, and A. Clark, "Generalising event forensics across multiple domains." School of Computer Networks Information and Forensics Conference, Edith Cowan University, 2004.
20. F. Amato, A. Mazzeo, A. Penta, and A. Picariello, "Building rdf ontologies from semi-structured legal documents," in *2008 International Conference on Complex, Intelligent and Software Intensive Systems*, 2008.
21. B. Schatz, G. Mohay, and A. Clark, "Rich event representation for computer forensics'," in *Proceedings of the Fifth Asia-Pacific Industrial Engineering and Management Systems Conference (APIEMS 2004)*, vol. 2, no. 12. Citeseer, 2004, pp. 1–16.
22. D. Kahvedžić and T. Kechadi, "Semantic modelling of digital forensic evidence," in *International Conference on Digital Forensics and Cyber Crime*. Springer, 2010, pp. 149–156.

Efficient Batch and Online Kernel Ridge Regression for Green Clouds

Bo-Wei Chen[1], Seungmin Rho[2†], Naveen Chilamkurti[3]

[1] School of Information Technology, Monash University, Australia
bo-wei.chen@monash.edu; dennisbwc@gmail.com
[2] Department of Media Software, Sungkyul University, South Korea
smrho@sungkyul.ac.kr
[3] Department of Computer Science and Computer Engineering, La Trobe University, Australia
n.chilamkurti@latrobe.edu.au

Abstract. This study presents an energy-economic approach for incremental/decremental learning based on kernel ridge regression, a frequently used regressor on clouds. To avoid reanalyzing the entire dataset when data change, the proposed mechanism supports incremental/decremental processing for both single and multiple samples (i.e., batch processing). Experimental results showed that the performance in accuracy of the proposed method remained as well as original design. Furthermore, training time was reduced. These findings thereby demonstrate the effectiveness of the proposed method.

1 Introduction

Ridge regression extends linear regression techniques, in which a ridge parameter is imposed on the objective function to regularize a model [1]. Such regularization uses $\mathcal{L}_2$ norm, or Euclidean distance, as the criterion to constrain the searching path of objective functions. Kernel Ridge Regression (KRR) further advances ridge regression by converting feature space into hyperspace with the use of kernel functions, for example, polynomial functions and radial basis functions. As one of the popular classifiers and regressors, however, KRR has more computational loads than Support Vector Machines (SVMs).

Although KRR has a closed-form solution, which involves the inverse of matrices, calculating these matrices degrades computational speed. Literature reviews [1] showed that the complexity of KRR [2] was as high as $O(N^3)$, whereas that of SVMs was $O(N^2)$, in which N represented the number of instances in data. Such a characteristic is a burden on cloud servers, which consume too much power for computation, not to mention incremental analysis.

Unlike typical learning, incremental analysis allows the system to add new training

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_40

samples. Additionally, no retraining is required. Both single incremental and multiple incremental mechanisms are conducive to relief of computational loads as earlier calculation results can be reserved for updating the new system in the future. Furthermore, if the size of data is far beyond the capability of one machine, especially when the memory space cannot accommodate the entire data at once, incremental analysis is a feasible solution.

In kernelized learning, many efforts have been devoted to incremental classifiers [1, 3-9]. For SVMs, Cauwenberghs and Poggio [3] established a milestone for kernelized learning as they discovered the equilibrium between old Lagrangian multipliers and new ones. A differential form was derived from the cost function of SVMs and the Karush–Kuhn–Tucker (KKT) conditions. Such a differential form supported single incremental and decremental learning. The model in [3] has inspired subsequent studies, for example [4], [7], and [8]. Karasuyama and Takeuchi [8] advanced the approach proposed by [3] and developed a strategy for multiple incremental/decremental learning. Their idea subsequently influenced the development of our work.

For KRR, incremental and decremental solutions become easier compared with those of SVMs because KRR has a closed-form solution. Recent works, such as [6], [9] and [1], were examples for single-instance incremental regressors. Based on kernel concepts, Engel *et al*. [6] developed a kernel recursive least squares algorithm, or incremental kernel regression. Their fundamental idea was equivalent to ordinary least squares (OLS) or linear least squares in statistics, but performed in hyperspace. The same algorithm was employed by Vaerenbergh *et al*. [9]. They furthered incremental kernel regression and integrated it into uncertainty analyses. In [1], rather than focusing on kernel regression, a recursive version of KRR was introduced. It used a single incremental mechanism to update weight vectors of the cost function. Moreover, a forgetting factor was integrated into the recursive form, where old and new training samples had different weights.

The contributions of this study are listed as follows.

- The proposed method supports decremental learning rather than merely incremental processing. Decremental learning is conducive to stabilization of the system by removing unnecessary outliers.
- Both incremental and decremental learning are capable of handling multiple samples in a batch.
- The proper size of a batch for incremental and decremental learning is derived in this article. Large datasets can be divided into several batches, which can be accommodated by a machine, for processing.

The rest of this paper is organized as follows. Section 2 introduces KRR, whereas Section 3 then describes the multiple incremental/decremental learning for KRR. Section 4 shows experimental results. Conclusions are finally drawn in Section 5.

2 Incremental/Decremental KRR

Let $\{(\mathbf{x}_i,y_i)|\ i = 1,\ldots,N\}$ denote a pair of an M-dimensional feature vector $\mathbf{x}_i$ and its corresponding label y_i, where i specifies the indices of N training samples. The objective of a linear regressor is to minimize the following cost function of least squares errors (LSEs).

$$\min_{\mathbf{u},b} E_{\mathrm{KRR}}(\mathbf{u},b) = \min_{\mathbf{u},b}\left\{\sum_{i=1}^{N}\left(\mathbf{u}^{\mathrm{T}}\phi(\mathbf{x}_i)+b-y_i\right)^2+\rho\|\mathbf{u}\|^2\right\} \tag{1}$$

where E_{KRR} is the cost function, $\mathbf{u}$ represents a J-by-1 weight vector, $\phi(\mathbf{x}_i)$ denotes the feature vector of $\mathbf{x}_i$ in the hyperspace, b is a bias term, and ρ specifies the ridge parameter. Besides, T means the conjugate operator, and $\|\cdot\|^2$ calculates the norm-two distance. Notable, J is the degree of the hyperspace when feature vectors are transformed by a kernel function.

Equation (1) can be rewritten as a matrix form, i.e.,

$$E_{\mathrm{KRR}}(\mathbf{u},b) = \left\|\mathbf{\Phi}^{\mathrm{T}}\mathbf{u}+b\mathbf{e}-\mathbf{y}\right\|^2+\rho\|\mathbf{u}\|^2. \tag{2}$$

Individually differentiating (16) with respect to $\mathbf{u}$ and b followed by zeroing both equations gives

$$\begin{aligned}\mathbf{u} &= \left(\mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}+\rho\mathbf{I}\right)^{-1}\mathbf{\Phi}\left(\mathbf{y}-b\mathbf{e}\right)\\ &= \mathbf{S}^{-1}\mathbf{\Phi}\left(\mathbf{y}-b\mathbf{e}\right)\end{aligned} \tag{3}$$

and

$$b = \frac{1}{N}\left(\mathbf{e}^{\mathrm{T}}\mathbf{y}-\mathbf{e}^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}}\mathbf{u}\right). \tag{4}$$

Unlike the solution to kernel regression, i.e., $\mathbf{u} = (\mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}})^{-1}\mathbf{\Phi}(\mathbf{y}-b\mathbf{e})$ where $\mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}$ could be singular, KRR avoids such a problem by adding a ridge term inside. The solution to (3) and (4) can be obtained by solving a system of linear equations.

$$\begin{bmatrix}\mathbf{u}\\ b\end{bmatrix} = \begin{bmatrix}\mathbf{S} & \mathbf{\Phi}\mathbf{e}\\ \mathbf{e}^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}} & N\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{\Phi}\mathbf{y}\\ \mathbf{e}^{\mathrm{T}}\mathbf{y}\end{bmatrix} \tag{5}$$

Based on the Schur complement theory,

$$\begin{aligned}\begin{bmatrix}\mathbf{S} & \boldsymbol{\Phi}\mathbf{e}\\ \mathbf{e}^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}} & N\end{bmatrix}^{-1} &= \begin{bmatrix}\mathbb{A} & \mathbb{U}\\ \mathbb{V} & \mathbb{N}\end{bmatrix}^{-1}\\ &= \begin{bmatrix}\left(\mathbb{A}-\mathbb{U}\mathbb{N}^{-1}\mathbb{V}\right)^{-1} & -\left(\mathbb{A}-\mathbb{U}\mathbb{N}^{-1}\mathbb{V}\right)^{-1}\mathbb{U}\mathbb{N}^{-1}\\ -\mathbb{N}^{-1}\mathbb{V}\left(\mathbb{A}-\mathbb{U}\mathbb{N}^{-1}\mathbb{V}\right)^{-1} & \mathbb{N}^{-1}\mathbb{V}\left(\mathbb{A}-\mathbb{U}\mathbb{N}^{-1}\mathbb{V}\right)^{-1}\mathbb{U}\mathbb{N}^{-1}+\mathbb{N}^{-1}\end{bmatrix}\end{aligned} \tag{6}$$

where

$$\begin{aligned}\left(\mathbb{A}-\mathbb{U}\mathbb{N}^{-1}\mathbb{V}\right)^{-1} &= \mathbb{A}^{-1}+\mathbb{A}^{-1}\mathbb{U}\left(\mathbb{N}-\mathbb{V}\mathbb{A}^{-1}\mathbb{U}\right)^{-1}\mathbb{V}\mathbb{A}^{-1}\\ &= \mathbf{S}^{-1}+\mathbf{S}^{-1}\boldsymbol{\Phi}\mathbf{e}\left(\mathbf{e}^{\mathrm{T}}\mathbf{e}-\mathbf{e}^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{S}^{-1}\boldsymbol{\Phi}\mathbf{e}\right)^{-1}\mathbf{e}^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{S}^{-1}\end{aligned}. \tag{7}$$

This form becomes useful in the following incremental and decremental processes.

3 Single Incremental and Decremental Processes

Single incremental and decremental processes are straightforward. Let $\mathbf{S}$ denote $\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} + \rho\mathbf{I}$. During the incremental phase, given a new training sample $(\mathbf{x}_c, y_c)$, the update of (3) becomes

$$\mathbf{u}[\ell+1] = \mathbf{S}[\ell+1]^{-1}\boldsymbol{\Phi}[\ell+1]\left(\mathbf{y}[\ell+1]-b[\ell+1]\mathbf{e}[\ell+1]\right) \tag{8}$$

and

$$b[\ell+1] = \frac{1}{N+1}\left(\mathbf{e}^{\mathrm{T}}[\ell+1]\mathbf{y}[\ell+1]-\mathbf{e}^{\mathrm{T}}[\ell+1]\boldsymbol{\Phi}^{\mathrm{T}}[\ell+1]\mathbf{u}[\ell+1]\right) \tag{9}$$

where

$$\begin{cases}\mathbf{S}[\ell+1]^{-1} = \left(\mathbf{S}[\ell]+\phi(\mathbf{x}_c)\phi(\mathbf{x}_c)^{\mathrm{T}}\right)^{-1}\\ \boldsymbol{\Phi}[\ell+1] = \left[\boldsymbol{\Phi}[\ell] \quad \phi(\mathbf{x}_c)\right]\\ \mathbf{y}^{\mathrm{T}}[\ell+1] = \left[\mathbf{y}^{\mathrm{T}}[\ell] \quad y_c\right]\end{cases}. \tag{10}$$

In (8)–(10), ℓ denotes the current state of the system. To save the computational time of $\mathbf{S}^{-1}$, the Sherman-Morrison formula and Woodbury matrix identity [10] indicate that the first equation of (10) can be converted into

$$\mathbf{S}\left[\ell+1\right]^{-1} = \mathbf{S}\left[\ell\right]^{-1} - \frac{\mathbf{S}\left[\ell\right]^{-1}\boldsymbol{\phi}\left(\mathbf{x}_c\right)\boldsymbol{\phi}\left(\mathbf{x}_c\right)^{\mathrm{T}}\mathbf{S}\left[\ell\right]^{-1}}{1+\boldsymbol{\phi}\left(\mathbf{x}_c\right)^{\mathrm{T}}\mathbf{S}\left[\ell\right]^{-1}\boldsymbol{\phi}\left(\mathbf{x}_c\right)}. \tag{11}$$

Regarding the decremental phase, given an index r of a sample, where $r \in \{1,\ldots,N\}$, a recursive form is created by considering the rth sample:

$$\mathbf{S}\left[\ell-1\right]^{-1} = \mathbf{S}\left[\ell\right]^{-1} + \frac{\mathbf{S}\left[\ell\right]^{-1}\boldsymbol{\phi}\left(\mathbf{x}_r\right)\boldsymbol{\phi}\left(\mathbf{x}_r\right)^{\mathrm{T}}\mathbf{S}\left[\ell\right]^{-1}}{1-\boldsymbol{\phi}\left(\mathbf{x}_r\right)^{\mathrm{T}}\mathbf{S}\left[\ell\right]^{-1}\boldsymbol{\phi}\left(\mathbf{x}_r\right)}. \tag{12}$$

For $\boldsymbol{\Phi}[\ell - 1]$ and $\mathbf{y}[\ell - 1]$, we can simply remove the corresponding column and row from $\boldsymbol{\Phi}[\ell]$ and $\mathbf{y}[\ell]$, respectively, to generate a recursive form.

4 Multiple Incremental and Decremental Processes

For multiple incremental and decremental processes, assume the system is about to add $|C|$ new samples and remove $|R|$ existing data. The operator $|\cdot|$ denotes the size. Then, (11) and (12) respectively become

$$\mathbf{S}\left[\ell+1\right]^{-1} = \mathbf{S}\left[\ell\right]^{-1} - \mathbf{S}\left[\ell\right]^{-1}\boldsymbol{\Phi}_C\left(\mathbf{I}+\boldsymbol{\Phi}_C^{\mathrm{T}}\mathbf{S}\left[\ell\right]^{-1}\boldsymbol{\Phi}_C\right)^{-1}\boldsymbol{\Phi}_C^{\mathrm{T}}\mathbf{S}\left[\ell\right]^{-1} \tag{13}$$

and

$$\mathbf{S}\left[\ell-1\right]^{-1} = \mathbf{S}\left[\ell\right]^{-1} + \mathbf{S}\left[\ell\right]^{-1}\boldsymbol{\Phi}_R\left(\mathbf{I}-\boldsymbol{\Phi}_R^{\mathrm{T}}\mathbf{S}\left[\ell\right]^{-1}\boldsymbol{\Phi}_R\right)^{-1}\boldsymbol{\Phi}_R^{\mathrm{T}}\mathbf{S}\left[\ell\right]^{-1}. \tag{14}$$

To facilitate multiple incremental and decremental processes at once, combination of (13) and (14) is necessary.

Let $\boldsymbol{\Phi}_{\mathcal{H}} = [\boldsymbol{\Phi}_C \mid \boldsymbol{\Phi}_R]$ represent the concatenation of all the column vectors in $\boldsymbol{\Phi}_C$ and $\boldsymbol{\Phi}_R$. Also denote $\boldsymbol{\Phi}'_{\mathcal{H}} = [\boldsymbol{\Phi}_C \mid -\boldsymbol{\Phi}_R]^{\mathrm{T}}$ as the concatenation of all the column vectors in $\boldsymbol{\Phi}_C$ and $-\boldsymbol{\Phi}_R$. Therefore, combination of (13) and (14) becomes

$$\begin{aligned}\mathbf{S}\left[\ell+1\right]^{-1} &= \left(\mathbf{S}\left[\ell\right]+\boldsymbol{\Phi}_C\boldsymbol{\Phi}_C^{\mathrm{T}}-\boldsymbol{\Phi}_R\boldsymbol{\Phi}_R^{\mathrm{T}}\right)^{-1}\\ &= \left(\mathbf{S}\left[\ell\right]+\boldsymbol{\Phi}_{\mathcal{H}}\boldsymbol{\Phi}'_{\mathcal{H}}\right)^{-1}\\ &= \mathbf{S}\left[\ell\right]^{-1} - \mathbf{S}\left[\ell\right]^{-1}\boldsymbol{\Phi}_{\mathcal{H}}\left(\mathbf{I}+\boldsymbol{\Phi}'_{\mathcal{H}}\mathbf{S}\left[\ell\right]^{-1}\boldsymbol{\Phi}_{\mathcal{H}}\right)^{-1}\boldsymbol{\Phi}'_{\mathcal{H}}\mathbf{S}\left[\ell\right]^{-1}.\end{aligned} \tag{15}$$

For $\mathbf{\Phi}[\ell + 1]$ and $\mathbf{y}[\ell + 1]$, $\mathbf{\Phi}[\ell]$ and $\mathbf{y}[\ell]$ simply remove the corresponding column(s) and row(s) from themselves, respectively. Subsequently, new samples are appended to the end of $\mathbf{\Phi}[\ell]$ and $\mathbf{y}[\ell]$ to generate $\mathbf{\Phi}[\ell + 1]$ and $\mathbf{y}[\ell + 1]$.

The batch sizes of $\mathbf{\Phi}_C$ and $\mathbf{\Phi}_R$, i.e., $|C|$ and $|R|$, can be different. Notably, the left-hand side of the two equations needs $O(J^3)$, whereas the inverse on the right-hand side requires $O(|C|^3)$ for (13) and $O(|R|^3)$ for (14), respectively [11]. To ensure performance, when the number of samples in a batch is smaller than the size of hyperspace features (i.e., $|C| < J$ or $|R| < J$), the system should perform an update. This implies a suitable batch size for time-series data, where new samples are rapidly generated and accumulated.

4 Experimental Results

Experiments on an open dataset were carried out for evaluating the performance. The information of the open dataset is listed in Table 1. The first column shows the name. The rest columns specify the number of classes, samples, and dimensions, respectively. Dataset "MIT/BIH ECG" is available at PhysioNet (www.physionet.org).

Table 1. Attributes of the dataset.

Name	#Classes	#Samples	#Dimensions
ECG	2	10403	21

The experiment used approximately 80% of the data for training and 20% of the data for testing. Furthermore, +4 and -2 samples were randomly selected for incremental and decremental learning at the same time. Table 2 summarizes the incremental and decremental settings. The system used Poly2 and Poly3 as kernels. The ridge parameter was 0.5.

Table 2. Settings of the incremental/decremental learning.

Name	Basic Training Size	Incremental/ Decremental Size
ECG	8323	+4 / -2

As the proposed algorithms generate the same optimal solutions as the original algorithm does, performance in accuracy remains unchanged. Thus, the metric for evaluating incremental and decremental algorithms is computation time. The accuracy for Poly2 was 94.47%, and that for Poly3 reached 94.14%.

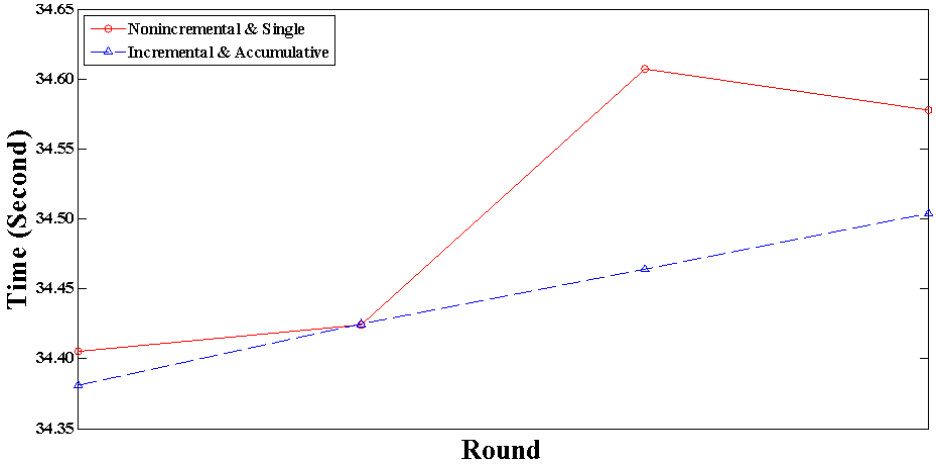


Fig. 1. Comparison between the incremental learning and the original algorithm with the use of Poly2 kernel. The dashed line uses accumulative time.

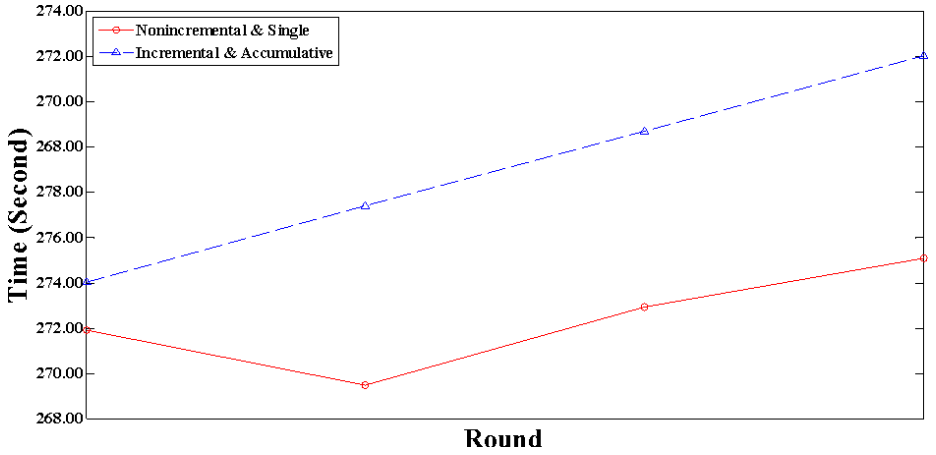


Fig. 2. Comparison between the incremental learning and the original algorithm with the use of Poly3 kernel. The dashed line uses accumulative time.

Figs. 1 and 2 display the multiple incremental/decremental results, where the horizontal axis is the round, and the vertical axis represents the computational time. The curve with "○" indicates the result by using the nonincremental algorithm, whereas "Δ" signifies the proposed algorithm. The solid lines are single-round time, and the dashed lines represent accumulative time since the initial training. Computational time all include conversion time from feature space to hyperspace. Examining these figures reveals that incremental and decremental mechanisms indeed saved the computational load. The numerical results in Tables 3 and 4 show that the proposed multiple incremental/decremental mechanism improved the efficiency by

850 and 161 times on average with the use of Poly2 and Poly3, respectively.

Table 3. Computational time by using Poly2 kernel

	Initial	Round 1	Round 2	Round 3	Round 4
Size	8323	+4 / -2	+4 / -2	+4 / -2	+4 / -2
Incremental	34.3413	0.0398	0.0436	0.0392	0.0397
Nonincremental	-	34.4050	34.4241	34.6070	34.5775

Table 4. Computational time by using Poly3 kernel

	Initial	Round 1	Round 2	Round 3	Round 4
Size	8323	+4 / -2	+4 / -2	+4 / -2	+4 / -2
Incremental	269.3652	1.6459	1.6940	1.6475	1.6564
Nonincremental	-	269.9599	268.7287	270.4560	271.5381

5 Conclusion

This work presents an efficient incremental/decremental mechanism for KRR. It can update the weight vector of the cost function without any retraining process. The proposed mechanism uses a recursive structure to replace the original computation of weight vectors. Such a structure allows the system to utilize earlier results to update the system when either single or multiple samples are changed. This mechanism is conducive to improvement of computational loads on cloud servers. This study also suggests an appropriate batch size during multiple incremental/decremental analyses. The Mathematical model shows that the size of each batch should be smaller than the feature dimensional size. Otherwise, it saves no computation.

An open dataset was used to evaluate the computational performance. Compared with the original algorithm without incremental/decremental learning mechanisms, the computational speed of the proposed method was enhanced by 850 and 161 times on average by using Poly2 and Poly3, respectively, far faster than the original algorithm. Such findings have established the effectiveness of the proposed incremental/decremental analyses.

Acknowledgments. This work is sponsored by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2013R1A1A2061978).

References

[1] S.-Y. Kung, *Kernel Methods and Machine Learning*. Cambridge, UK: Cambridge University Press, Jun. 2014.

[2] S.-Y. Kung and P.-Y. Wu, "On efficient learning and classification kernel methods," in *Proc. 2012 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2012)*, Kyoto, Japan, 2012, Mar. 25–30, pp. 2065–2068.

[3] G. Cauwenberghs and T. Poggio, "Incremental and decremental support vector machine learning," in *Proc. 14th Annual Conf. Neural Information Processing System (NIPS 2000)*, Denver, Colorado, United States, 2000, Nov. 28–30, pp. 409–415.

[4] C. P. Diehl and G. Cauwenberghs, "SVM incremental learning, adaptation and optimization," in *Proc. International Joint Conference on Neural Networks (IJCNN 2003)*, Portland, Oregon, 2003, Jul. 20–24, pp. 2685–2690.

[5] J. Kivinen, A. J. Smola, and R. C. Williamson, "Online learning with kernels," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2165–2176, Aug. 2004.

[6] Y. Engel, S. Mannor, and R. Meir, "The kernel recursive least-squares algorithm," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2275–2285, Aug. 2004.

[7] P. Laskov, C. Gehl, S. Krüger, and K.-R. Müller, "Incremental support vector learning: Analysis, implementation and applications," *Journal of Machine Learning Research*, vol. 7, pp. 1909–1936, 2006.

[8] M. Karasuyama and I. Takeuchi, "Multiple incremental decremental learning of support vector machines," *IEEE Transactions on Neural Networks*, vol. 21, no. 7, pp. 1048–1059, Jul. 2010.

[9] S. V. Vaerenbergh, M. Lázaro-Gredilla, and I. Santamaría, "Kernel recursive least-squares tracker for time-varying regression," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 8, pp. 1313–1326, Aug. 2012.

[10] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes 3rd Edition: The Art of Scientific Computing*, 3rd. ed. Cambridge, UK: Cambridge University Press, 2007.

[11] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge, MA, US: MIT Press, 2012.

Queuing-Oriented Job Optimizing Scheduling In Cloud Mapreduce

Ting-Qin He [1], Li-Jun Cai [1], Zi-Yun Deng [2], Tao Meng [1], XuAn Wang[3]
[1] College of Information Science and Engineering, Hunan University, Changsha, China
{hetingqin, ljcai, mengtao }@hnu.edu.cn
[2] College of Electrical and Information Engineering, Hunan University, Changsha, China
{dengziyun}@hnu.edu.cn
[3] Department of electronic technology, Engineering University of CAPF, Xi'an, China
wangxazjd@163.cn

Abstract. Cloud MapReduce, as an implementation of MapReduce framework on Cloud for big data analysis, is facing the unknown job makespan and long wait time problem, which have seriously affected the service quality. The Inefficient virtual machine allocation is one critical causing factor. Based on the M/M/1 model, a new queuing equation is built to ensure the virtual machine with the high efficiency. By jointing queuing equation and objectives function, a two variables equation group is designed to compute the desired virtual machine number for different jobs. According to the desired virtual machine number of each job, we developed a queuing-oriented job optimizing scheduling algorithm, called QTJS, to optimal job scheduling and enhance the resource utilization in Cloud MapReduce. Extensive experiments show that our QTJS algorithm consumes less job execution time and performs better efficiency than other three algorithms.

Keywords: map-reduce; cloud computing; queuing theory; job scheduling;

1 Introduction

With the continuous development of information technology, user need to process more and more data sets for knowledge discovering [1],[2]. In recent years, MapReduce [3], one of most popular distributed parallel programming model, has already been widely used on analysis of big data. Cloud MapReduce (CMR), one type of web service, is an implementation of MapReduce framework on Cloud [4],[5]. CMR integrates the MapReduce framework to the Cloud environment which consists a lot of distributed physical resources from multiple data centers in different geographical positions, CMR provides a public web service of MapReduce, the new platform plays a very important commercial role for individuals and enterprises.

The virtual machine optimization in CMR environment has already merged some achievements trying to handler the low working efficiency [6]. The existing work mainly focus on physical resource optimization, completion time optimization and energy optimization. Most scholars study focused on physical resources (CPU, memory, and storage) to improve resource utilization and increase data center profit.

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_41

More others focus on job costs and execution time scheduling, a plenty of recent work on traditional MapReduce job scheduling by using multiple different models and methods have been proposed [7],[8],[9],[10]. However, there are still need to further improve the job scheduling in CMR environment which is seldom considered in traditional optimal algorithms, a trouble situation is that users are still facing the fact to frequently login in CMR and query the job status. Some scholars attempt to work on the energy optimization through virtual machine distribution process scheduling. Literature [11],[12],[13],[14] address genetic, greedy, heuristic algorithm, FIFO, FAIR and Capacity Scheduler to consider energy optimization to some extents. However, they are mostly trying to reduce the number of physical hosts, and no consider the problem on CMR environment, when the scale of data sets is very large, the process time may reach several hours, even several days.

The motivation of this paper is to propose efficiency virtual machine allocation approach and job scheduling algorithm for cloud MapReduce environment. In the scheduling model, each data center is seen as a queuing system, virtual machine (VM) as the server, tasks as the waiter. One virtual machine usually only deals with single task at one moment, only the current task is completed, then the next task starts, so the computing process of each virtual machine satisfies the M/M/1 model. Based on the M/M/1 model, if the service rate of one virtual machine is higher than the arrival rate of task, then the tasks could finished in timely. Otherwise, some tasks which arrive later would be delayed. Therefore, when the service rate of one virtual machine is equal to the arrival rate of task, the virtual machine reaches the highest work efficiency, the job costs and job execution time would be optimal. In summary, based on the above positive connection between the service rate of one virtual machine and the arrival rate of task, it available to build a two element equation to dynamically compute the virtual machines set for jobs, so as to minimize job costs and execution time.

The other sections of this paper are organized as follows: section 2 presents an overview of our CMR Infrastructure. Section 3 presents the queuing-oriented job optimizing scheduling in CMR. Section 4 gives the performance evaluation of the proposed methods. Finally, we conclude the paper.

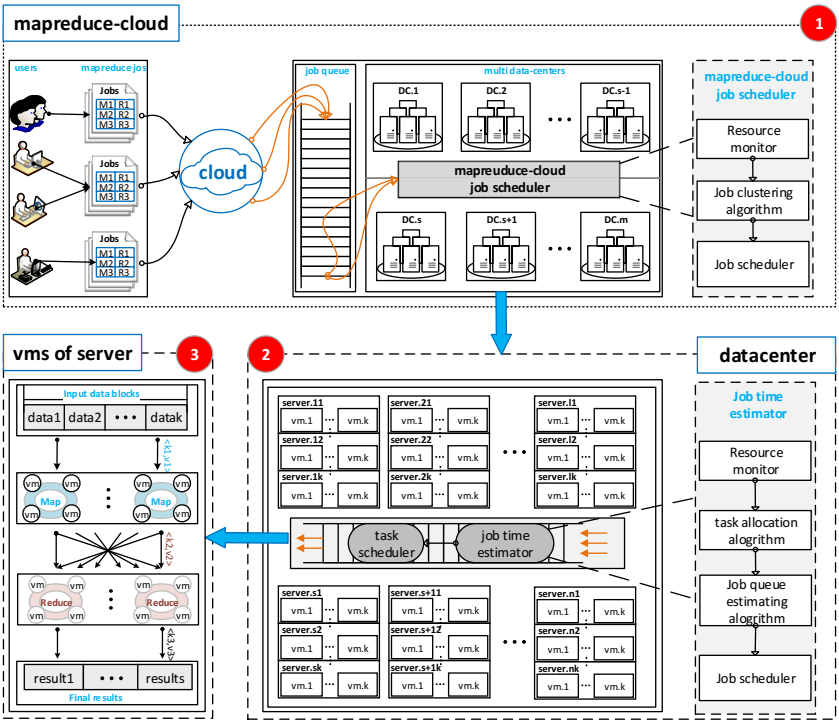
2 CMR Infrastructure

The CMR usually consists of multiple data centers, and integrates heterogeneous resources as the resource pool for computation. The CMR infrastructure is described in figure 1.

From the Fig.1, we can see that CMR is consists of four entities:

- **Job Requests**: each user may submit multiple jobs at one time, CMR receive various job requests in random. All jobs are independent with each other. Each job request contains three parameters: resource requirements, input data and compute functions (Map and Reduce functions).
- **Data Center**: Each data center is an independent resource pool which integrates underlying heterogeneous resources, including servers, networks, storages, and software. Each data center is considered as a job allocation unit

and waiting for the jobs from CMR. After jobs arrival, the available data center allocates sufficient VMs to execute Map and Reduce tasks of each job, and return results to user. Multiple distributed date centers in different geographical position compose A CMR. Multiple data centers are connected by the high-speed networks.



- **Fig. 1.** Infrastructure of Cloud MapReduce

- **Servers**: servers are the physical resources and responsible for the computation. In each data center, multiple heterogeneous servers consist of the resource pools. Each server has different resource capability properties, including CPU, RAM, I/O, and Bandwidth. The local high-speed network is bridge to connect multiple servers for data transmission. Generally, to compute large scale data set more elastically, each server is connected to the public large storage space, such as SAN.
- **VMs**: Virtual machine (VM) is as the computing unit of single task. When each job arrives to data center, it will be firstly split into multiple Map and Reduce tasks, and then allocated to multiple suitable VMs. Therefore, VM is the unit computing unit of data center resource pool. To deal with different resource requirement of each job, VMs usually are heterogeneous in data center.

According to the CMR infrastructure, the process of job scheduling is formulated as follows: multiple MapReduce job requests *JobS* are submitted to CMR job queue through online model. According to the properties of each job request, CMR sent each job request to job queue for evaluating the job cost *JC*. Based on the evaluation results of multiple data centers, CMR would send the job request *job* to the job queue of data center with best evaluation results. According to the "first come first served" way, each data center splits each job into three MapReduce tasks, and allocates to VMs for parallel executing. In the whole scheduling process, there are two level scheduling: cloud scheduling and data center scheduling. Cloud scheduling is only to dispatch each job request to job queue of data center according to the evaluation results. Data center scheduling is the core, it is responsible for selecting reasonable VMs to parallel execute the tasks using MapReduce model. The figure 2 shows the two level scheduling process of CMR, as we can see, all jobs requests in the cloud queue are waiting for execution firstly. Next, CMR sends job request to *job time estimator* of each data center to predict job execution results. Finally, each job is allocated to one data center based on the evaluation results.

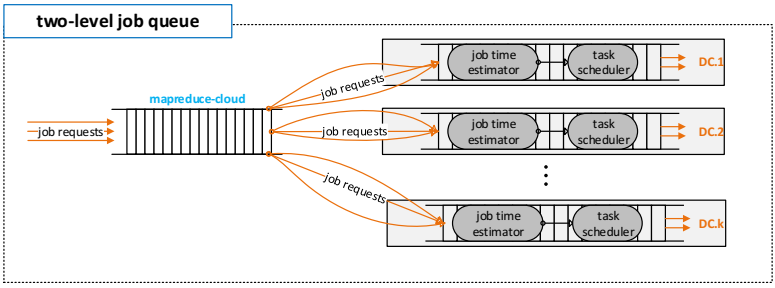


Fig. 2. Two Level Scheduling

3 Queuing-oriented Job Optimizing Scheduling in CMR

In this section, we give the scheduling modeling and propose the optimizing virtual machine allocation algorithm using queuing theory.

3.1 Scheduling problem and modeling

To build the scheduling model more clearly, we take an example to show the scheduling process: N job requests are submit to CMR and allocated to the same data center. Split Each job request job_i into M data trunk, M map tasks, and R reduce tasks. By using evaluation method, each job request job_i needs to wait for $JTmw_i$ millisecond before job execution, the all tasks are allocated to a group virtual machines set with the number of used virtual machine S, and the whole prediction execution time is $JTme_i$. In the virtual machines set, each virtual machine vm_j needs to

process P map tasks and P reduce tasks. The transmission time of $k\ th$ data trunk from public storage to virtual machine vm_j is TTD_{ijk}. The execution time of $k\ th$ map task on virtual machine vm_j is TTM_{ijk}. The intermediate data size of computing $k\ th$ map task on virtual machine vm_j is $idata_{ijk}$. The execution time of $k\ th$ reduce task on virtual machine vm_j is TTR_{ijk}. Otherwise, the rent cost of virtual machine vm_j is VC_j per millisecond, and the data transmission cost is DTC_j per megabyte data.

The whole task execution time of multiple tasks on one virtual machine TT_j is formulated as follows:

$$\begin{cases} TTS_{ijk} = TTD_{ijk} + TTM_{ijk} + TTR_{ijk} \\ TT_{ij} = \sum_{k=1}^{P} TTS_{ijk} \end{cases} \tag{1}$$

Where the whole task execution time contains three parts, including data transmission time, map execution time, and the last is reduce execution time. The term TTS is the single task execution time, the term P is the task number, and TT is the all task execution time of current job on the current virtual machine. So, the real execution time of job request job_i is defined as: $JT_i = \max\left(TT_{ij} \mid j \in 1,...,S\right)$, where the job real execution time is the maximum task execution time on one virtual machine because all virtual machines parallel deal with the multiple tasks of one job. Also, the job prediction time (makespan) is formulate as $JTm_i = JTmw_i + JTme_i$. After computing the job real execution time, the cost of job request job_i is formulated as follows:

$$JC_i = \sum_{j=1}^{S} VC_{ij} * TT_j + 64 * P * DTC_{ij} \tag{2}$$

In the equation 4, one job request contains two parts, respectively the rent cost and data transmission cost, where the term (64*P) is the data transmission size and the 64 is the default size of each data trunk. The model goals is to minimize the job cost and job execution time, so the goal function of this paper are defined as:

$$Min(JC_i * JT_i) \ \ Subject\ to: (i \in 1,...,N) \cap (\sum_{1 \le j \le S} P_j = M) \tag{3}$$

3.2 Using queuing theory for VMs allocation

In this section, we address queuing theory to optimal job scheduling and virtual machine allocation according to each job's attributes. A MapReduce job usually contains a data set and the number is M, a map task set and the number is M, a reduce task set and the number is R. To optimal job scheduling, we must decide the idea virtual machine set to compute current job for minimizing the job cost and job execution time, while the capacity of each virtual machine is larger than the job requirements. When the virtual machine set is less than requirements, the job cost could be decreased and the job execution time could be increased. When the virtual machine set is larger, the job cost could be increased and the job execution time could be decreased. In the CMR, each data center is a dynamic system with many heterogeneous virtual machines (compute node). Each virtual machine is considered as waiter, each task is seen as customer. The service process of each virtual machine is a single channel queuing process, which satisfies the M/M/1 model of queuing theory. In the M/M/1 model, when the arrival time of task is equal to the service time

of virtual machine, the resource utilization of virtual machine is the highest. Therefore, we try to ensure the arrival time of task equal to the service time of each virtual machine, improve the resource utilization of whole data center. Based on the above analysis, we assume that using S virtual machine to execute current job could get the best job cost and execution time. Then, each virtual machine would transmit M/S data trunk, compute M/S map task, and run M/S reduce task. In addition, we assume each virtual machine need to use TTD millisecond to transmit one data trunk, use TTM millisecond to compute one map task, and need TTR millisecond to run one reduce task. So, the single task execution time of each virtual machine is TTS, and equal to the sum of TTD, TTM and TTR. Therefore, the service speed of each virtual machine is $\mu = \frac{1}{TTS}$. Because of each virtual machine of S virtual machine set may be heterogeneous, so we assume that the whole job execution time is JT and equal to the maximal task execution time of virtual machine set. Due to all tasks must be completed before the newest task arrives, so the task arrival speed of current job is $\lambda = \frac{M/S}{JT - TTS}$. To assure each task is processed in real time, so the task arrival speed is less than the service speed of virtual machine, we could get following constraint:

$$\begin{aligned} &\because \ \lambda \le \mu \\ &\therefore \ \frac{M/S}{JT - TTS} \le \frac{1}{TTS} \\ &\therefore \ \frac{M}{S} \le \frac{1}{TTS}(JT - TTS) \\ &\therefore \ \left(1 + \frac{M}{S}\right) * TTS \le JT \end{aligned} \tag{4}$$

In the above constraint, when the arrival speed of task is equal to the service speed of virtual machine, the resource utilization gets highest value, so we can get an equation:

$$JT = \left(1 + \frac{M}{S}\right) * TTS \tag{5}$$

Where the connection of job execution time and used virtual machine number is cleared. With the number of virtual machine set increases, the job execution time would be decreased, otherwise, on the contrary.

In addition, based on the goal model of this paper (equation 3), a goal function is further defined to describe the goals of minimizing the job cost JC and job execution time JT, and is defined as:

$$f(JC, JT) = JC * JT = \left(\sum_{j=1}^{S} VC_j * TT_j + 64 * P * DTC_j\right) * JT \tag{6}$$

To remove the difference of virtual machine rent cost and data transmission cost, we use average cost to further improve the goal function:

$$f(JC, JT) = JC * JT = \left(S * JT * VC_{avg} + M * 64 * DTC_{avg}\right) * JT \tag{7}$$

Where the term VC_{avg} is the average rent cost of whole virtual machine set, and the term DTC_{avg} is the average transmission cost of whole virtual machine set. Otherwise, the M is the map task number of current job and 64M is the default size of one data trunk.

3.3 Optimizing queuing-oriented job scheduling

In the Section 3.2, two functions both have two variables: *S* and *JT*. Therefore, we could construct a two variables equation group for optimizing the job scheduling, and the equation group is defined as:

$$\begin{cases} JT = \left(1+\frac{M}{S}\right) * TTS \\ f(JC, JT) = \left(S * JT * VC_{avg} + M * 64 * DTC_{avg}\right) * JT \end{cases} \tag{8}$$

Based on the above equation group, we could further simply the goal function with only one variable S.

$$\begin{aligned} f(JC,JT) &= JC * JT \\ &= \left(S * JT * VC_{avg} + M * 64 * DTC_{avg}\right) * JT \\ &= S * VC_{avg} * \left(1+\frac{M}{S}\right)^2 * TTS^2 + M * 64 * DTC_{avg} * \left(1+\frac{M}{S}\right) * TTS \\ &= VC_{avg} * TTS^2 * S + \left(VC_{avg} * TTS^2 * M^2 + M^2 * 64 * DTC_{avg} * TTS\right) * \frac{1}{S} + \left(M * 64 * DTC_{avg} * TTS + VC_{avg} * TTS^2 * 2M\right) \\ &= C_1 * S + C_2 * \frac{1}{S} + C_3 \end{aligned} \tag{9}$$

$$\text{where, } \begin{cases} C_1 = VC_{avg} * TTS^2 \\ C_2 = \left(VC_{avg} * TTS^2 * M^2 + M^2 * 64 * DTC_{avg} * TTS\right) \\ C_3 = \left(M * 64 * DTC_{avg} * TTS + VC_{avg} * TTS^2 * 2M\right) \end{cases}$$

The goal is to minimize the job cost and job execution time, so it can consider the problem of getting the extreme value for the variable *S*. To get the extreme value, we need to find the derivative of goal function for the variable *S*.

$$\text{then, } f'(JC,JT) = C_1 - C_2 * \frac{1}{S^2} \tag{10}$$

$$\text{where, } \begin{cases} C_1 = VC_{avg} * TTS^2 \\ C_2 = \left(VC_{avg} * TTS^2 * M^2 + M^2 * 64 * DTC_{avg} * TTS\right) \end{cases}$$

When the function satisfies $f'(JC,JT) = 0$, it archives the minim value and the virtual machine number S is:

$$\because C_1 - C_2 * \frac{1}{S^2} = 0 \quad \therefore \rightarrow S = \sqrt{\frac{C_2}{C_1}} \tag{11}$$

$$\text{where, } \begin{cases} C_1 = VC_{avg} * TTS^2 \\ C_2 = \left(VC_{avg} * TTS^2 * M^2 + M^2 * 64 * DTC_{avg} * TTS\right) \end{cases}$$

Based on the equation 11, it can decide the best virtual machine allocation number considering the different job properties. Thus the process of QTJS algorithm consists of four steps. Firstly, when user submits jobs which contain three type of tasks, including an input data queue, a map task queue, a reduce task queue to CMR and wait for the scheduler command. Secondly, according to the job properties and status, resource monitor will search all virtual machine status and form the idle VM set, then build the scheduling goal function and set redundant coefficient and relative parameters. Thirdly, construct the two element equation to dynamically compute the

virtual machines set S according to equation 11. Finally, task scheduler allocates the tasks to virtual machines according to the scheduling results. The pseudo code is showed as algorithm 1.

Algorithm 1: queuing-oriented job optimizing scheduling

Input: (i)job request *job*, (ii)redundant coefficient γ ,
(iii) all virtual machine list of each data center *VL*.
Output: the final job scheduling results.
Procedure queuing_oriented_scheduling (*job*, γ ,*VL*)
// scheduling modeling
1: search all virtual machine idle status and form the idle list *VL*
2: initial *JTmw*, *JTme*, and *JTMr,* and set redundant coefficient and relative parameters.
3: for each $job_i \in jqueuelist$
build goal function $Min(JC_i * JT_i)$
// calculate the desired virtual machine number *S*
4: construct the two element equation and solve the optimal S (equation 11).
//job execution
5: while(virtual machine set is available)
Allocate the queuing tasks to *VL* and execute the jobs tasks based on S.
6: end while
7: end for
8: return the job results.

4 Experiments

As the purpose of the scheduling algorithm QTJS is to save the job waiting time, decrease the task complete time, and improve the resource utilization, this section we compare our proposed method to the traditional schedule algorithms like FIFO(First In First Out)[12], FAIR(Fair Scheduler)[13] and CS(Capacity Scheduler)[14]. Through measuring the jobs complete time, the performance of throughput and the resource utilization to evaluates the effects and the performance.

4.1 Experimental environment

We have implemented QTJS in Hadoop platform with version 1.20, the JDK version is 1.7. Due to hardware limits, we deployed QTJS in a compute cluster which consists of 9 compute nodes. In order to simulate heterogeneous environment, we selected four different type of host to build a heterogeneous hadoop cluster. The 9 physical nodes include 3 Dell OptiPlex3010 hosts, 2 Sugon A440-G hosts, 2 HP DL320E Gen8 hosts, and 2 HP ProLiant DL160 hosts, one of the Dell OptiPlex3010 hosts is a master node with large public storage, and others are respectively four different datacenters. Table 1 shows the details of the physical hosts, including the host type, amount, CPU and MEM information etc. All the nodes are connected through high-speed switch and communicate with others via 1Gb/s link, there are 8 compute nodes and 1 master node. To better simulate the jobs scheduling process, we extended the hadoop source code and customized six classes (CrmQuquer,

JobDataTransmission, TaskSynchronizer, JobTrainer, CloudConfig, VmGenerator), and modified JobTracker, TaskTracker and LaunchTaskAction three class on hadoop1.20, so as to achieve job scheduling optimization. Where, CrmQuquer is to build CMR job queue and plays the master node role, thus different granularity of computing could be employed. JobDataTransmission is responsible for data transfer. TaskSynchronizer collects status of each datacenter node to provide decision support. JobTrainer refers to test job and saving job history data. CloudConfig manages cloud scheduling information, VmGenerator generates heterogeneous virtual machines for each datacenter. A job trigger monitors job status on JobTracker.

Table 1. Physical machine parameter.

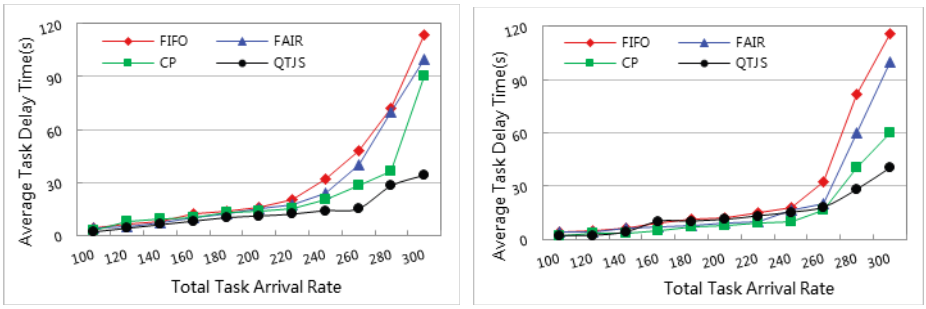
Type	Number	Cores	ARM(G)	Character
G1:Dell OptiPlex3010	1	4 CPUs,Core3.3GHz	4	Cloud queue
G1:Dell OptiPlex3010	2	4 CPUs,Core3.3GHz	4	DC1
G2:Sugon A440-G	2	8 CPUs,Xeon2.6GHz	8	DC2
G3:HP DL320E Gen8	2	16 CPUs,Xeon2.6GH	16	DC3
G4:HP ProLiant DL160	2	24CPUs,Xeon2.7GHz	24	DC4

The process of experiments is that when new jobs have been submitted, the input data will be firstly stored at public storage. When there are underlying physical resource available, the input data will be transited from public store to data node (HDFS), and the map and reduce tasks will be allocated to VMs at multiple datacenters, the final results will be written back to public storage. In additional, we chosen two classic benchmarks to evaluate the performance of QTJS: WordCount and Terasort. Two benchmarks are included in the Hadoop and usually used to evaluate the performance of MapReduce schedulers. For all experiments, every benchmark is running ten times and use the average value as the results.

4.2 Experimental Results

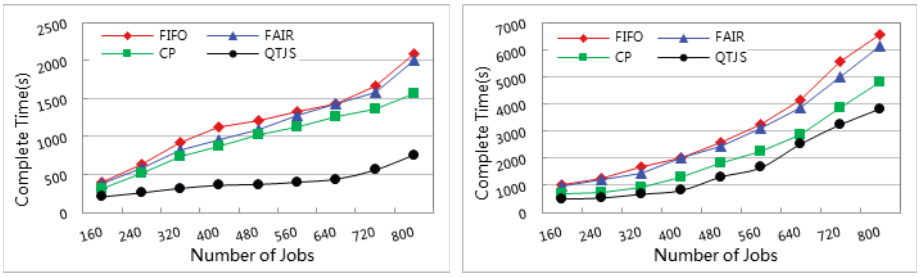
Throughout Performance. Throughout performance measures the task delay time, the task arrival rate is an essential factor of queuing theory, we run the two instances for ten times, using the average time as for comparing the task delay time with the task arrival rate varies. It is easy to find naturally similarity trend from figure 3, for both instances, the total task arrival speed rate varies from (100-300), four algorithm show the different task delay time, when the rate is less than 240, the four algorithm show the similar delay time, when the speed rate is over 240, our proposed algorithm causes lest delay time than other algorithm, FIFO is highest. As QTJS using queuing theory, it's suitable for virtual machine in CMR environment.

Average job complete time. The performance of our proposed algorithm are evaluated by comparing with other three traditional algorithm by running instances of WordCount, Terasort respectively. We plot the average complete time for the two kinds of jobs scheduled. As Figure 4 shows, when applied QTJS to WordCount and Terasort benchmark, the job complete time of FIFO performs worst, it takes much more time in all situation. FAIR and CP follows the average job complete time at two benchmarks jobs, QTJS is less than the other three algorithms.



(a) WordCount benchmark (b) Terasort benchmark

Fig. 3. Throughput performance



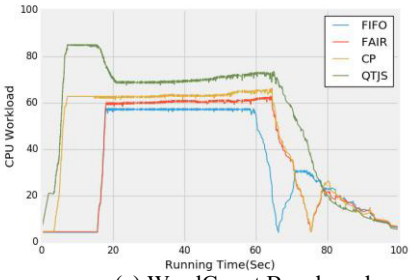
(a) WordCount benchmark (b) Terasort benchmark

Fig.4. Average job complete time

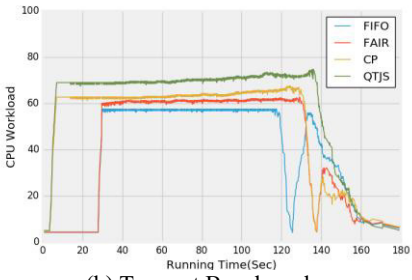
CPU Workload. Figure 5 shows the average CPU workload of four algorithms. As can be seen from the figure, for WordCount benchmark, with time going on (0-100), four algorithms perform different CPU workload fluctuations. When the time is within 20 seconds, all the four algorithm show high CPU workload, our QTJS show the highest especially, which means it's can more fully use physical host resources. As time runs to 80 seconds, the overall average CPU utilization decreases with the computation task decrease. At the final stage of job execution, the CPU Worklaod becomes similarity and decrease gradually. Such results can also be viewed on Terasort benchmark (0-180), the QTJS performs higher CPU workload, when time runs over than 140 seconds, and the same trend emerges and decrease gradually. The QTJS dynamic computer the virtual machine S, it helps to allocate the tasks to the S virtual machines on same physical host or similar host, such achieves the goal of increasing the resource utilization.

Network Workload. The WordCount and Terasort benchmark are executed for 10 times on network workload test. Figure 6 shows the comparison results of network workload on whole cluster physical hosts during shuffling intermediate data with the four scheduler. As we can see, QTJS shows highest network workload at the beginning time, with the time evolves, all the workload decreases and closes to similarity value. We know, in CMR scheduling process, it's need to transfer margin data at first, so the network is high, as time involves, the tasks are parallel executed in the VMs, the workload trend is becoming relatively smooth, from the tests, we know

that our proposed algorithm shows better network throughout, which helps to improving workload efficiency.

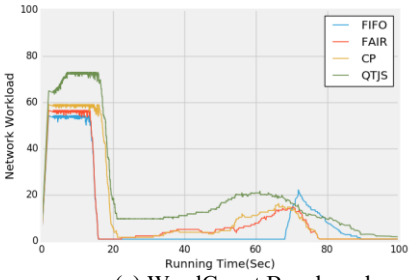


(a) WordCount Benchmark

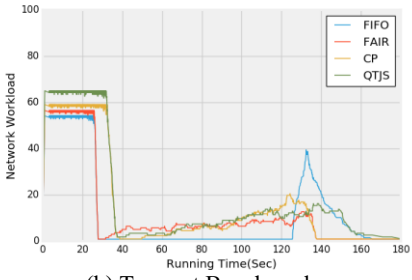


(b) Terasort Benchmark

Fig. 5. CPU Workload



(a) WordCount Benchmark



(b) Terasort Benchmark

Fig. 6. Network Workload

5 Conclusions

In this paper, we analyzed the necessity of the CMR job scheduling, and proposed QTJS algorithm to optimal virtual machines allocation and jobs scheduling. We model the computing process of each virtual machine which satisfies the M/M/1 model. Based on the queue theory, a two variables equation is built to compute the desired virtual machine number for different jobs, such for minimizing the job costs and job execution time, the experiments verify our proposed method contributing to ensure the virtual machine with the high workload efficiency.

As ongoing work, we plan to study the job makespan prediction to further optimal job scheduling, based on the two variables equation group, we already get the desired virtual machine number for different jobs, then through the prediction of job makespan to ensure the deadline and minimize the job costs.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (61174140, 61472127, 61272395).

References

1. J. Dean and S. Ghemawat, “MapReduce: Simplified data processing on large clusters,” in Proc. 6th Symp. Oper. Syst. Design Implementation (OSDI), Dec. 2004, pp. 137–150.
2. Thabet M, Boufaida M, Kordon F. An approach for developing an interoperability mechanism between cloud providers [J]. International Journal of Space-Based and Situated Computing 27, 2014, 4(2): 88-99.
3. Afrati F N, Ullman J D. Optimizing joins in a map-reduce environment[C]//Proceedings of the 13th International Conference on Extending Database Technology. ACM, 2010: 99-110.
4. Verma A, Cherkasova L, Campbell R H. Orchestrating an ensemble of MapReduce jobs for minimizing their makespan[J]. Dependable and Secure Computing, IEEE Transactions on, 2013, 10(5): 314-327.
5. Xilong Q, Peng X. An energy-efficient virtual machine scheduler based on CPU share-reclaiming policy [J]. International Journal of Grid and Utility Computing, 2015, 6(2): 113-120.
6. Hacker T J, Romero F, Nielsen J J. Secure live migration of parallel applications using container-based virtual machines[J]. International Journal of Space-Based and Situated Computing 1, 2012, 2(1): 45-57.
7. Hashem I A T, Yaqoob I, Anuar N B, et al. The rise of “big data” on cloud computing: review and open research issues [J]. Information Systems, 2015, 47: 98-115.
8. Manoharan M, Selvarajan S. An efficient methodology to improve service negotiation in cloud environment [J]. International Journal of Grid and Utility Computing, 2015, 6(3-4): 150-158.
9. Dahiphale D, Karve R, Vasilakos A V, et al. An advanced MapReduce: cloud MapReduce, enhancements and applications [J]. Network and Service Management, IEEE Transactions on, 2014, 11(1): 101-115.
10. Liu H, Orban D. Cloud MapReduce : a MapReduce implementation on top of a cloud operating system[C]//Proceedings of the 2011 11th IEEE/ACM international symposium on cluster, cloud and grid computing. IEEE Computer Society, 2011: 464-474.
11. Verma A, Kaushal S. Deadline constraint heuristic-based genetic algorithm for workflow scheduling in cloud [J]. International Journal of Grid and Utility Computing, 2014, 5(2): 96-106.
12. Verma A, Cho B, Zea N, et al. Breaking the MapReduce stage barrier [J]. Cluster computing, 2013, 16(1): 191-206.
13. Saez J C, Pousa A, Castro F, et al. ACFS: a completely fair scheduler for asymmetric single-isa multicore systems[C]//Proceedings of the 30th Annual ACM Symposium on Applied Computing. ACM, 2015: 2027-2032.
14. Shengjun X, Delong W, Suhong S. PCSP: A Preemptive Capacity Scheduler Policy for Scheduling Hadoop Jobs[J]. International Journal of Grid and Distributed Computing, 2015, 8(5): 33-46.

A Novel Cloud Computing Architecture Oriented Internet of Vehicles

He XU[1,2,a], Ye DING[1,2,b], Peng LI[1,2,c], Ruchuan WANG[1,2,d]
[1] School of Computer Science & Technology/School of software, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
[2] Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks, Nanjing, 210003, China
[a]xuhe@njupt.edu.cn,[b]2736036251@qq.com,[c]lipeng@njupt.edu.cn,[d]wagnrc@njupt.edu.cn

Abstract. Internet of Vehicles is one of the applications of the Internet of things technology in the field of intelligent transportation, and cloud computing is a new computing model which can provide services to users. The road safety requirements are very important in Internet of Vehicles, such as road accidents, motor vehicle violations of traffic signs, the driver drinking or fatigue driving and other issues which are needed to be timely warning and performing related processing. For this purpose, a novel cloud computing application framework is proposed, which can deal with the problem of road traffic safety in time, and solve the problem of real-time multimedia service and security authentication service on the vehicle terminal, making the Internet of Vehicles users to enjoy the network service at anytime and anywhere.

Key Words. Internet of Vehicles, Cloud Computing, Electric Power Cloud, Internet of Things

1 Introduction

Intelligent transportation is a transportation-oriented service based on present electronic information technology systems [1]. The prominent feature is to provide a diversity of services for traffic participants, which is based on the information collection, process and dissemination, exchange and analysis. Internet of vehicles technology is one of the key technologies of intelligent transportation systems. On August 27, 2013, "Internet of vehicles industry technology innovation strategic alliance" initiated by the Chinese automotive engineers society was formally established in Beijing. The alliance includes 15 vehicles companies. Its members covered vehicles manufacturers, mobile communication operators, manufacturers of hardware device, software providers and related research institutes. The alliance aims to unite forces together in related industries, cooperate and work together. In addition

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_42

to promoting in-car application service, they focused on enhancing performance of Internet of vehicles technologies for vehicles safety and economy. Internet companies are also concerned about Internet of Vehicles, Alibaba set up “Internet auto” foundation with SAIC company; Tencent released the “WeDrive” which is the whole solution of Internet of Vehicles following the publication telematics solution “Road PO Box”; “CarLife” is formally launched by Baidu, which takes the overall layout field of Internet of Vehicles. Experts predicted that China’s Internet market will reach 200 billion by 2020. Views about accelerating the construction of ecological civilization contain “developing distributed energy, building the smart grid, improving the operation and management system”, which is published On April 25, 2015 by the CPC Central Committee and the State Council. Vigorously developing energy-saving and new energy vehicles, improving innovation capability and industrial level, strengthening the construction of infrastructure, increasing the popularizing intensity", which indicates that China’s smart grid construction has once been pushed to new heights, the smart grid is the key technology which is critical to achieving new energy vehicles and telematics technology.

Cloud computing [2] is a new mode of computing technology presented in recent years[3-6], which brings the resources supply according the demand by adopting technologies of virtualization and distribution, parallel processing. With the development of mobile communication networks, mobile cloud computing has been widely studied and applied with solving the problem of the shortage of mobile terminal resource-constrained. Mobile cloud application aims to provide a seamless way giving mobile users a wealth of function applications, and Web services will not be affected by the resource constraints of mobile devices. Although the research is still in its infancy, it will become the mainstay mode of mobile applications in the future [7, 8]. Early in 2011, State Grid Information & Telecommunication Co.,Ltd. have launched a new project named “Power Cloud Simulation Laboratory Construction”, which is applied in Construction of the power of cloud applications. In China, State Grid Corporation carried out “power cloud simulation laboratory construction" project[9]. State Grid Corporation of smart grid strengthened the Smart Grid and the construction of the “Three Five” system, which creatively puts forward the power of cloud-building programs, covering electricity cloud platform, cloud electricity distributed database, electricity cloud resource virtualization management platform and typical applications based on cloud power operating platform.

Thus, there are many important applications significance for the study of electricity for vehicle networking cloud computing system. Researchers have also carried out related attempts.

Literature [10] proposed Ad Hoc network architecture which is suitable for vehicle networking. Literature [11] researched generic cloud network computing model of Ad Hoc [12] for Internet of vehicle. In connection with the problems of resource scheduling in the cloud computing, literature[13] research the task scheduling algorithm and virtual machine deployment algorithm. In order to achieve intelligent grid transmission, dispatch and storage, literature[14] applied the cloud computing technology into the large-scale smart grid management systems. Literature [15] designed the Internet of Vehicles mobile cloud systems which are based on cloudlet by combining Internet of Vehicles with mobile cloud computing, and built three-level

network model: center cloud, roadside cloud and car cloud. Literature [16] researched that users in the car how to use the mobile to control the terminal application through the cloud computing technology.

April 1, 2015, Ali cloud computing and PV system solution provider entered into a strategic cooperation agreement, together with connecting to Sunshine Power issued the wisdom of photovoltaic cloud "iSolarCloud". In the future, sun power will introduce the operation and maintenance of the solar power plant for nearly 20 years and design system capability and the product technology. Through the massive computing, storage and network connectivity provided by the alibaba computing, it provided intelligent operation and maintenance services to solar power plants [17].

In the Internet of vehicles environment, there are many equipments about car-terminals, such as GPS, wireless transceiver, radar, etc.. What's more, different vehicles will be equipped with different abilities (such as processor speed, memory capacity, memory size, etc.) for these different devices. Because general encryption and decryption algorithms of security mechanisms are required with a certain hardware resources, and on-board equipment normally has some hardware limitations, so it is difficult to apply the application of security in these heterogeneous devices. Therefore, researchers attended to researching the light-weight cloud computing model which is suitable for Internet of vehicles. Literature [18] proposed an Eco-Routing routing protocol which is used to prevent transferring malicious data in the Car Ad Hoc Network. Literature [19] proposed a cluster-based model of distributed dynamic public key infrastructure used in automotive networking, which is a good solution to the problem of key management vehicle network. Literature [20] designed collaboration cloud model which can be used to provide security services for the car networking, through defined traffic information and warnings as a cloud service. Literature [21] proposed security incentives for cloud vehicles. Literature [22] proposed anonymous authentication mechanism based on the cloud security service in the V2I vehicle network. It designed Vehicles and RSU local certificate revocation mechanism so as to reduce certificate revocation lists and revocation system service access token size. Literature [23] proposed an incentive TCBI (threshold credit-based incentive mechanism) based on the Gates credit for the car DTN network based on the cloud vehicle, interested car networking nodes making as large or great as possible and achieving fairly trust management of vehicle network data transmission.

In this paper, we will study and construct power cloud computing architecture oriented Internet of vehicles. The purposes are integrating fully cloud computing, power system resources, and data, storage and computing resources of the infrastructure of the internet of vehicles, enhancing grid scalability and security service capabilities, reducing infrastructure costs for internet of vehicles systems, grid systems and cloud computing hardware, and providing important technical support for achievement of internet of vehicles and applications of cloud computing.

2 Network Architecture Design

Internet of Vehicles cloud network systems architecture proposed in this paper is shown in Figure 1, which includes cloud servers, information collection terminals, charging stations, new energy electric cars.

Cloud servers: it mainly provides Internet-based cloud computing, cloud storage, cloud services. Such as power adjustment of schedules, traffic lights, multiple media information services, etc.

Information collection terminals: it collects information of passing cars, such as traffic information, status of drivers and the status of vehicles, etc.

Charging stations: it provides power to electric vehicles which need charging, and communicates with the vehicle terminal by wireless channel. The distance between the posts and communication network can cover all vehicles on the road between the two.

New energy electric cars: they are the cars which mainly use electric power. The mobile terminal systems consists of Android, IOS, embedded Linux systems.

In the network system, vehicles, information collection terminals and cloud servers, and charging stations, they can communicate each other. Cloud servers can make some dynamic adjusts on the signal lights based on traffic statistics collected by information collection terminals. Once there is an accident in the road, the system can inform the passing vehicles instantly so as to avoid the accident cars areas.

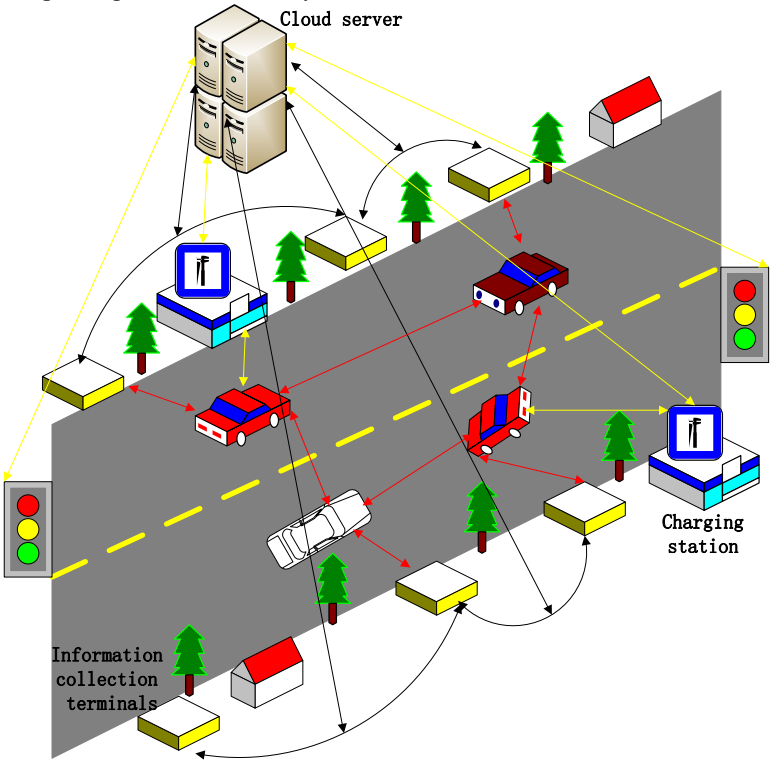


Figure 1. Network structure of cloud computing system oriented Internet of Vehicles

3 Model Design

In connection with the need of internet of vehicles, this paper provides power cloud architecture oriented internet of vehicles in Figure 2.

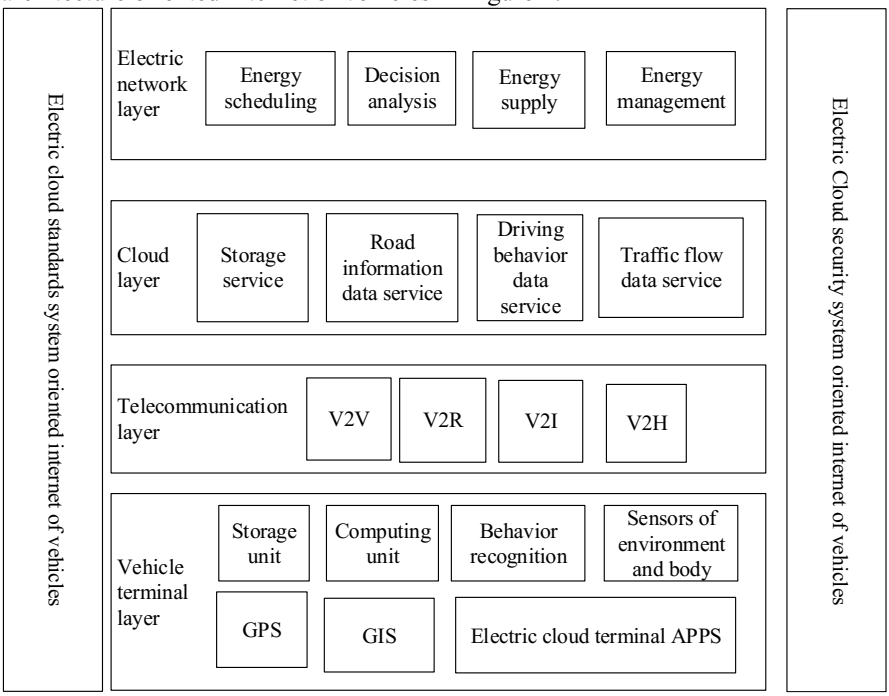


Figure 2. Power cloud architectures oriented internet of vehicles

Electric network layer: it mainly includes energy control, energy supply, energy management and decision analysis, etc.

Cloud layer: it mainly provides information storage service, road information data services, drivers' behavior and traffic flow data.

Communication layer: it includes communications services, such as vehicles with vehicles (V2V), vehicles and roads(V2R), and vehicles and instructions(V2I), vehicles and human(V2H) , etc.

Vehicle terminals layer: it includes storage, calculation modules, GPS, GIS, environment and the body sensor, behavior recognition and electric cloud App applications and other components. Electric Cloud App runs a variety of power systems and software applications related with internet of vehicles.

Electric cloud standards system oriented internet of vehicles: it mainly covers the standards of power devices of internet of vehicles, operational specification between the users of internet of vehicles and the power cloud computing systems, specification of telecommunication in the terminals of the internet of vehicles and integration specification of software resources of power cloud systems.

Electric Cloud security system oriented internet of vehicles: it mainly includes power cloud security system oriented with internet of power, organizational norms, security policies, etc.

4 Application Cases

At present, about 140 million cars have insurances in China, ranking second in the world. With the increase in vehicle, traffic order and security faces an even greater challenge, the development of the internet of vehicles is a smart and efficient direction. Telematics can make interactions between human, vehicles and environmental connections, and make traveling more efficient. Telematics for road traffic safety requirements are very significant to our daily life, such as road accidents, motor vehicle in violation of a traffic sign, the driver drinking alcohol or fatigue driving, which these warning and related matters are supposed to be carried out in a timely manner. The flows of power cloud applications of the internet of vehicles designed in this paper are as follows:

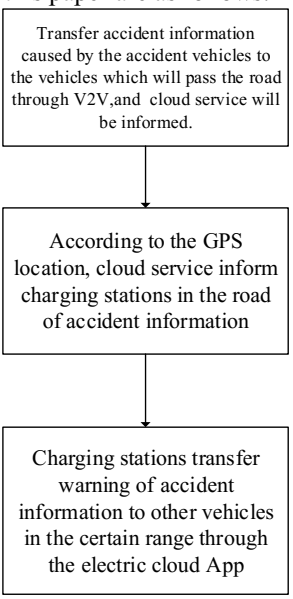


Figure 3. Accident in Road

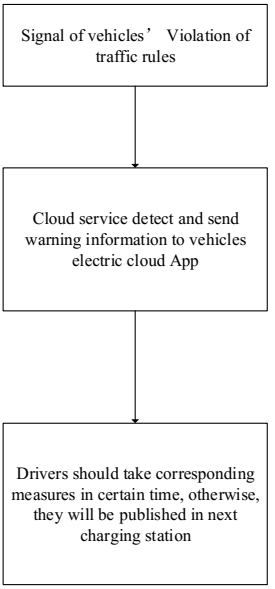


Figure 4. Vehicles' violation of Traffic Rules

After road traffic accident occurred, the Internet of vehicles system would quickly monitor the accident and notify the other upcoming vehicles via the road equipment so that they can ensure traffic safety or change the traffic route. Information circular flow system is shown in Figure 3 after road accident occurred. Signs of motor vehicles violating traffic ban would produce safety influence on the road and pedestrians, which caused that the Internet of vehicles environment need a good solution. Various types of sensors would be arranged surrounding roads and car terminals in the vehicle network environment. These sensors can detect driving behavior in real time. Once the vehicle operator violate traffic rules, power cloud App will promptly notify the traffic management department and remind drivers to regulate their driving behavior. The work flow is shown in Figure 4 after the motor vehicles violate traffic rules.

If drivers have fatigue driving and drunk driving, system will take certain measures. The processes are shown in Figure 5:

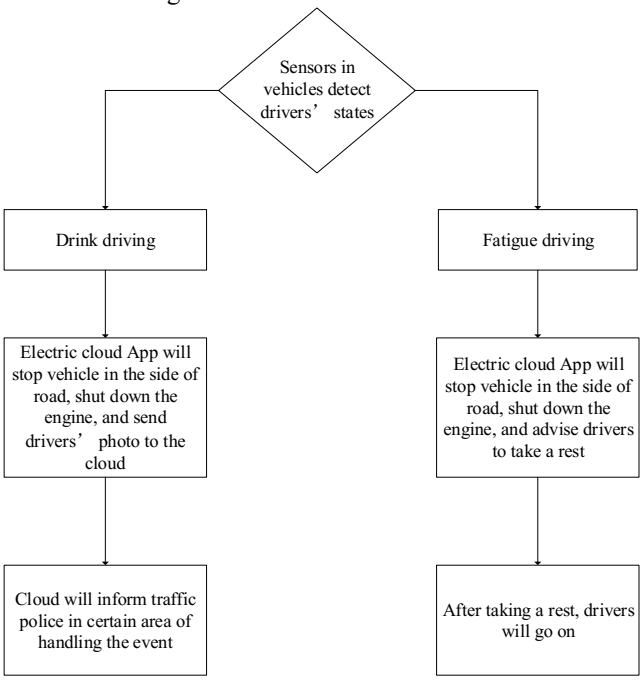


Figure5. Systematic Flow After drivers having fatigue driving or drunk driving

Multimedia Information Service in the existing Internet network is applied very widely. However, it hasn't been applied in the Internet of vehicles environment. In current road environment, a large number of mobile terminals (such as mobile phones, PAD, text reader, etc.) provide service through its own mobile communications network. If a plurality of mobile terminals use the same multimedia information service, the flow of information needs to be sent multiple times. If it is a video service, a lot of network bandwidth will be taken up, thus the network will be unavailable. We proposed a method which is suitable for multimedia information service in the Internet of vehicles environment. The method mainly uses peer computing technology, it considers the cloud storage server and vehicle terminal as the peer clients. What's more, it can also provide multimedia information services for vehicles roundly and exchange own information resource. Because its support from peer computing technology, cloud can maintain a list of information resources so that it can get a balance between servers in network and to ensure that the vehicle terminal can enjoy multimedia network services in real-time. Flow of multimedia information network service in Internet of vehicles is shown in Figure 6.

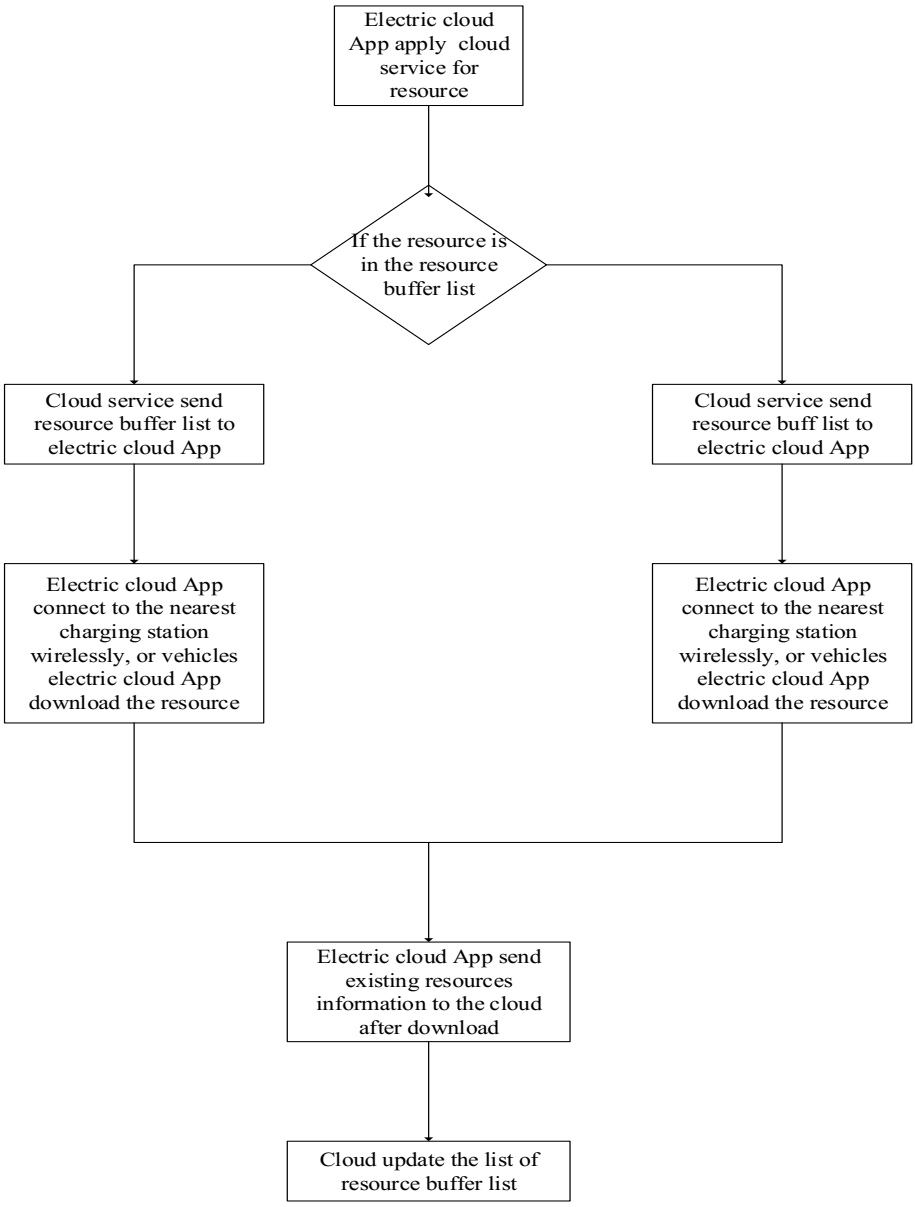


Figure 6. Flow of multimedia information network service in Internet of Vehicle environment

5 Security Mechanism

Under the circumstance of Internet of Vehicles, a large number of vehicles have moved from their own vehicle-registered network to another-unregistered network, in order to ensuring that the Internet of vehicles system will provide the safe and reliable service, we need to study the appropriate key management mechanisms. PKI(Public Key Infrastructure) is the infrastructure which uses the public key theory and technology to provide information security service[24]. PKI is the universal infrastructure which provides information security service. This system provides online authentication on the basic of unified safety certification standards and specifications, which is the collection of CA Certificates, Digital Certificates, Digital Signatures and related security application component modules. As a technical system, PKI can be used as the technical basis of Support Certification, Integrity, Confidentiality and non-repudiation, which solved the security issue of authentication, message integrity and non-repudiation technically and provide reliable security for network applications. Therefore, this paper considered power cloud security authentication scheme oriented Internet of Vehicles on the base of PKI. The specific work of this scheme is shown in Figure 7.

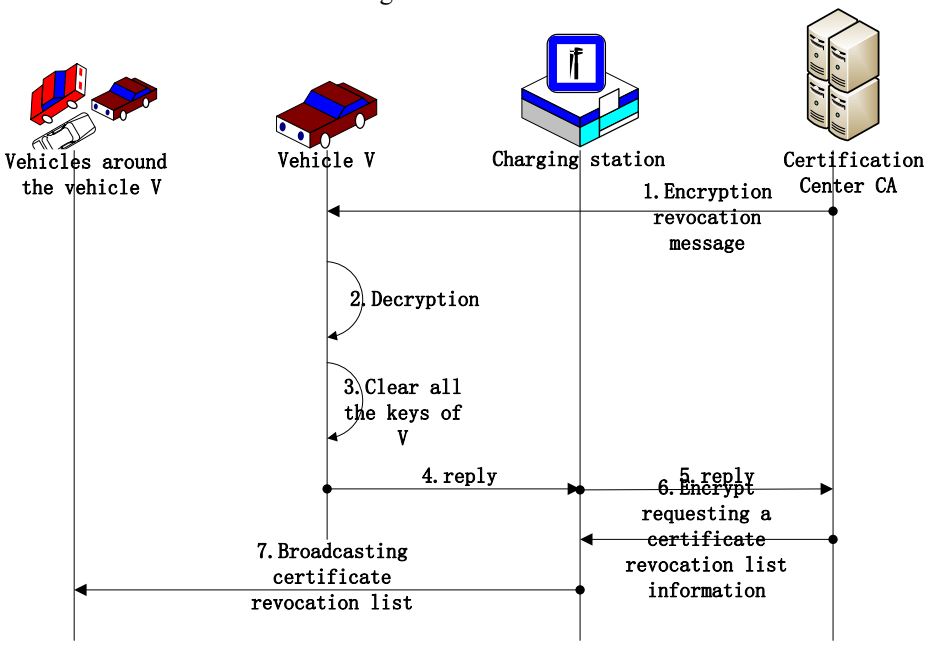


Figure 7. Flow of certificate revocation

Certification Authority (CA) publishes Keys: Public and private key of the vehicle are published in the CA, and the distribution of Key is in a certain period of time matching with an unique ID.

Authentication Key: The effectiveness of Public Key of the vehicle will be

certified by CA. When the vehicle i requests the Public Key to the CA center Cj, Pi will award the Public Key of this vehicle in Cj.Then Cj will produce a certificate on the basis of the Public Key of this vehicle and its own ID number, specifically: Ceri[Pi]=Pi|SignPrCj(Pi|Cj). Where Cer[Pi] is the Public Key that Cj published to the vehicle i, PrCj is the Private Key of CAj, the ID and (Pi|Cj) of Certification Center Cj will be signed with the Private Key.

Certificate Revocation: To protect the data of Internet of vehicles from attack, usually the expired certifications need to be revoked. The system mentioned above will revoke Certificate Key first, then build the certificate revocation list, broadcast and transport it to the vehicles around. Broadcast information is broadcast by charging station to other vehicles around. The specific processes of Certificate Revocation is shown as Figure 7.

6 Conclusions

The development of Internet of Vehicles provides a broad energy demand for power systems and cloud computing. Power cloud computing architecture oriented internet of vehicles combined the advantages of the distributed computing resources of cloud computing application, in order to provide the demand services to the users of Internet of Vehicles, which can solve the safety problems that exist in the convenient traffic and enable the demand of large bandwidth and high-volume to download on-board video under the circumstance of Internet of Vehicles. Meanwhile, we put forward that power cloud security authentication scheme oriented Internet of Vehicles on the base of PKI. Due to the constant evolve of cloud computing and Internet of Vehicles technology, problems with its corresponding scheduling algorithm of power system, the resource management of power cloud computing need to do in-depth study in our later work. Moreover, the safety research that adapted to the access mechanisms [25] of Power Cloud Computing on the basis of Internet of Vehicles is also very important.

Acknowledgments. This work is financially supported by the National Natural Science Foundation of P. R. China (No.61373017, No.61572260, No.61572261, No.61672296, No.61602261), the Natural Science Foundation of Jiangsu Province (No.BK20140886, No.BK20140888), Scientific & Technological Support Project of Jiangsu Province (No. BE2015702, BE2016185, No. BE2016777), Natural Science Key Fund for Colleges and Universities in Jiangsu Province (No.12KJA520002), China Postdoctoral Science Foundation (No.2014M551636, No.2014M561696), Jiangsu Planned Projects for Postdoctoral Research Funds (No.1302090B, No.1401005B), Natural Science Foundation of the Jiangsu Higher Education Institutions of China(Grant No. 14KJB520030), Jiangsu Postgraduate Scientific Research and Innovation Projects (SJLX15_0381, SJLX16_0326), Project of Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks(WSNLBZY201509), and NUPTSF(Grant No. NY214060, No. NY214061).

References

1. Barrero F, Guevara J A, Vargas E, et al. Networked transducers in intelligent transportation systems based on the IEEE 1451 standard[J]. Computer Standards & Interfaces. 2014, 36(2): 300-311.
2. Erl T, Puttini R, Mahmood Z. Cloud Computing: Concepts, Technology, & Architecture[M]. Pearson Education, 2013.
3. Yuriyama M, Kushida T. Integrated cloud computing environment with IT resources and sensor devices[J]. International Journal of Space-Based and Situated Computing, 2011, 1(2/3):163 - 173.
4. Kopetzky R, Nther M, Kryvinska N, et al. Strategic management of disruptive technologies: a practical framework in the context of voice services and of computing towards the cloud[J]. International Journal of Grid & Utility Computing, 2013, 4(4):47-59.
5. Shen Q, Yang Y, Wu Z, et al. Securing data services: a security architecture design for private storage cloud based on HDFS[J]. International Journal of Grid & Utility Computing, 2013, 4(4):242-254.
6. Ma K, Zhang L. Bookmarklet-triggered unified literature sharing services in the cloud[J]. International Journal of Grid & Utility Computing, 2014, 5(4):217-226.
7. Fernando N, Loke S W, Rahayu W. Mobile cloud computing: A survey[J]. Future Generation Computer Systems. 2013, 29(1): 84-106.
8. Dinh H T, Lee C, Niyato D, et al. A survey of mobile cloud computing: architecture, applications, and approaches[J]. Wireless communications and mobile computing. 2013, 13(18): 1587-1611.
9. Zhang Xiao. The power of cloud simulation laboratory project through the acceptance of China ICT companies [N].State Grid News,2011-02-11.
10. Zeadally S, Hunt R, Chen Y, et al. Vehicular ad hoc networks (VANETS): status, results, and challenges[J]. Telecommunication Systems. 2012, 50(4): 217-241.
11. Bitam S, Mellouk A, Zeadally S. VANET-cloud: a generic cloud computing model for vehicular Ad Hoc networks[J]. Wireless Communications, IEEE. 2015, 22(1): 96-102.
12. Perkins C E. Ad hoc networking[M]. Addison-Wesley Professional, 2008.
13. Huang Xiaoyuan. Cloud Platform Resources Scheduling Problem Research Based on Internet of Vehicles[D]. PhD Thesis, China: Dalian University of Technology, 2014.
14. Markovic D S, Zivkovic D, Branovic I, et al. Smart power grid and cloud computing[J]. Renewable and Sustainable Energy Reviews. 2013, 24: 566-577.
15. Xia Wenlong. Mobile Cloud-based Vehicular Networks Design And Key Technologies[D]. PhD Thesis, China: Guangdong University of Technology, 2014.
16. Nkenyereye L, Jang J W. Cloud Computing Enabled External Applications to Car Users using Nomadic Smartphones[C]. International Conference on Computer Information Systems and Industrial Applications, 2015. Atlantis Press, 2015.
17. PRESS. Cloud computing will be the driving force to drive the world [J]. Information System Engineering, 2015,5(20): 6-7.
18. Basaras P, Maglaras L, Katsaros D, et al. A Robust Eco-Routing Protocol Against Malicious Data in Vehicular Networks[J]. arXiv preprint arXiv:1506.05017. 2015.
19. Gazdar T, Benslimane A, Belghith A, et al. A secure cluster‐based architecture for certificates management in vehicular networks[J]. Security and Communication Networks. 2014, 7(3): 665-683.
20. Hussain R. Cooperation-aware VANET clouds: providing secure cloud services to vehicular ad hoc networks[J]. Journal of information processing systems. 2014, 10(1): 103-118.
21. Lim K, Abumuhfouz I M, Manivannan D. Secure Incentive-Based Architecture for Vehicular Cloud[M]. Ad-hoc, Mobile, and Wireless Networks, Springer, 2015, 361-374.

22. Park Y, Sur C, Rhee K. Pseudonymous authentication for secure V2I services in cloud-based vehicular networks[J]. Journal of Ambient Intelligence and Humanized Computing. 2015: 1-11.
23. Zhou J, Dong X, Cao Z, et al. Secure and Privacy Preserving Protocol for Cloud-based Vehicular DTNs[J]. IEEE Transactions on Information Forensics and Security. 2015, 10(6): 1299-1314.
24. Housley R, Polk T. Planning for PKI: best practices guide for deploying public key infrastructure[M]. John Wiley & Sons, Inc., 2001.
25. Yan G, Rawat D B, Bista B B. Towards secure vehicular clouds[C]. 2012 Sixth International Conference on Complex, Intelligent and Software Intensive Systems (CISIS), 2012. IEEE, 2012: 370-375.

Design and implementation of Data Center Network Global Congestion Price Calculating method based on Software Defined Networking

Chen Xiao Long[1], Shan Chun[2]
[1] School of Information Engineer ,Jinhua Polytechnic,Jinhua Zhejiang 321000
mmc518@126.com
[2] School of Accountancy of Guangdong Polytechnic Normal University, Guangzhou Guangdong 510665
xlycxl@21cn.com

Abstract. Aim the defect that the calculating algorithm for the link price is based on the proportion of the link state such as queuing length、queuing delay etc.Howerer the proportion is invariable and is unable to meet the dynamical character of data center network. Therefore, the figure that the controller of SDN architecture can grasp the global resource information will be made full use of , a kind of two layer network gloabal congestion price calculating algorithm is put forward.The upper algorithm is deployed in the opendaylight controller and give proportion paramater vector B according to the global resource information(including network topology、link state、queue length). The lower algorithm is deployed in OpenFlow switch, compute the link congestion price according to proportion paramater vector B gived by the upper algorithm..

Keywords： opendaylight; two layer algorithm ;proportion;congestion price

1 Introduction

Compared to the characteristics of Internet such as self-organizing, low bandwidth and high latency, the traffic of the data center network possess centralized, high bandwidth,low latency and high throughput, etc. Therefore, the traditional TCP protocol directly applied to the data center network face the shortage of functionality and performance, etc [1,2]。

Multiple source and multiple link network utility maximization framework model was established based on optimization theory in in [3].The idea of such distributed optimization theory framework is congestion pricing of the link is dynamically adjusted and the sending rates of source end adjutsted at the same time, when the system is in equilibrium, network utility achieve optimality.By far, end-to-end transmission control protocol (TCP) and the deformation various congestion control mechanism for the current data center network and the active queue

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_43

management algorithm in the link were contained in this distributed optimization theory framework. The key problem of Network congestion control are how to calculate congestion price based on the network load state and how to adjust sending window size according to network congestion pricing information. According to the characteristics different from Internet traffic of data center network, DCTCP[4]、ICTCP[5]、D3[6]、D2TCP[7]、PDQ[8] and DeTailDCTCP[9] was proposed to meet the need of data center network ,but these improved TCP is still under optimization theory framework of [3] and is still calculting the link congestion price according to the the proportion of the congestion link state such as queue length、queuing delay etc. and the proportion is invariable. And calculating the congestion pricing algorithm can be summarized as the following formula:

$$p = f(B,Q) \tag{1}$$

where B is proportion paramater vector，Q is queuing length,queuing delay status vector etc.

For example the very typical RED algorithm, which proportion paramater vector is $B = \{p_{\max}, b_{\min}, b_{\max}\}$， where $p_{\max}$ is the maximum packet drop probability，p is packet drop probability. $b_{\min}$ is the lower threshold of queue length and $b_{\max}$ is the upper threshold of queue length.It is well known that this RED algorithm possess the open problem is difficult to choose appropriate parameters vector B based on the actual link status.

In data center, the operator knows the network topology structure and the demand of core application . So they can significantly improve the utilization rate of network bandwidth. Google recently announced that the using SDN(soft defined network) technology can improve the bandwidth utilization for the core data center network. But the current research mainly focused on flow control about the network traffic load balancing based on the research of SDN architecture. Namely update the flow table in the openflow switch and forward the flow to the link which load is free[10]. But when the load of all link are heavy in the data center network ,the method to update flow table is powerless. So to rise the link congestion price and restrain send rate of the source end ,combinae with the traffic control again , is really a good way to improve transmission performance of data center network[10]. Therefore,we will make full use of the character that the controller of SDN architecture can grasp the global resource information[11], put forward a kind of two layer network gloabal congestion price calculating(GCPC) algorithm based on optimization theory framework of distributed network. The upper algorithm deploy in the central controller and the controller give proportion paramater vector B according to the global resource information(including network topology、link state、queue length). The lower algorithm deployment in OpenFlow switch, compute the link congestion price according proportion paramater vector B gived by the upper algorithm, And regularly feedback the current link status to the central controller.

2 Two layer Algorithm General Framework

The GCPC algorithm operation mechanism of the SDN frame structure is build as shown in the figure 1:

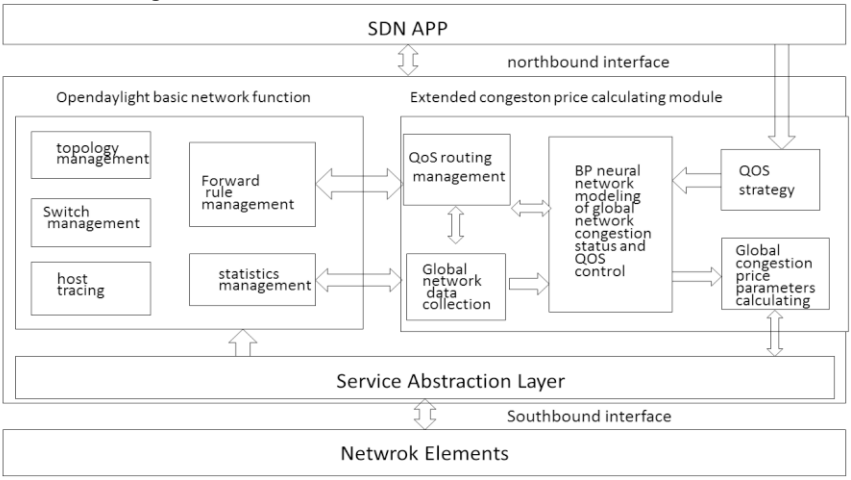


Fig.1.optimization control framework based on the SDN data center network resource management

The Opendaylight controller is adopted to realize the two layer network congestion price calculating algorithm. The control layer of OpenDaylight mainly include Service Abstrction Layer(SAL) and basic network service function.The basic netwrok function include topology management module, statistics management module, swithc management module, switches forward rule management module and host tracing module.At the same time ,the extended congestion price calculting module mainly include global network data collecting module,BP neural modeling of global network congestion status and global congestion price parameters calculting module.

In SDN network framework, the information in the openflow switch including new flow arriving and link congestion state will be feedbacked to the upper layer algorithm in center controller.The topology management module, statistics management module collect link state information online,and submitted thess link state statistical data to the global network data collecting module in the extended module，so the upper layer algorith in central congroller can understand real-time global network link state.When a fraction of link state are in heavy load,the central controller can update forward flow table of openflow switchs to realize link flow load balance,and the network transmission will not interrupt.Then it is not necessary to raise link congestion prices and to adjust the send rates of source TCP .So the upper layer algorithm can give less conservative proportion paramater vector B.Howerver when all link state are heavy load, it is powerless to update the forward flow table. So before all link state enter the heavy load state,the central control will give the no conservative proportion paramater vector B and raise the ling congestion prict to supress the send rate of the TCP source and to avoid network congestion.

3 Two layer congestion price algorithm

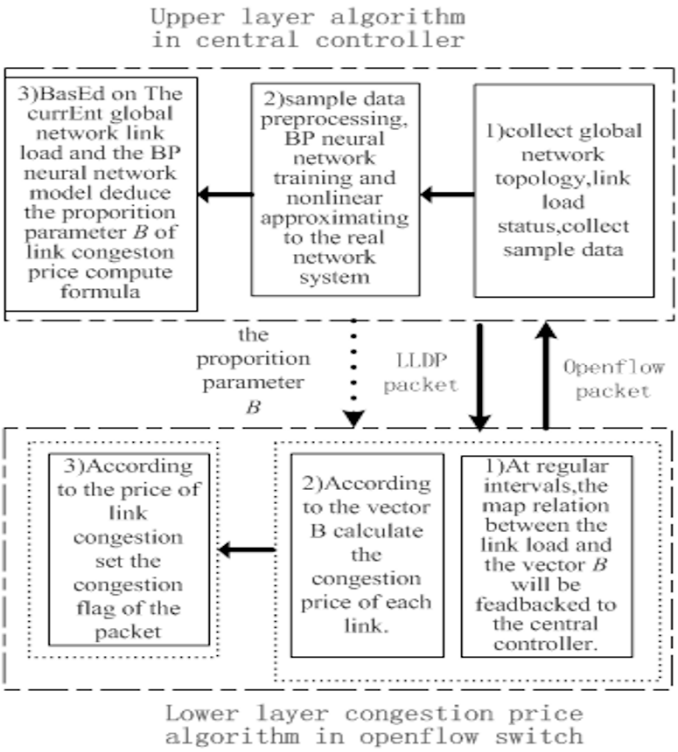


Fig.2. two layer global congestion price calculating algorithm

The core idea of the improved algorithm is that the upper layer algorithm in central controller give the proportion paramater vector B of switch link congestion price calculting algorithm by making full use of the global network link state. When part link state are not heavy ,the central controller can control the packet walk through the not heavy link and give less conservation proportion paramater vector B .Howerver when all link state are heavy load, it is powerless to update the forward flow table . So before all lind state enter the heavy load state,the central control will give the no loose proportion paramater vector B and raise the ling congestion prict to supress the send rate of the TCP source and to avoid network congestion.

3.1 Upper layer algorithm in central controller

1)Global link congestion state analysis

As shown in Fig.2, the upper layer algorithm in central Controller send regularly the packet of link layer discovery protocol(LLDP) to all switch equipment and collect the connection information of all switch equipment. Then the central controller can build the global network topology. By through OpenFlow channels between the central controller and switch, we modify exchange of three kinds of exchange packet include controller - to - switch, asynchronous and symmetric packets and add new exchange information to the data structure of these three kinds of exchange packets.The above modified information of data structure include the lind load $s_1, s_2, s...., s_n$,proportion paramater vector B for link congestion price calculating algorithm and practical running effect of switch in data center. The key informtion in the modified data structure is the mapping relationship between global network congestion state and the proportion paramater vector B of calculated congestion price parameter vector .

2) Algorithm initialization

At the beginning of the algorithm running, accordding to the openness and programming of the SDN framework, , the mapping relationship between the proportion paramater vector B of calculated congestion price parameter vector and global network congestion state are set by the network administrator according to experience. When part link state are not heavy ,the upper layer algorithm in central controller give less conservation proportion paramater vector B .Howerver when all link state are heavy load, ,the central control will give the no conservation proportion paramater vector B and raise the ling congestion price to supress the send rate of the TCP source and to avoid network congestion.

3) BP neural network parameter setting

As shown in Fig.2,In the upper layer algorithm of central controller,we design three layer BP neural network which the number of input layer number is n, the number of hidden layer neurons is m, the number of output layer is one. The each layer connection weights of BP neural network first randomly is initialized to the value between [0, 1].In the process of operation, the efficient mapping relationship data (including high utilization,good stability) will be collected as the samples of neural network training data.The input data of samples are the the lind load $s_1, s_2, s...., s_n$, the output data of samples is proportion paramater vector B.

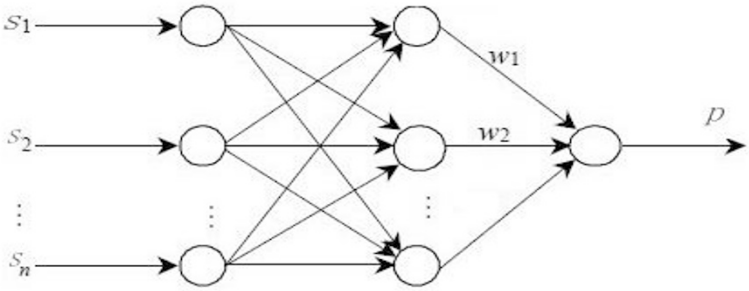


Fig .3. BP neural network model

Where $w_{i,j}$ is the connection weights between input layer and hidden layer. v_j is connection weights is the between hidden layer to output layer， θ_j is he threshold of hidden layer node, γ is the threshold of output layer node threshold $i = 1,...,n, j = 1,...,m$ .The transition function is sigmoid function as bellow:

$$f(x) = 1/(1+\exp(x)). \tag{2}$$

The output of hidden layer is as bellow:

$$y_j = f(\sum_{i=1}^{n} w_{i,j} s_i - \theta_j), j = 1,...,m \tag{3}$$

The value of output layer is as bellow:

$$X = f(\sum_{i=1}^{m} \nu_i y_i - \theta_i) \tag{4}$$

The error function is the square error function:

$$E = \frac{1}{2}(B - X)^2 \tag{5}$$

4) BP neural network training and modeling

The BP neural network's is training according to the sample data and get the optiomal connection weights and threshold. According to the ture link load of submitted by each switch and the trained of neural network model,the central controller can calculate the global network congestion status and give the suitable parameter vector B.

3.2 Lower layer algorithm in openflow switch

1)At regular intervals,the map relation between the current link load and the vector b will be feadbacked to the central controller and collected as the samples training data for BP neural network.

2) As shown in Fig.2,based on the vector B gived by the upper layer algorith,according to the $p = f(B, \mathrm{Q})$,calculate the congestion price of the each link,where Q is the current link congestion status.

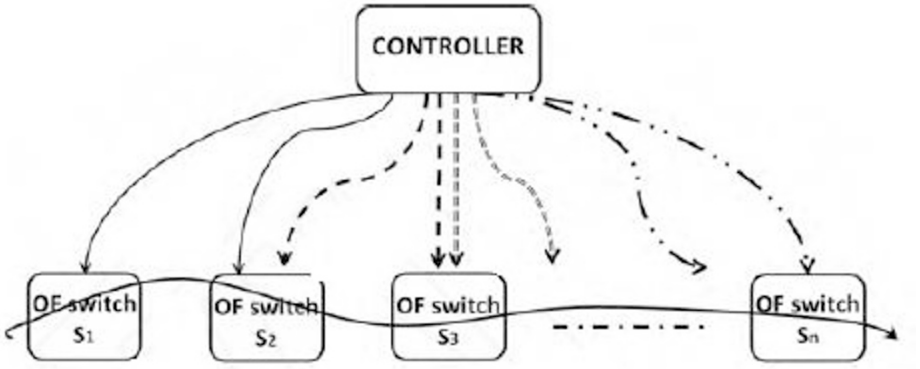


Fig.4. SDN neural network model

(3)As shown in Fig.4, standard TCP NEW Reno is used as the baseline of our evaluation.The initial window is 12KB,and openflow switches use RED queues and FCFS scheduling.The RED congestion price is calculated as follows:

$$p_i = p_{\max}(\frac{s_i - b_{\min}}{b_{\max} - b_{\min}}) \tag{6}$$

the calcultion result is shown in Table 1.

Table 1. the map relation between the load and the parameter B

Num.	s1	s2	s3	s4	load	*p*max	*b*min	*b*max
1	0.2	0.3	0.2	0.3	lighter	0.01	0.1	0.8
2	0.4	0.5	0.7	0.6	light	0.05	0.1	0.8
3	0.7	0.6	0.8	0.5	low heavy	0.08	0.1	0.7
4	0.8	0.9	0.7	0.6	heavy	0.1	0.1	0.7

4 Conclusion

We present the design and evaluation of two layer algorithm.An efficacious approach that give proportion paramater vector B according to the global resource information(including network topology、link state、queue length) and improve

the network throughput. Experimental results demontrate that it remarkly improves the networkthroughput up to morke 80% of bisection bandwidth .In future work,we will implement GCPC in real datacenter networks for further research.

Acknowledgments

This work is suppourted by the Public welfare research project of Guangdong Province under Grant No. 2014A010103031.

References

1. HuaWei Huang,Song Guo,Peng Li. Cost Minimization for Rule Caching in Software Defined Networking[J] IEEE Transactions on Parallel and Distributed Systems.2015.12
2. Krister Jacobsson,Lachlan L,Steven H.Low. An Improved Link Model for Window Flow Control and Its Application to FAST TCP[J].IEEE Transactions on Automatic Control,2009,54(3):551-564.
3. H. Kim, N. Feamster. Improving Network Management with Software Defined Networking. IEEE Communications Magazine, Februry 2013: 114-119.
4. ALIZADEH M, GREENBERG A, MALTZ D A, et al. Data center TCP (DCTCP)[C]. Proc of SIGCOMM'10, New Delhi, India, 2010.
5 VAMANAN B, HASAN J, VIJAYKUMAR T N. Deadline-aware datacenter TCP (D2TCP)[C]. Proc SIGCOMM'12 , Helsinki, Finland,2012.
6 WU H, FENG Z,GUO C, et al. ICTCP: incast congestion control for TCP in data center networks[C]. ACM CoNEXT'10, Philadelphia, USA, 2010.
7 WILSON C, BALLANI H, KARAGIANNIS T, et al. Better never than late: meeting deadlines in datacenter networks[C]. Proc SIGCOMM'11, Toronto, Ontario, Canada, 2011.
8 HONG C, CAESAR M, GODFREY P B. Finishing flows quickly with preemptive scheduling[C]. Proc of SIGCOMM'12, Helsinki, Finland,2012.
9 ZATS D, DAS T, MOHAN P, et al. DeTail: reducing the flow completion time tail in datacenter networks[C]. Proc of SIGCOMM'12 Helsinki, inland,2012.
10 R. Soule, S. Basu, P. J. Marandi, et al. Merlin: A Language for Provisioning Network Resources.International Conference on emerging Networking Experiments and Technologies (CoNEXT),Sydney: ACM Press, 2014.
11 A. Gupta, L. Vanbever, M. Shahbaz, et al. SDX: A Software Defined Internet Exchange. Proc. Of the SICOMM, Chicago: ACM Press, 2014.Appendix: Checklist of Items to be Sent to Conference Proceedings Editors (see instructions at conference webpage)

Business Process Mining based Insider Threat Detection System

Taiming-ZHU[1,2], Yuanbo-GUO[1,2], Jun-MA[1,2], Ankang-JU[1,2]
[1] Information Engineering University, Zhengzhou, China
[2] State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou, China
ryanztm@126.com

Abstract. This paper introduces a Business Process Mining Based Insider Threat Detection System. The system firstly establishes the normal profiles of business activities and the operators by mining event logs, and then detects specific anomalies by comparing the content and the order of execution logs with the corresponding normal profile in order to find out the insiders and the threats they have brought. The anomalies concerned are defined and the corresponding detection algorithms are presented. We have performed experimentation using the ProM framework and Java programming with five synthetic business cases, and found that the system can effectively identify anomalies of both operators and business activities that may be indicative of potential insider threat.

1 Introduction

The insider threat is a long-term problem that faced by most organizations. It usually results in significant damage and could range from financial theft and intellectual property theft to the destruction of property and business process. The main causes of insider threats are as follows: First, part of employees may lack security awareness and violate the safety regulations by accident. Second, part of employees intentionally bypasses the security measures for their own convenience and efficiency in the works. Last but not least, some employees choose to leak the organization's confidential information or sabotage the systems because of their resentment or other's inducement. In general, insider threat a comprehensive problem, which consists of human factors and systemic factors.

In nowadays, more and more organizations begin to use various business systems to accomplish business activities for higher efficiency. However, it could also bring many chances to spiteful insiders due to the lack of design on system security. Therefore, we see this problem from the perspective of business activities and try to detect insider threat by a comprehensive analysis of operators' abnormal behavior and anomalies emerged during business process execution.

Since the actual business activity involves many factors, its process model must also be multidimensional, not only to reflect the sequence between business events, but also to reflect the behavior information of operators, the features of business cases and the time and frequency information of business events. To establish a good model,

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_44

most organizations turn to the log-based process mining method. As far as we know, many of the current process mining methods establish a business process model only from the perspective of control-flow [8-12]. In this paper, we present a business process mining based insider threat detection system shown in Figure 1, which includes the operators' behavior into business process mining. The system mines normal business process models from the logs generated during the execution of business activities under a normal environment, and then detects anomalous behavior of operators, logical anomalies and performance anomalies in execution logs by comparing them with the corresponding models. To test the performance of the approach, the system was implemented in ProM framework [14] and Java Programming and the experimentation was conducted with synthetic data for evaluation. It was found that the system performed well for detecting the insider threats.

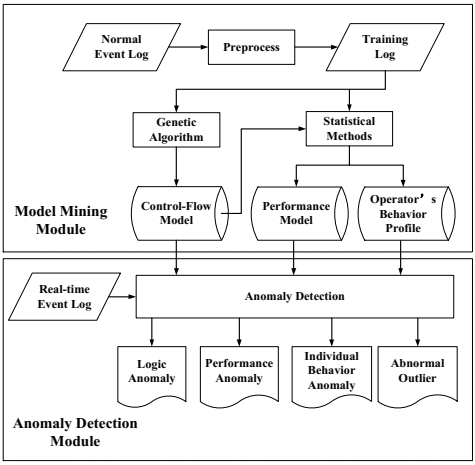


Fig. 1. Business Process Mining Based Insider Threat Detection System

The remainder of this paper is organized as follows. The related works are introduced in Section 2. Section 3 briefly introduces the presented insider threat detection system and the details of each components are described in Section 4. Section 5 presents the process of conducting experimentation and discusses the results, and Section 6 concludes this paper.

2 Related Work

In this section, we introduce the related work in two research areas: insider threat detection and business process mining.

Insider threat has always been an important factor in harming the security of organizations and enterprises. Early in 2000, Anderson et al. [1] presented a variety of challenging problems in researching this area, which further induced attention and exploration of insider threats. From then on, researchers have proposed a variety of specific methods to detect insider threats. For instance, Spitzner et al. [2] try to decoy insider attacks using a honeypot. However, with the way of attack becoming more

subtle and advanced, people have begun to seek more sophisticated methods. In [3], the author use role-based access control model to establish user behavior rules and detect insider threats by looking for behavior that violates the rules. The authors in [4] extend the RBAC model by focusing on generalized attributes of people and data, and placing the insider threat in the context of modeling policies. Moreover, Greitzer and Frincke [5] combine traditional cyber security audit data with psychosocial data, so that people can detect and predict insider threats. Similarly, Brdiczka et al. [6] present an insider threat detection method that combines structural anomaly (SA) detection with personal psychology (PP), and identifies insider threats by integrating SA and PP and sorting the results of integration. Parveen et al. [7] use stream mining and graph mining to detect insider threat and extend it in [8]. Unlike these approaches, the detection system proposed in this paper not only considers the anomalous behavior of insiders, but also detects anomalies in business activities to help managers to evaluate the influence to the business system.

Business process mining is a business process reconstruction technique developed from the workflow management area. Many researchers have proposed relating mining algorithms. It is worth mentioning that Aalst et al. [9] point out a complete business process modeling process should consider three perspectives, i.e. the control-flow perspective, the organization perspective and the case perspective. The control-flow perspective focuses on the ordering of activities. The organization perspective focuses on the originator field. The case perspective focuses on properties of cases. Most existing process mining methods only consider the control-flow perspective. For example, the α-algorithm proposed in [10] builds control-flow structure by discovering the binary relations between activities, but cannot capture the dependency in non-free-choice constructs. The α++-algorithm [11] extend the α-algorithm but cannot deal with noisy logs. The genetic mining algorithm [12] can overcome the limitations of current process mining techniques and it uses global optimization strategies to ensure the outcome is global optimal, which is better than local optimal in process modeling. In this paper, we first use the genetic mining algorithm to discover the business control-flow model, and then mine the business performance model and operators' behavior profile based on the control-flow model and the information in business logs in order to detect possible insider threats in business activities.

3 System Design and Implementation

3.1 Log Preprocessing

An event log generated during the execution of business activities contains many events, each event consists of the identifier of the business activity it belongs to, the task name, the timestamp, the operator, the operations and the changes of data or file etc. As the data source of process mining, the completeness and correctness of the event logs are directly related to the accuracy of the mining results, so that it is necessary to preprocess the event logs before using it. We will give some formal definitions of the relevant concepts and then describe the process of log preprocessing in the following part.

Definition 1：Event

An Event E is a sextuple (Id, TaskName, Type, Operator, Timestamp, ExtraInfo), where Id is an identifier of the kind of business activity, TaskName is the name of the task in this business activity, Type is the execution state of this task, Timestamp is the time the event is recorded, ExtraInfo is some other relevant information about this task.

Definition 2：Event Sequence

An Event Sequence ES is a sequence of events recorded in an execution of a business activity, i.e. $ES = (E_1, E_2, \ldots E_n)$, where E_1 and E_n are the starting and ending event of ES respectively, and any two events in ES satisfy the following condition:

$$\forall E_i, E_j \in ES\ (i<j),\ E_i.\text{Timestamp} \leq E_j.\text{Timestamp} \quad (1)$$

Definition 3：Event Log

An Event Log EL is a set of several event sequences, i.e. $EL = (ES_1, ES_2, \ldots ES_n)$.

The first step of preprocessing is to have an overall understanding of the log's statistic information, like the number of different business activities, the occurrence number and the ratio of each kind of event in the event sequence and so on. Secondly, we remove the irrelevant event sequence according to the Id. Then, the illegal starting event and ending event are determined. Finally, choosing the event sequences that have correct starting event and ending event and each intermediate event is completed to be the input of process mining algorithm. The preprocessed log is called training log in this paper.

3.2 Business Process Mining

Since the complex structure may result in different event sequences of a same process model, so the model mined from a certain event sequence is just a particular solution of all the possibilities. To enhance the robustness of the mined model, we should collect enough different event sequences of the business activities and select the appropriate mining algorithm. After that, we can further mine the performance model and each operator's behavior profile by using the control-flow model and the information in training logs.

Mining Process Model in Control-Flow Perspective

We can consider process mining as a search for the most appropriate process out of the search space of candidate process models. In order to get an optimal process, selecting an appropriate algorithm is very important. Usually, people can take global strategies or local strategies to measure the degree of optimization of a model [12]. The common drawback of local strategies is that they cannot guarantee the outcome of the locally optimal steps will result in a globally optimal process model. In contrast, global strategies primarily based on a one strike search for the optimal model.

The genetic algorithm is a typical example. In addition, compared with traditional methods, genetic algorithm can solve more problems, such as non-free-choice, invisible tasks and duplicate tasks. Therefore, we use it for process mining, which is roughly divided into four phases: initialization, selection, propagation and

termination, which is shown in Figure 2.

After getting the control-flow model, we need to determine the events or sub-processes that have specific time or frequency constraints based on expert knowledge. Then we could obtain the statistical information about these events or sub-processes through relevant fields in training logs, such as the time interval between events, the timestamp and the occurrence number of specific events etc. Due to the slight normal variation of the above indicators in each event sequence, we create a multi-set for each indicator in order to calculate the average value, which makes up the performance model.

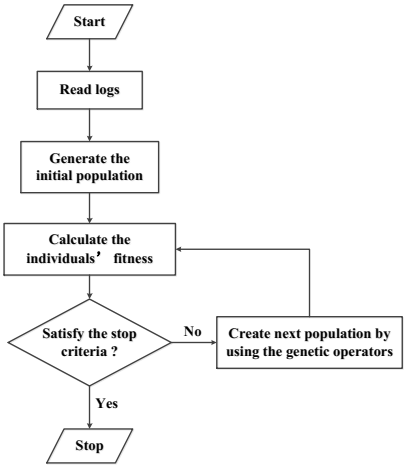


Fig. 2. Main Steps of Genetic Algorithm

Mining Normal Profile of Operator

Through statistical analysis of training logs and observation of control-flow model, we can easily determine who is involved in the business activity, what his tasks are, and the order of these tasks, their time interval and frequency. The role of an operator can be known from the training log, or it can be reasonably assumed if the corresponding information is not contained in the training log. More details can be get from the ExtraInfo field, such as the devices and the software, the operations on these objects and so on. After that, we use a multifurcating tree to represent the normal behavior profile of an operator. The outline of the tree is shown in Figure 3. The root of the tree is "Id-Operator", which indicates the business case and the operator's name. The other layers of the tree are as described in the following:

- Role level, which contains all the roles of this operator.
- Task level, which lists all the tasks that may be performed based on a specific role.
- Device level, which lists all the devices that are operated in the corresponding task.
- Data level, which lists all the software, files and data that are involved in the corresponding device.
- Operational level, which lists the operations and their frequency.

To facilitate the subsequent calculation, we use the behavior vector to represent these operations and their frequency range.

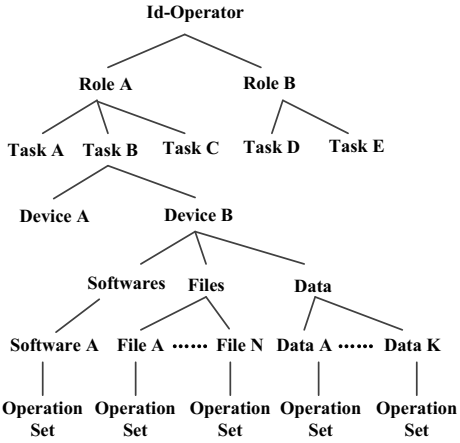


Fig. 3. An Example of Tree-Structured Profile of Operator Behaviors

Definition 4: Behavior Vector

A Behavior Vector BV is a tuple that indicates the operations on a device and their frequency of an operator in the execution of a certain task, i.e. BV(Operator, Task, Device) = ($f(op_1)$, $f(op_2)$ … $f(op_n)$), where $f(op_i)$ is the frequency of operation op_i, which satisfies the following equation:

$$f(op_i) = \#(op_i)/\sum\#(op_i) \tag{2}$$

where $\#(op_i)$ is the occurrence number of op_i.

In particular, BV_N represents the behavior vector in the normal profile, and BV_E represents the behavior vector in practical execution process.

3.3 Anomaly Detection

In the following, we will introduce anomaly detection method from two aspects. We list the types of anomalies concerned in this paper and then give the pseudo code of each anomaly detection algorithm.

Operator's Anomalous Behavior Detection

The formal definitions of operator's behavior anomalies and the pseudo codes of detection algorithms are given below.

Definition 5: Individual Anomalous Behavior

Let σ be the threshold, the Individual Anomalous Behavior is the situation that BV_E has some operations that do not belong to BV_N or the difference of frequency of a same operation between BV_E and BV_N is greater than σ.

Definition 6: Abnormal Outlier

An Abnormal Outlier is the operator whose current behavior vector significantly deviates from all the remaining peers' behavior vector.

The pseudo code of individual behavior anomaly detection algorithm is presented below according to the above definition:

```
Individual-Behavior_AD(Normal Profile, Execution Log)
for each task T and device D
   for each BV_E in Execution Log
      find the corresponding BV_N in Normal  Profile
      if {op|op ∈ BV_E}≠{op|op ∈ BV_N}
         trigger an Operation-Content alarm
      end if
      if ∃ op_i,∃|BV_E.f(op_i)-BV_N.f(op_i)|> σ
         trigger an Operation-Frequency alarm
      end if
   end for
end for
```

When detecting abnormal outliers, we also inspect the behavior vector both on the content and frequency of operation, and use distance-based clustering method to find out abnormal outlier. The pseudo code of detecting abnormal outlier is presented below:

```
Outlier_AD(Execution Log)
for each role R
   for each task T and device D
      set op-set to ∅
      for each operator O belongs to R
         op-set = op-set∪O.BV_E.op_i
      end for
      unify the op of each BV according to op-set
      for each operator O belongs to R
         for each op which is newly added into O.BV_E
            set f(op_i) = 0
         end for
         for each O.BV_E
            set the weight of each op
         end for
      end for
      cluster the BV_Es according to the Euclidean distance
      between each BV-pair
      for each cluster of BV_Es
         find out the abnormal cluster through expert
         knowledge and trigger an Outlier alarm
      end for
   end for
end for
```

Control-Flow Anomaly Detection

In the perspective of control-flow, we focused on the logical anomaly and the performance anomaly.

Definition 7: Logical Anomaly

Logical Anomaly refers to the situation that business process cannot follow the normal process structure so that causes abnormal termination or incorrect results.

Conformance checking is usually used to detect the logical anomaly in a business process. The pseudo code of detecting logical anomaly is presented below:

```
Logic_AD(Normal Model, Execution Log)
for each ES in Execution Log
   if ES can't be parsed by Normal Model
     trigger a Control-Flow-Logic alarm and log the fault
     event in ES
   end if
end for
```

Definition 8: Performance Anomaly

The Performance Anomaly refers to the situation that the indicators of some specific events in the current event sequence are beyond the normal scope in the performance model.

In detail, the Performance Anomalies include:

- Timestamp Anomaly: the timestamp of a specific event is beyond the scope of threshold τ_1;
- Time Interval Anomaly: the time interval of two specific events is beyond the scope of threshold τ_2;
- Frequency Anomaly: the occurrence number of a specific event is beyond the scope of threshold μ.

where τ_1, τ_2 and μ are the average value of the corresponding indicator in the performance model. The pseudo code of detecting performance anomaly is presented below:

```
Performance_AD(Time-Frequency Constraint Table, Execution
Log)
for each ES in Execution Log
   if event E has timestamp-constraint
     if E.Timestamp ∉ τ1
       trigger a Timestamp-Anomaly alarm and log E
     end if
   end if
    if event E1 and E2 have time-interval constraint
     if |E1.Timestamp-E2.Timestamp|> τ2
       trigger a Time-Interval-Anomaly alarm and log E1
       and E2
     end if
    end if
    let S be the set of events which have frequency
    constraint
    if #(S) ∉ μ
       trigger a Frequency-Anomaly alarm and log S
    end if
end for
```

4 Experimental Evaluation of Our Approach

4.1 Generating Dataset

To generate experimental data set, we used the Process Log Generator (PLG) to simulate the execution of five business cases. For a more accurate result, we simulate twenty times of execution for each business case, the event sequences of each business case are recorded into a log file with the format of MXML. This format can be supported by the $ProM_{import}$ plugin and be easily imported into the mining tool ProM.

After that, some anomalies are injected into the data set to simulate the insider threats the system suffered. Each injected anomaly corresponds to a pre-designed insider threat scenario, which describes the attacker's role, operating information and the affected activities. Once a scenario is created, then an operator in the correct role is selected, and the attack data are injected into the log files. In order to simulate the attack scenario as real as possible, the anomalies are directly related to some operators and the tasks they perform, and the operator's abnormal behavior is associated with the control-flow anomalies and the performance anomalies. The details of original data and the injected anomalies are shown in Table 1, where OCA represents Operation-Content Anomaly, OFA represents Operation-Frequency Anomaly, LA represents Logical Anomaly, TA represents Timestamp Anomaly, TIA represents Time-Interval Anomaly and FA represents Frequency Anomaly.

Table.1. Details of Five Business Process Cases and Injected Anomalies

No.	Average number of events	# Roles	# Operators	Injected anomalies
1	574	3	6	5 OCA, 2 LA
2	1227	4	9	9 OCA, 3 OFA, 3 LA, 3 TIA
3	2072	4	12	13 OCA, 5 OFA, 5 LA, 4 TA
4	3356	5	14	16 OCA, 7 OFA, 2 FA
5	5184	6	18	21 OCA, 12 OFA, 6 LA

4.2 Model Implementation

Model mining begins with log preprocessing. First, the log file of a business case is imported by the $ProM_{import}$ plugin of the ProM framework, which contains the event sequences generated during several executions. Then, the simple log filtering tools are used to filter the event sequences in the imported log file and retain those complete ones. These tools can help define expected starting and ending events and thus remove the incomplete ones. Next, the genetic algorithm plugin is used to deal with the preprocessed log file and discover the proper control-flow model of the business

case. The output is in the form of C-net. Finally, a Java program is written to discover the performance model and the operators' behavior profile. The relative parameters are collected and calculated during the process. To reduce experimental error and get a more stable and reliable model, we take the 10-fold cross validation method and the original dataset is partitioned to ten equally folds. The value of predefined metrics of a business case's model are average of the ten results gained from ten mining procedures.

4.3 Results

Table 2 shows the results from the detection system:

Table.2. Anomaly Detection Results

No.	# Alerts of Anomalies
1	5 OCA, 1 LA
2	9 OCA, 4 OFA, 3 LA, 2 TIA
3	14 OCA, 4 OFA, 4 LA, 4 MA
4	16 OCA, 8 OFA, 3 FA
5	22 OCA, 12 OFA, 7 LA

For evaluating the detection results, we use the F1 score as the evaluation indicator, which is commonly used in the fields of information retrieval and classification. The value of P, R and F1 score of each scenario are shown in Table 3. From these results, scenario 1 has the best precision of 100% and scenario 4 has the best recall of 100%. Due to the limited data size and injected anomalies, a slight deviation in FP and FN will result in a great impact on the precision and recall, so we judge the overall effect of the method by the F1 score, which is the harmonic mean of precision and recall. We can see from Table 1 and Table 3 that, regardless of the number of business events, the F1 score can maintain a high level of more than 90%, which indicates that the proposed methods have a good performance for detecting specific anomalies.

Table.3. Values of P, R and F_1 score

No.	P	R	F_1 score
1	100%	85.7%	92.3%
2	94.4%	94.4%	94.4%
3	96.2%	92.6%	94.4%
4	92.6%	100%	96.2%
5	92.7%	97.4%	95.0%

In the test of detecting abnormal outliers, we choose the fifth business case as test data, and use the agglomerative hierarchical clustering to cluster the behavior vectors of operators that belong to the same role, for discovering the abnormal outlier. The results are shown in Table 4:

Table.4. Results of Outlier Detection

Role	# Operator	# Cluster	# Outlier
1	3	2	1
2	4	1	0
3	5	2	1
4	3	1	0
5	3	2	1

Apparently, the first, third and fifth role each has an outlier and is consistent with our injected anomalies. Since the experimentations are conducted in a static environment and have no legal dynamic change of business activities, so the operator with individual anomaly may also be an abnormal outlier, which makes the number of abnormal outlier in Table 4 is equal to the number of injected anomalies in Table 2. In practical cases, if business activities legally adjust in the level of role, the operators' normal behavior vectors of a same role will have similar change, thus the abnormal outliers may be fewer than the operators with individual anomaly in a same role.

5 Conclusion

In this paper, we have proposed an effective insider threat detection system based on business process mining. The system takes the normal business logs as data source, discovers the control-flow model of the business process by genetic mining method. On the basis of this, the system further mines the tree-structured operator's behavior profile. In addition, it also gets the normal data about performance on specific events through statistical method and expert knowledge. Subsequently, we analyze the possible abnormal behaviors that a malicious operator may perform when conducting an insider attack, as well as the influences of these behaviors on the business activities. The anomaly detection methods are described in detail. The results of experimentation show that the proposed system performs well for detecting the injected anomalies and discovering the malicious operators.

However, the proposed methods also have some limitations. Once the business activities and the appointment are legally changed, the original model will be no longer applicable and new models must be mined again. Moreover, due to the lack of real business audit log, it is inevitable for some deviations in understanding the details of business activities. In future work, we will focus on how to build a dynamic model that with greater resilience for business process and operator's behavior. At the same time, we will research how to detect the collusive insider attack by discovering the social relationships between the employees within an organization.

References

1. Anderson R H, Bozek T, Longstaff T, et al. Research on mitigating the insider threat to information systems-# 2. RAND NATIONAL DEFENSE RESEARCH INST SANTA MONICA CA(2000).
2. Spitzner L. Honeypots: Catching the insider threat. Computer Security Applications Conference, 2003. Proceedings. 19th Annual. IEEE. (2003) 170-179.
3. Hu N, Bradford P G, Liu J. Applying role based access control and genetic algorithms to insider threat detection. Proceedings of the 44th annual Southeast regional conference. ACM (2006) 790-791.
4. Bishop M, Engle S, Peisert S, et al. We have met the enemy and he is us. /Proceedings of the 2008 workshop on New security paradigms. ACM (2009) 1-12.
5. Greitzer F L, Frincke D A. Combining traditional cyber security audit data with psychosocial data: towards predictive modeling for insider threat mitigation. Insider Threats in Cyber Security. Springer US (2010) 85-113.
6. Brdiczka O, Liu J, Price B, et al. Proactive insider threat detection through graph learning and psychological context. Security and Privacy Workshops (SPW), 2012 IEEE Symposium on. IEEE (2012) 142-149.
7. Parveen P, Evans J, Thuraisingham B, et al. Insider threat detection using stream mining and graph mining. Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on. IEEE (2011) 1102-1110.
8. Parveen P, Thuraisingham B. Unsupervised incremental sequence learning for insider threat detection. Intelligence and Security Informatics (ISI), 2012 IEEE International Conference on. IEEE (2012) 141-143.
9. Van der Aalst W M P, de Medeiros A K A. Process mining and security: Detecting anomalous process executions and checking process conformance. Electronic Notes in Theoretical Computer Science (2005) 121: 3-21.
10. Van der Aalst W, Weijters T, Maruster L. Workflow mining: Discovering process models from event logs. Knowledge and Data Engineering, IEEE Transactions on (2004) 16(9): 1128-1142.
11. Wen L, Wang J, Sun J. Detecting implicit dependencies between tasks from event log. Frontiers of WWW Research and Development-APWeb 2006 (2006) 591-603.
12. De Medeiros A K A, Weijters A. Genetic process mining. Applications and Theory of Petri Nets 2005, volume 3536 of Lecture Notes in Computer Science (2005)
13. Van der Aalst W M P, van Dongen B F, Günther C W, et al. ProM: The Process Mining Toolkit. BPM (Demos) (2009) 489: 31.
14. Burattin A, Sperduti A. PLG: A Framework for the Generation of Business Process Models and Their Execution Logs. Business Process Management Workshops. Springer Berlin Heidelberg (2011) 214-219.

Part II
Workshop SMECS-2016: 9th International Workshop on Simulation and Modelling of Engineering & Computational Systems

Design of Intermediate Frequency digital processing module based on AD9269

Li Mei
Engineering university of army police force，Xi ’an, China
E-mail: limei0117@126.com
TEL: 13519126070

Abstract. Digital Intermediate Frequency (DIF) is the key technology in software radio area. The DIF processing module of shortwave base on software radio technology has DIF processing to input signal, then complete Digital signal processing in digital area. In the DIF processing module, A/D and D/A are very important. It asks the A/D converter have enough working bandwidth (more than 2 GHz) and higher sampling velocity (60MHz), and higher conversion digit to improve dynamic range. The new A/D converter AD9269 can meet the needs. This paper introduces the design of DIF processing module built on AD9269, and explains the interface design of DIF processing module.

1 Introduction

With a rapid development of A/ D conversion technology and FPGA technology, the digital processing is increasingly closer to the antenna and applies the software to perform the radio function as much as possible. The DIF processing technology is one of important technologies to promote the performance of modern communication receiver, turning the analog signal into the DIF signal, dealing them with the digital down converter, the filter, A/D and so on, and finally converting them into the digital baseband signal. As the key component of the DIF processing, the early A/D converter is limited by its speed and precision and struggling to meet the requirement of the high- speed digitalization of the DIF receiver[1]. The article takes the DIF receiver of short- wave radio based on software radio technology as an example, and provides a design scheme of the DIF l processing module based on the new ADC component of AD9269.

2 Gain Analysis of DIF Processing

In the DIF processing system, the gain is mainly composed of three parts: the front-end low noise amplifier LNA gain, A/D gain, digital conversion gain, as shown in Figure1. We analyze the gain of each part in detail as follows.

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_45

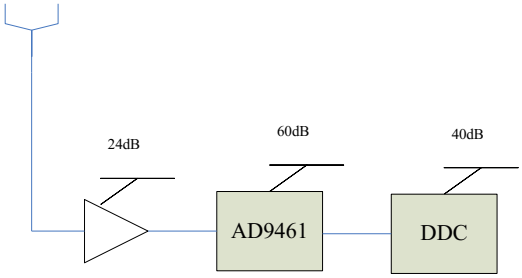


Fig. 1. Frequency conversion process

2.1 Low Noise Amplifier LNA Gain

The amplifier gain between unit board inputting to A/D sampling $G = 24dB$, the noise factor $NF = 2.4dB$. Based on the noise factor and gain, the noise level of the system front end adding to the A/D input end can be calculated [3]. The noise temperature of the system converting into the RF input end $T_e = T_0(10^{\frac{NF}{10}} - 1)$, among which $T_0 = 290^o K$, so the noise power $P_e = kT_eB$ $k = 1.38 * 10^{-23}$ is Boltzman constant, the system bandwidth $B = 3KHz$, so $P_e = -140dBm$. And because $G = 24dB$, the noise power presented at the A/D input end is $-140dBm + 24dB = -116dBm$.

2.2 A/D Gain

The noise power of the A/D input end is obtained according to the front endgain and the noise figure, then the system receiving sensitivity is determined by the performance of the A/Dconverter [3]. The A / D converter is 16 bits, 130MSPS, when the input signal is 30MHz, input amplitude is －1dBFS (compared with the full range), it can provide a signal to noise ratio (SNR)which is greater than 75dB and a spurious free dynamic range (SFDR) which is greater than85dB. The full range input of the A / D converter is ＋11dBm/50Ω.

Then we analyze the impact of the signal to noise ratio on the receiving sensitivity. It can be drawn that when the input signal amplitude is －1dBFS, SNR＝75dB, the noise level generated by the interior AD device is 16dBm－1dB－75dB＝－60dBm. Furthermore, based on the front end noise power-140dBm, through the A / D sampling, together with the noise generated by the interior AD device －60dBm, superposition of the two, the noise level of the entire bandwidth at the output end is －60dBm, when the signal is in full range input, the signal to noise ratio is 16dBm＋60dBm＝76dB。

3. The structure of the DIF processing system

The structure of the DIF processing system is shown in Figure 2. The system is mainly composed of the RF front- end, the DIF pre- processing and the DIF processing[2]. The radio- frequency signal received by the antenna is amplified by the prefilter amplifier circuit of 2- 30MHz. The DIF pre- processing scheme of 70 MHz is applied to effectively restrain the interference of combined frequency and sub- channel. The signal is sampled in A/ D after passing through the amplifier circuit of band- pass filter. The sampling clock is selected as 40MHz. After the under sampling of DIF signals of 70 MHz, the signal is going to be dealt with down- conversion mixer in the FPGA, and processed further in DSP when the signal has turned into the baseband signal at the rate of 80 KHz.

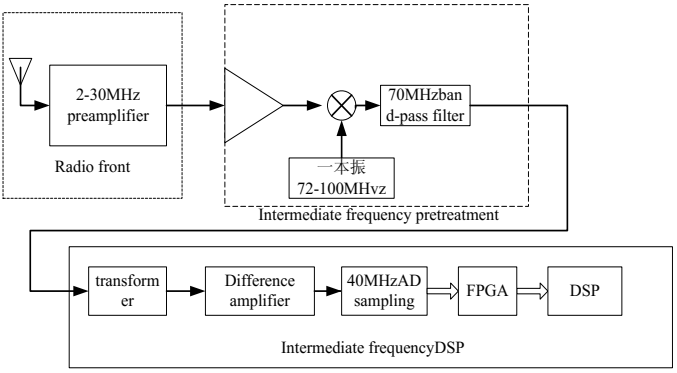


Fig. 2. Structural graph of the DIF processing system

4. The DIF processing unit

4.1 The digital down converter

On the principle comparison, the digital down converter is in agreement with the analog down converter[3]. While, two- way multiplier and the low- pass filter of the analog down converter have a bad consistency, leading an unevenly I and Q channel amplitude, with a great phase quadrature error, temperature drift and bad stability. In the digital down converter, the multiplier and the low- filter can be perfectly consistent with each other, which is less vulnerable to the temperature drift and noise, with a better precision and flexibility.

The digital down converter is shown in Figure 3. Aftr the under sampling in ADC, the DIF of 70 MHz accesses to the down conversion module at the rate of 40 Mbps, passing through the mixer, the CIC filter(extracted at 10 times), the half- band filter(extracted at 2 times), the Fir filter(extracted at 25 times) and finally accessing in the DSP at the rate of 80 Kbps.

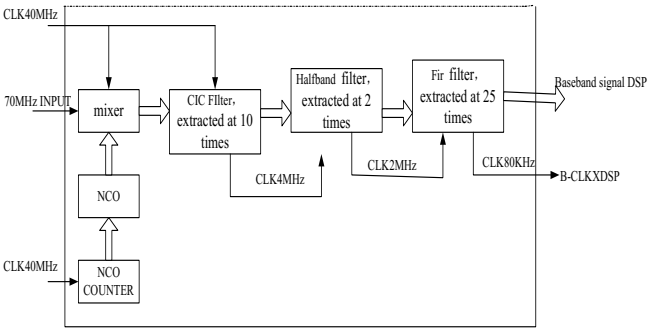


Fig. 3. The block diagram of the structure of down converter

To reduce the system noise and improve the operation stability, the whole system applies the same clock source of 80 MHz, with 40 MHz, 4 MHz, 2 MHz,80 KHz shared by the frequency of 80MHz.

4.2 The design of automatic gain control (AGC)

The DIF signal level received by the system relies on the radio- frequency input signal level and the front- end gain of receiver. Because the channel of radio signal is a complicated time- varying channel, the input signal of radio frequency changes greatly. The changing range is called as the dynamic range of receiver. The parameter of signal channel has a large difference sometimes, and it is bad for the rear processing [4]if there is not a good AGC controlling system, especially the frequency hopping system.

When the signal is very weak, it needs to increase the gain to make the DIF sampling signal perform in a full scale range, which benefits the precision of rear signal. When the signal is so strong, it needs to decline the gain or will result in the overflow and distortion. The text of AGC is to adjust automatically the gain of receivers, control the DIF signal amplitude within a certain range. In fact, the dynamic range compression amplifies the entire dynamic range of receivers.

The AGC system in this passage is a negative feedback system with a closed loop, as shown in Figure 4:

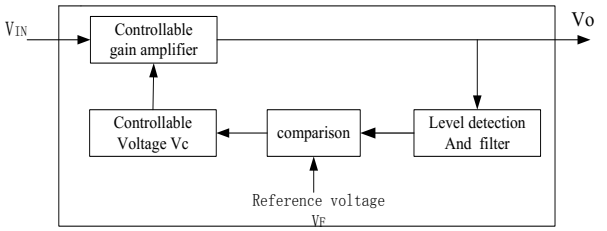


Fig. 4. The diagram of AGC circuit

The VC in the diagram is the control voltage, controlling the amplifying times of controlled gain amplifier. The gain range of the module is 0~ 50 dB with the foot path is 1dB/30mV, and when VC=0V, the gain is the smallest(0dB); when VC=1.5V, the gain is the biggest(50 dB).

4.3 The sampling of the band-pass signal

DIF input signal is 70 MHz. According to the Nyquist Law, the sampling rate must be above 140 MSPS, while the highest sampling rate of general ADC can not meet the demand. However, because the DIF is the narrow band signal only with several KHz~ hundreds KHz, we can apply the band pass sampling technology.

According to the band pass sampling law, with a frequency band pass signal of x(t) and a frequency bandwidth within (fL, fH), if the sampling frequency of fs can meet the demand of:

$$fs = \frac{2(f_L + f_H)}{2n+1} \quad (1)$$

In the formula, n is taken as a positive integer satisfying with $f_s \geq 2\ (f_H - f_L)$, using fs to perform the equal interval sampling. The signal x（nTs） can precisely restore the original signal x（t）. We can suppose the center frequency(i.e. the carrier frequency) as $f_C = \frac{f_L + f_H}{2}$, and the formula above can also be expressed as:

$$fs = \frac{4f_C}{2n+1} \quad (2)$$

n can be taken as a positive integer satisfying with fs ≥ 2（fH - fL）.

In this system, f_C=70MHz, n is taken as 3,and then:

$$fs = \frac{4\times 70}{2\times 3+1} = 40\text{MHz} \quad (3)$$

It is lower than the highest sampling rate of ADC. The signal of band pass sampling passes through CIC filter, half- band filter and FIR filter in FPGA and is sent to DSP, which not only decreases the signal rate, reduces the burden of DSP, but also prevent the aliasing of signals.

5. ADC components

According to the requirement of ADC sampling rate and the dynamic range, we use the AD9269 components[5] of AD cooperation in America.

The AD9269 has the accuracy of 16- bit with the highest sampling rate of 80 MSPS. The main characteristic includes that the sample- and- hold circuit still can

keep an excellent performance at the highest input frequency of 200 MHz, with a low cost, low power consumption and easy usage; the available function of SPI selective DC calibration and quadrature error calibration(QEC) can correct the DC offset, the gain and the phase mismatch between the two channels; the standard serial port interface(SPI) supports the feature and function of each product; the signal- to- noise ratio is 71dBFS(with the input value of 200MHz), and the spurious free dynamic range(SFDR) is 93dBc(with the input value of 9.7MHz) or 80dBc(with the input value of 200MHz), with the DNL of -0.5/+1.1 LSB.

References

1. Zhang Hongyu,: The PCB plate of digital receiver in the digital signal processor of DIF processing of software radio AD conversion, master degree thesis, University of Electronic Science and Technology of China (2006)
2. Zhuo Xingwang: The design and FPGA implementation of the demodulation system of digital correlator, master degree thesis, University of Electronic Science and Technology of China (2005)
3. Yang Xiaoniu, Lou Caiyi: The principle and application of software radio. Beijing, Tsinghua University Press (2003)
4. Attri, S.; Sohi, B.S; Chopra, Y.C: Efficient design of Application Specific DSP Cores Using FPGAS. ASSIC.2001. Proceeding 4th international conference on 23-25 Oct(2001)462-466.
5. AlalogDevicedInc.AD9269Datasheet.
6.Wepman A. Analog-to-digital Converters and Their Applications in Radio Receivers. IEEE Comu. Mag. 1995,33(5):46-54
7.Akos D M. Direct Bandpass Sampling of Multiple Distinct RF Signals.IEEE Transactions on Communications, 1999,47(7): 983-988
8. PN F901SC Data Sheet.1999-12
9. Donadio M P. CIC Filter Introduction. http://www.dspguru.com
10. F.Sheikh,S.Masud.Sample Rate Conversion Filter Design for Multi-standard software Radios. Digital Signal Processin, 2010,20(1):3-12
11. G.Dolecek,Simple Wideband CIC compensator. Electronics Letters,2009,45(24) 1270-1271
12. Lanying Zhang, Sijie Yuan. A Novel PCM/FM Demodulation Method Using DSTFT Analysis. Digital Object Identifier, ICOSP.2008:1775-1779
13. Research on the Method of Demodulation for PCM/FM Signal Based on Instantaneous Frequency Measurement. Digital Object Identifier,CESA.2006:93-96
14. P.Zicari, P.corsonello and S.Perri. An Efficient Bit-detection and Timing Recovery Circuit for FPGAs.Electronics, Circuits and Systems. FRANCE, 10-13 DEC 2006,1-3:168-171

Research on Adaptive Cooperative Spectrum Sensing

Ruilian TAN[1,2]
[1] College of Equipment Management and Safety Project, Air Force Engineering University, Xi'an 710051,China
[2] Armed Police Engineering University Electronic Technology Laboratory of Information Engineering Department, Xi'an 710086, China
madamtan@126.com

Abstract. Aiming at the fact that the traditional single threshold energy detection in cognitive radio is likely to misjudge situations and then lead to the shortage of missing the spectrum access opportunities due to the low signal to noise ratio caused by single threshold, this paper proposes adaptive cooperative spectrum sensing method based on double threshold energy detection. The model is made based on additive white Gaussian noise channel model, by improving the traditional double threshold energy detection method. For the size of the received energy statistics, different methods of decision are used, and the verdict is determined by the judgment methods directly. The simulation results show that under the condition of low signal-to-noise ratio, compared with the traditional single threshold energy detection model, the adaptive cooperative spectrum sensing model has better recognition performance, and with the increase of number of users, recognition performance is better.

Key Words. Cognitive Radio, Energy Detection, Double Threshold, Cooperative Spectrum Sensing

1 Introduction

Rapid growth of the wireless services makes spectrum resource gradually in shortage. The FCC spectrum strategy task group in the literature [1] points out that the allocated spectrum utilization rate is low and it is quite important to conduct full utilization of spectrum under the premise that it does not affect the main user communication; one of the most critical technologies is Cognitive Radio (CR) [2]. An important premise of CR is spectrum detection; only by the spectrum detection to find spectrum hole, can it realize the access of adaptive dynamic spectrum [3]. Currently, the idle spectrum detection method mainly includes matched filtering detection, energy detection and cyclostationary feature detection, and cooperative spectrum detection put forward in recent years. Literature [4] shows that the matched filter detection is the detection method of optimal signal, but requires knowing the priori

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_46

information of main users, such as modulation method, etc., and the prior information is hard to get in the actual process of communication. Literature [5] demonstrates that cyclostationary feature detection method can detect the category of the main user's signal, but the algorithm is complex, the testing time is too long. Literature [6] proves that the energy detection don't need any spectrum detection method of priori information, and the algorithm is simple and can be used for simple spectrum detection.

The quality of system modeling has greater influence on energy detection performance. Currently, there are two main types of model used for the detection. One is the detection method of certainty of main user's signal in white Gaussian noise channel put forward by the famous scholar Urkowitz [6]; the other is the model for detecting signal under the establishment of the fading channel put forward by the scholars such as Digham [7]. Spectrum detection is the key technology to improve spectrum efficiency under the premise that it does not interfere with the main user's normal communication, however in the practical communication, performance is often constrained by multipath fading and shadow effect, etc., making the performance of the single user detection greatly decline; aiming at this problem, in literature [8], based on cooperative communication and distributed detection, it proposes spectrum sensing algorithm of multi-user collaboration, and the simulation is used to prove the superiority of multi-user cooperation awareness.

In this paper, based on energy detection algorithm and the multi-user collaborative decision criterion, the first detection model is used to propose an adaptive cooperative spectrum sensing model based on double threshold energy test. The model is done with two steps based on double threshold energy detection method: firstly, through the energy detector, each user's energy statistics is received and then sent into double threshold decision center; cognitive user's energy statistics outside the threshold range directly takes hard decision to get the verdict; cognitive user's energy statistics outside the threshold range is directly accumulated and sent into the data fusion center to make collaborative integration judgment, mainly adopting "or" criterion, and then the integration center gives the final result through judgment. The simulation results show that with low signal-to-noise ratio and the same false alarm rate, the performance of the model is better than the traditional single threshold energy detection model; with the increase of the number of users, the detection probability is higher.

2 Single-user Energy Detection Model

Energy detection is the most basic method of spectrum detection, with simple implementation and low algorithm complexity. Judgment method is to preset a threshold, and through comparing the energy detector with the set threshold, if it is beyond the decision threshold, it is considered the main users exist in this frequency. We can define energy detection as a binary hypothesis problem:

$$\begin{cases} H_0: y(k)=\omega(k) & H_0 \\ H_1: y(k)=s(k)+\omega(k) & H_1 \end{cases} \quad (1)$$

Wherein, $y(k)$ is the signal received by cognitive users. $s(k)$ is detected signal, namely, the signal launched by the main user. $\omega(k)$ is additive white gaussian noise. k = 1, 2... , n; n is the used number of receiving signals. H_0 indicates idle state, namely main user does not exist. H_1 is the occupying state, namely the main user exists. When the main user does not exist in frequency, $s(k)$ is zero, to detect the situation of H_0 , and the energy detector at this time can be written as P[2]:

$$p = \sum_{k=0}^{n} \left| y(k) \right|^2 \tag{2}$$

Wherein, n is the sampling sequence vector dimension. Because signal energy of the channel in occupying state must be greater than the energy in the idle state; at this time, it can determine whether the main user exists; by comparing P and preset threshold λ_E , if it is more than λ_E , it belongs to main user in occupying state, otherwise in idle state, which can be used by accessing to the idle spectrum.

Spectrum detection performance can be measured by two probabilities [2] : detection probability P_d and false alarm probability P_f . P_f refers to the probability that the main user occupies the channel when the channel is in the idle state; the increase of P_f will make cognitive users lose the opportunity to access to the idle spectrum. We can express it as:

$$P_d = prob(P > \lambda_E \left| H_1 \right.) \tag{3}$$

$$P_f = prob(P > \lambda_E \left| H_0 \right.) \tag{4}$$

P follows chi square distribution with the degree of freedom of 2υ

$$P \sim \begin{cases} \chi^2_{2\upsilon} & H_0 \\ \chi^2_{2\upsilon}(2\eta) & H_1 \end{cases} \tag{5}$$

wherein, υ =TW, T is time length of observation, W is band width of observation, η is signal to noise ratio.

P probability distribution function is:

$$f_P(m) = \begin{cases} \frac{1}{2^{\upsilon}\Gamma(u)} m^{\upsilon-1} e^{-\frac{m}{2}} & H_0 \\ \frac{1}{2}\left(\frac{m}{2\eta}\right)^{\frac{\upsilon-1}{2}} e^{-\frac{2\eta+m}{2}} I_{\upsilon-1}\left(\sqrt{2\eta m}\right) & H_1 \end{cases} \tag{6}$$

wherein, $\Gamma(\bullet)$ is gamma function, $I_{\upsilon}(\bullet)$ is Bessel function of the first category of deformation of υ level.

Assuming cognitive radio network is not declined, detection probability P_d and false alarm probability P_f can be gained under additive white gaussian noise as follows:

$$P_d = prob(P > \lambda_E \left| H_1 \right.) = Q_{\upsilon}\left(\sqrt{2\eta}, \sqrt{\lambda_E}\right) \tag{7}$$

$$P_f = prob(P > \lambda_E \left| H_0 \right.) = \frac{\Gamma(\upsilon, \lambda_E / 2)}{\Gamma(\upsilon)} \tag{8}$$

Wherein, $Q_{\upsilon}(a,x)$ is universe Marcum Q function.

In accordance with the above assumptions, the flow chart of energy detector is shown in figure 1. The signal y (t) received by cognitive users first selects the corresponding bandwidth and center frequency by band-pass filter, and then calculate the energy of the signal through the square law device, and implement energy accumulated over a period of time through the integrator, finally get the energy detector statistics P which is sent to the judgment and preset threshold λ_E for comparison, if it is more than λ_E, it belongs to main user in occupying state, otherwise in idle state, which can be used by accessing to the idle spectrum.

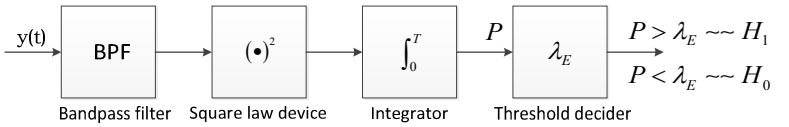


Figure 1. Flow chart of energy detector

3 Adaptive Cooperative Spectrum Sensing Model

Through the above analysis, we know that the traditional energy detection is based on the single threshold, namely, the threshold decision device has only one decision threshold, as shown in figure 2. When the main user's signal power is bigger and noise is smaller, single-threshold energy detection can well work; when the unknown noise signal power is bigger, namely under the condition of low SNR, the main user's signal might not be identified, and single-threshold energy detection may make a wrong decision, so as to make the false alarm rate increase, and miss the opportunity to involve in idle spectrum. In view of this, based on the shortage of the single-threshold energy detection, the literature [9] [10] put forward double-threshold energy detection of different algorithms, and prove the feasibility of this method through the computer simulation. The basic idea of double-threshold detection is as shown in figure 3. The test sets up two decision thresholds, λ_0 and λ_1; when the signal energy statistic value is greater than λ_1, the main user exists, and the local decision 1 is sent to the data fusion center. If the signal energy statistic value is less than λ_0, the main user does not exist and the local decision 0 is sent to the data fusion center. If the signal is between λ_0 and λ_1, the signal energy statistics needs to be sent to the data fusion center for another decision. As you can see, the decision is made for the choice of λ_0 and λ_1 and how to make decision in fusion center, which is the key for whether the right judgment can be made. Based on the double-threshold detection algorithm, this paper puts forward an adaptive cooperative spectrum sensing model, and its algorithm process is shown in figure 4.

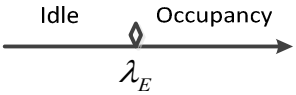


Figure 2. Single-threshold energy detection

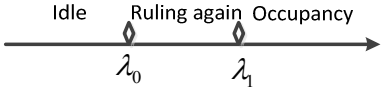


Figure3. Double-threshold energy detection

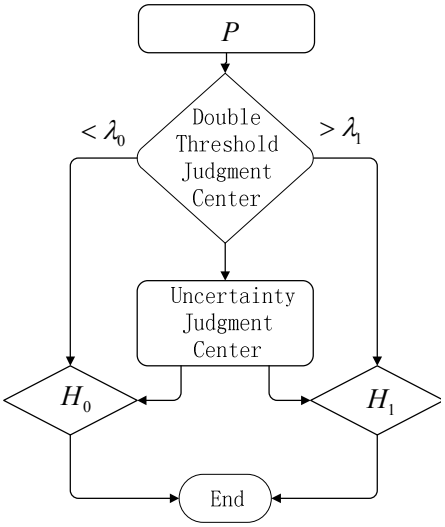


Figure 4. Adaptive cooperative spectrum sensing model

The energy statistics P of each user signal obtained by energy decision device directly gets into double-threshold decision device; if it is greater than λ_1, the main user exists and cannot be accessed for use, then the decision is ended, namely, a new spectrum access opportunity should be selected; if it is less than λ_0, it is directly judged as idle spectrum, and the decision is ended, namely it can begin to access and use the spectrum hole; if it is between λ_0 and λ_1, it needs to enter a state of uncertainty decision center for judgment, decision center at this time has only a kind of data, set as: R_i, $\lambda_0 < R_i < \lambda_1$, set that the data fusion center receives signal energy statistics in uncertain state sent by n cognitive users, the decision is made in accordance with the world's largest merger principle:

$$D=\begin{cases}0\ ,\ 0\le\sum_{i=1}^{N}w_i\bullet R_i\le\lambda\\ 1\ ,\ \sum_{i=1}^{N}w_i\bullet R_i>\lambda\end{cases} \tag{9}$$

Wherein, $w_i=\frac{\gamma_i}{\sqrt{\sum_{i=1}^{N}\gamma_j{}^2}}$ is the maximum coalescence coefficient. It can be seen that n statistics also almost follow chi square distribution. According to the formula (5), it can get

$$\sum_{i=1}^{N}w_i\bullet R_i\sim\begin{cases}\chi^2_{2N}\ , & \mathrm{H}_0\\ \chi^2_{2N}\left(2\bullet\sum_{i=1}^{N}(w_i\bullet\gamma_i)\right), & \mathrm{H}_1\end{cases} \tag{10}$$

The decision result after data center information is integrated as follows:

$$M=\begin{cases}H_0\ :\ \mathrm{D}=0\\ H_1\ :\ \mathrm{D}=1\end{cases} \tag{11}$$

In data fusion decision criteria, "or" strategy is the most conservative and optimal decision method; in this paper, the uncertain state decision center uses the "or" criterion for decision. Assuming N users that get into the uncertain state decision center are independent and identically distributed, so the detection probability $Q_{d.or}$ and false alarm probability $Q_{f.or}$ of multi-user collaborative detection in the uncertain state decision center are:

$$Q_{d.or}=1-\prod_{i=1}^{N}\left(1-P_{d.i}\right) \tag{12}$$

$$Q_{f.or}=1-\prod_{i=1}^{N}\left(1-P_{f.i}\right) \tag{13}$$

4 Simulation and Result Analysis

4.1 Single-threshold Energy Detection

From the formula (8), under the premise that false-alarm probability P_f is known, a single-user detection threshold can be can calculated, which is substituted into formula (7), then we can know the signal-to-noise ratio η is certain, namely, the detection probability can be obtained at this time. In AWGN channel, the simulation parameters are set as follows: T = 10, W = 5 * 104 hz, N = 1 = 0.01 = 0.05, 0.1 =, = -20, 0 dB, what is shown in figure 5 is the different SNR threshold energy detection performance. By the simulation results, it can be seen that by setting detection threshold λ_E and signal-to-noise ratio η, we can easily obtain the detection probability P_d. When λ_E is certain, the greater the signal-to-noise ratio is, the greater the detection probability is; when the signal-to-noise ratio is the same, the greater the detection false alarm rate P_f is, the greater the detection probability is, namely, P_d

increases as P_f increases , when SNR tends to -10dB, almost all can be checked out. P_f is too large, which means that cognitive users will lose the chance of more access to the spectrum hole, but if it is too low, which will make unable to test the main user; under the condition of low SNR, what is tested is noise, and it is error-prone, so the single-threshold detection is hard to balance detection probability and false alarm probability.

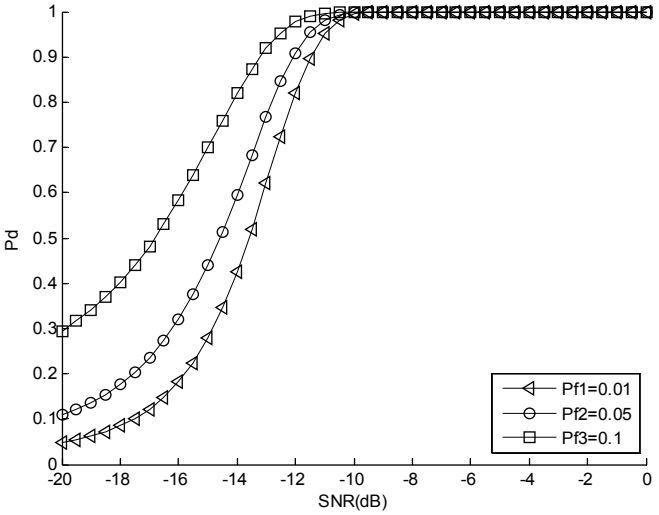


Figure 5. SNR and detection probability curve

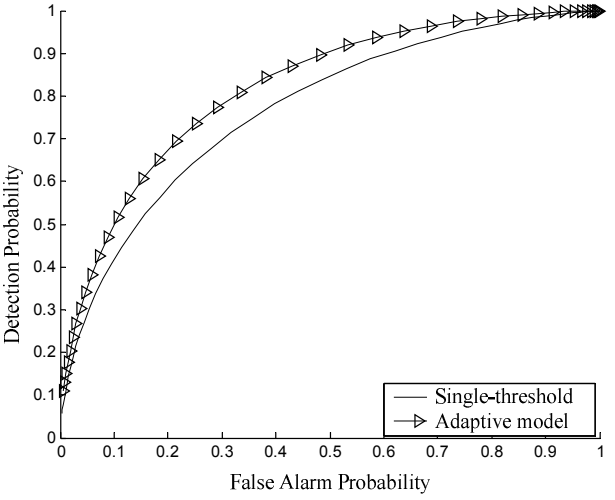


Figure6. False alarm probability and detection probability curve comparison

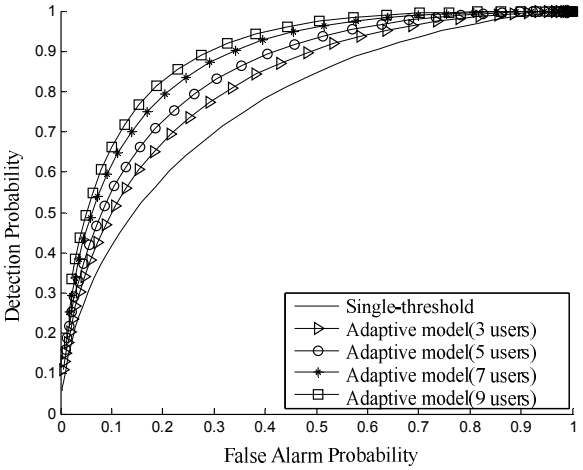


Figure 7. False alarm probability and detection probability curve comparison

4.2 Double-threshold Adaptive Energy Detection

In order to verify the performance of the proposed adaptive spectrum perception model, by using the above data to set the detection time T = 100, number of users N = 3, over AWGN channel to enter the energy detector, assuming that the channel is independent of each other, then each cognitive user takes this adaptive detection model for energy decision, the result is as follows:

Under SNR = -18dB when P_f = 0.1, from figure 6, it can be seen that the single-threshold energy detection probability is 0.4 or so, which is similar to figure 5 result, proving that the performance of the traditional single-threshold energy detection has great relation with the setting of detection threshold λ_E, but has little to do with the number of the detected users. Under the same P_f, when P_f is smaller, the performance of them is similar; with the increase of P_f, the detection probability by adaptive double threshold energy detection is obviously higher than that of single threshold energy detection.

Under the SNR = -18 dB, with adaptive double threshold energy detection model proposed in this paper, changing the number of users, the result in figure 7 is gained. As shown in figure 7, with the increase of the number of users, the performance of detection is better. The formula (13) shows that when $P_{f.i}$ is certain, the more the number N is, is smaller, the smaller $\prod_{i=1}^{N}\left(1-P_{f.i}\right)$ is, namely, the greater $Q_{f.or}$ is, the greater $Q_{d.or}$ is. This also verifies the accuracy of adaptive double threshold energy detection model proposed in this paper.

5 Conclusion

Based on the in-depth study of cognitive radio in traditional single threshold energy detection, aiming at the shortcomings of the single threshold energy detection on the reliability, this paper proposes a adaptive spectrum sensing model of double threshold energy detection, with hard decision method for certain energy statistics beyond the threshold range, and data fusion decision for energy statistics within threshold range, to better achieve the spectrum sensing. The simulation results show that the proposed adaptive spectrum perception model based on double threshold has better performance than single threshold energy detection model, and under the condition of low signal-to-noise ratio, it still has good recognition performance.

References

1. First Report and Order, Federal Communication Commission Std. FCC 02-48, Feb.2002.
2. P Qihang,Z Kun,W Jun,et al.A Distributed Spectrum Sensing Scheme Based on Credibility and Evidence Theory in Cognitive Radio Context. Proceedings of 17th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC) ,2006.
3. I F Akyildiz, W Y Lee, M C Vuran, S Mohanty. Next generation/dynamic spectrum access/cognitive radio wireless networks: a survey[J]. Computer Networks Journal(Elsevier),2006,50(13):2127-2159.
4. Cui Tao, Gao Feifei, Nallanathan A. Optimal ofcooperative spectrum sensing in cognitive radio. IEEETransactions on Vehicular Technology ,2011 .
5. Penna F, Garello R, Spirito M A. Cooperative spectrum sensing based on the limiting Eigen value ratio distribution in wishart matrices[J].IEEE Communications Letters. 2009,13(7):507-509.
6. Urkowitz H. Energy detection of unknown deterministic signals[J]. Proceedings of the IEEE.1967,55(4):523-531.
7. Kostylev V I. Energy detection of a signal with random amplitude[J]. ICC, New York, 2002:1606-1610.
8. PanLiang Deng,YuanAn Liu,Gang Xie,XiaoYu Wang,BiHua Tang,JianFeng Sun. Error exponents for two-hop Gaussian multiple source-destination relay channels[J]. Science China Information Sciences, 2012 (2) :388-391.
9. BOMFIN R C D V,GUIMARAES D A,SOUZA R A A D. On the probability of false alarm of the power spectral density split cancellation method. IEEE Wireless Communication Letters . 2016,8 (7):47-55.
10. Li Qunwei,Li Zan,Shen Jian,Gao Rui.A novel spectrum sensing method in cognitive radio based on suprathreshold stochastic resonance. 2012IEEE International Conference on Communications . 2012：2682-2686.

Research on Directional AF Cooperative Communication System Based on Outage Probability

Zhe LI[1,a], Ruilian TAN[1,2,b] , Kai SHI [3, c] , Baoxiang REN[1, d]
[1] College of Equipment Management and Safety Project, Air Force Engineering University, Xi'an 710051,China
[2] Armed Police Engineering University Electronic Technology Laboratory of Information Engineering Department, Xi'an 710086, China
[3]The Military Representative Office of PLA at Cheng du Aircraft Industrial Company, Cheng du 610092,China
[a] kongyanshi@126.com,[b] madamtan@126.com, [c] ganbing1981@126.com,
[d] renbaoxiang@126.com

Abstract. In order to improve outage property of cooperative communication system, directional antenna technology could be introduced. The paper proposes single relay AF cooperative communication protocol based on directional antenna and deduces the outage probability of system in fully directional mode theoretically so as to deduce the lower bound. With the minimum system outage probability to be the target, the paper analyzes OPA (Optimum Power Allocation) and makes detailed analysis and comparison of the system property of the established model by combining with specific simulation value. The simulation result shows that on the condition of high SNR (Signal to Noise Ratio) or greater difference of channel coefficient between two terminals and relay, the lower bound and accurate value are approximated and the outage property of fully directional cooperative communication is superior to that of traditional mode. Under the situation of limited power, OPA can enhance the outage property of system effectively.

Key Words. Cooperative Communication; Directional Antenna; Amplify and Forward; Outage Probability

1 Introduction

As the research topic in recent years, cooperative communication has gained many stage achievements[1-6], which mainly focus on: cooperative mode , capacity problem , network code , system synchronism , relay node selection , resource management , etc. Early in 1970s, Van has proposed, in the Literature [7], tri-terminal channel and deduced the upper and lower bound of the channel capacity, but cooperative communication, as a kind of diversity technique, was proposed clearly at

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_47

the beginning of this century [8,9]. Laneman et al firstly proposed a concept of two-stage synchronous transmission and gave the two currently common relay modes, namely amplify-and-forward and decode-and-forward[10]. But all these researches base on omnidirectional antenna. Directional antenna is able to focus on sending the energy to the direction needing communication so as to reduce the signal interference in non-communication direction, increase spatial multiplexing of channel and enhance the capacity of channel, so it has wide application prospects. In military field, directional antenna is more widely applied, especially in the occasion where RF (Radio Frequency) invisibility shall be considered, such as stealth fighter, the advantage of directional antenna is obvious and it has become the sole selection of the realization of communication and radar system. The researches on directional antenna and cooperative communication are very rich, but the researches on combination of them are fewer.

This chapter mainly considers AF cooperative communication system based on directional antenna on the single relay condition, supposes the communication mode to be TDMA mode and introduces directional antenna to cooperative communication system for exploratory research. The research focuses on the impact of outage probability in single node cooperative communication system on the whole performance of the system under the amplify-and-forward protocol. Firstly, through mathematical modeling and theoretical deduction, this chapter provides directional AF (DAF, Directional Amplify and Forward) cooperative communication system mode and its communication process based directional antenna; then deduces the outage probability in details, gets the approximate closed expression. On this basis, it gets OPA method; then, it, through simulation analysis, compares accurate outage probability, approximate outage probability and simplified outage probability, proves infinite approximation of them on the big SNR condition and analyzes the impact of each parameter on outage probability; at last, it analyzes the property change situation of system after introducing directional antenna, gets the relationship between SNR and outage probability on different conditions, analyzes the impact of power allocation on system property further and optimizes power allocation.

2 Cooperative Communication System Model based on Directional Antenna

2.1 Directional Antenna Model

Directional antenna receipt signal model is [10]:

$$P_f = \frac{P_t G_t G_r}{K r^{\nu}} \tag{1}$$

In the above formula, P_t represents transmitted power, G_t represents directive antenna gain at transmitting terminal, G_r represents directive antenna gain at receiving terminal, K represents the constant of atmospheric absorption and ohmic loss, ν is path decline coefficient, valued [2, 4] normally.

2.2 Cooperative Communication Model based on Directional Antenna

If transmitting and receiving terminals use directional antenna, then the cooperative system using directional antenna is directional transmitting-directional receiving mode, namely omnidirectional transmitting and omnidirectional receiving mode by comparing with omnidirectional antenna mode and the transmitting and receiving terminal of each joint increases two directive antenna gain, namely G_t and G_r. In order to study the system performance after leading in directional antenna conveniently, we mainly consider typical three-node cooperative communication mode in Rayleigh Flat Fading Channel. If $h_{i,j}$ expresses the channel coefficient of each node, then it obey the index distribution of parameter $1/\sigma_{i,j}^2$. It adopts directional antenna, so the beam forming signal energy can only cover a certain spatial angle, so compared with omnidirectional antenna model, it consumers one more communication time slot [10]. At this time, the whole communication process is completed by three time slots and two stages, as is shown in *Fig* 1:

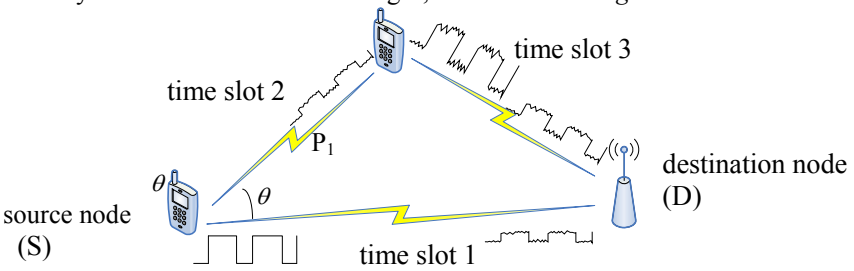


Fig.1 Directional and cooperative communication system model

Stage 1 (time slot 1 and time slot 2): source node transmits the radio information to destination node and relay node respectively.

Suppose the information received by destination node and relay node is $y_{s,d}$ and $y_{s,r}$ respectively, use G_s, G_r and G_d to express source node, relay node and destination node. Then:

$$y_{s,d} = \sqrt{P_1 G_s G_d} h_{s,d} x + \eta_{s,d} \quad (2)$$

$$y_{s,r} = \sqrt{P_1 G_s G_r} h_{s,r} x + \eta_{s,r} \quad (3)$$

In the above formula, x is the transmitted signal symbol, P_1 is the transmitted power, $\eta_{s,d}$ and $\eta_{s,r}$ are AWGN (Additive White Gaussian Noise) with the mean of zero and variance of N_O. $h_{s,d}$ and $h_{s,r}$ are channel coefficient from source to destination node and source to relay node and obey CGRV (complex Gaussian distribution of random variables) with mean of zero and variance of $\sigma_{s,d}{}^2$ and $\sigma_{s,r}{}^2$. The paper set the variance is not equal to zero.

Stage 2 (time slot 3): under AF protocol, relay amplifies and forwards the received signal to destination node.

Because we mainly study the performance of cooperative communication system under AF protocol after leading in directional antenna, in order to get theoretical

closed expression, we only consider ideal fixed AF protocol. Suppose the transmitted power of relay terminal is P_2, then in the case of amplification factor $\beta = \sqrt{P_2 G_r G_d} / \sqrt{P_1 G_s G_r \left|h_{s,r}\right|^2 + N_0}$, the signal received by destination terminal from relay terminal can be expressed as:

$$y_{r,d} = \beta h_{r,d} y_{s,r} + \eta_{r,d} \tag{4}$$

Substitute specific numerical value in Formula (3) and β into Formula (4), through organizing, we get:

$$y_{r,d} = \frac{\sqrt{P_1 P_2 G_s G_r{}^2 G_d}}{\sqrt{P_1 G_s G_r \left|h_{s,r}\right|^2 + N_0}} h_{r,d} h_{s,r} x + \tilde{\eta}_{r,d} \tag{5}$$

In the formula, $\tilde{\eta}_{r,d} = h_{r,d}\eta_{s,r}\sqrt{P_2 G_r G_d} / \sqrt{P_1 G_s G_r \left|h_{s,r}\right|^2 + N_0} + \eta_{r,d}$. Suppose $\eta_{s,d}$ and $\eta_{s,r}$ are independent, then equivalent noise is CGRV when mean is zero. The variance is:

$$\tilde{N}_0 = \left(\frac{P_2 G_r G_d \left|h_{r,d}\right|^2}{P_1 G_s G_r \left|h_{s,r}\right|^2 + N_0} + 1 \right) N_0 \tag{6}$$

In the case that channel coefficient is known, the destination terminal usually adopts MRC (Maximum Ratio Combining) method, the destination terminal will receive the signal of two links, the first is $y_{s,d}$, its SNR is γ_1, and the second is $y_{r,d}$, its SNR is γ_2. Then, the reception signal by adopting MRC can be expressed as:

$$y = A y_{s,d} + B y_{r,d} \tag{7}$$

In the formula, $A = \sqrt{P_1 G_s G_d} h_{s,d}^* / N_0$, $B = \sqrt{P_1 P_2 G_s G_d \beta} h_{s,r}^* h_{r,d}^* / \tilde{N}_0$. Suppose the average energy of transmitted symbol x is 1, then we can easily get:

$$\gamma_1 = \frac{\left|A\sqrt{P_1 G_s G_d} h_{s,d}\right|^2}{\left|A\right|^2 N_0} = \frac{P_1 G_s G_d}{N_0} \left|h_{s,d}\right|^2 \tag{8}$$

$$\gamma_2 = \frac{\left| B \dfrac{\sqrt{P_1 P_2 G_s G_r{}^2 G_d}}{\sqrt{P_1 G_s G_r \left|h_{s,r}\right|^2 + N_0}} h_{r,d} h_{s,r} \right|^2}{\left|B\right|^2 \tilde{N}_0} = \frac{\dfrac{P_1 G_s G_r}{N_0} \dfrac{P_2 G_r G_d}{N_0} \left|h_{s,r}\right|^2 \left|h_{r,d}\right|^2}{\dfrac{P_1 G_s G_r}{N_0} \left|h_{s,r}\right|^2 + \dfrac{P_2 G_r G_d}{N_0} \left|h_{r,d}\right|^2 + 1} \tag{9}$$

In the case that directive gain is given, $P_1 G_s G_d \left|h_{s,d}\right|^2 / N_0$ and $P_1 G_r G_d \left|h_{r,d}\right|^2 / N_0$ obey the index distribution. According to Formula (9), it can be seen that approximates the weighting harmonic mean of $\left|h_{s,r}\right|^2$ and $\left|h_{r,d}\right|^2$, approximate obedience parameter is:

$$\xi_2 = \frac{N_0}{P_1 G_s G_r \sigma_{s,r}{}^2} + \frac{N_0}{P_2 G_r G_d \sigma_{r,d}{}^2} \tag{10}$$

According to the above analysis, it can be seen instant mutual information in amplify-and-forward mode is the function of decline coefficient, that is:

$$I_{AF} = 1/2lb(1+\gamma_1+\gamma_2) \tag{11}$$

3 Fixed AF Outage Probability based on Directional Antenna

According to the definition of channel capacity, suppose $\lambda = P_2 / P_1$, $\gamma = P_1 / N_0$, substitute SNR of each link into it, get mutual information expression:

$$I_{AF} = 1/2lb(1+\gamma\left|h_{s,d}\right|^2 G_s G_d + f(\gamma\left|h_{s,r}\right|^2 G_s G_r, \lambda\gamma\left|h_{r,d}\right|^2 G_r G_d)) \tag{12}$$

In the equation, $f(x,y) = xy/(1+x+y)$, suppose all channels are Rayleigh channel, then $\gamma\left|h_{s,d}\right|^2 G_s G_d$, $\gamma\left|h_{s,r}\right|^2 G_s G_r$ and $\gamma\left|h_{r,d}\right|^2 G_r G_d$ comply with exponent distribution. The parameters are $\frac{N_0}{P_1 G_s G_d \sigma_{s,d}^{\ 2}}$, $\frac{N_0}{P_1 G_s G_r \sigma_{s,r}^{\ 2}}$ and $\frac{N_0}{P_1 G_r G_d \sigma_{r,d}^{\ 2}}$. Let $\gamma\left|h_{s,d}\right|^2 G_s G_d$, $\gamma\left|h_{s,r}\right|^2 G_s G_r$ and $\gamma\left|h_{r,d}\right|^2 G_r G_d$ be X_0, X_1 and X_2, and $\frac{N_0}{P_1 G_s G_d \sigma_{s,d}^{\ 2}}$, $\frac{N_0}{P_1 G_s G_r \sigma_{s,r}^{\ 2}}$ and $\frac{N_0}{P_1 G_r G_d \sigma_{r,d}^{\ 2}}$ are L_0, L_1 and L_2, the mutual information can be expressed as:

$$I_{AF} = 1/2lb(1+X_0+f(X_1,\lambda X_2)) \tag{13}$$

Then, the outage probability can be expressed as:

$$P_{out}(R) = P_r(I_{AF} < R) = E_{h_{s,d}h_{s,r}h_{r,d}}\left[1/2lb(1+X_0+f(X_1,\lambda X_2)) < R\right] \tag{14}$$

By calculating the integral of the above formula, we can get accurate outage probability on the condition of SNR. However, we can see, from the above formula, the accurate expression of outage probability under this situation is difficult to acquire, so the paper tries to utilize the following inequation to get the lower bound of outage probability:

$$\frac{xy}{x+y+1} < \frac{xy}{x+y} < \min(x,y) \tag{15}$$

Then:

$$P_{out}(R) = P_r((X_0+\lambda X_2) < 2^{3R}-1, X_1 > \lambda X_2) + P_r((X_0+X_1) < 2^{3R}-1, X_1 < \lambda X_2) \tag{16}$$

Expand the above equation, get:

$$\begin{aligned} & P_r((X_0+\lambda X_2) < 2^{3R}-1, X_1 > \lambda X_2) \\ &= \int_0^{2^{3R}-1} (X_0 < 2^{3R}-1-\lambda X_2, X_1 > \lambda X_2) f_{X_2}(x_2)dx_2 \\ &= \int_0^{2^{3R}-1}\int_0^{2^{3R}-1-\lambda X_2}\int_{\lambda X_2}^{\infty} f_{X_2}(x_2)f_{X_0}(x_0)f_{X_1}(x_1)dx_2dx_0dx_1 \end{aligned} \tag{17}$$

$$P_r((X_0+X_1) < 2^{3R}-1, X_1 < \lambda X_2) = \int_0^{2^{3R}-1} (X_0 < 2^{3R}-1-X_1, X_1 < \lambda X_2) f_{X_1}(x_1)dx_1 \tag{18}$$

Let $2^{3R}-1=n$, we can get closed expression as below:

$$P(R)=-(L_2+\lambda L_1)\left(\frac{e^{-L_0 n}-e^{-(L_1+L_2/\lambda)n}}{L_2+\lambda L_1-\lambda L_0}+\frac{e^{-(L_1+L_2/\lambda)n}-1}{L_2+\lambda L_1}\right) \quad (19)$$

According to Literature Attachment A, we can get the outage probability of SNR under big SNR situation:

$$P_r(I_{AF}<R)=\frac{n^2 L_0(L_1+L_2)}{2} \quad (20)$$

It is called simplified outage probability expression. In the above equation, R expresses spectrum efficiency, the unit is bps/Hz and Factor 3 in 3R is because cooperative process occupies three communication time slots. From Equation (20), we can see outage probability is proportional to γ^{-2}, so the gained diversity gain is 2.

4 Optimum Power Allocation(OPA)

Suppose total power to be P, and $P=(1+\lambda)P_1$, Formula (21) is rewritten to be expression about total power, as below:

$$P_r(I_{AF}<R)=\left(\frac{G_s G_r \sigma_{s,r}{}^2+\lambda G_r G_d \sigma_{r,d}{}^2}{2G_s G_d \sigma_{s,d}{}^2 G_s G_r \sigma_{s,r}{}^2 \lambda G_r G_d \sigma_{r,d}{}^2}\right)\left(\frac{2^{3R}-1}{P/N_O}\right)(1+\lambda)^2 \quad (21)$$

OPA is equivalent to

$$\min\left(\frac{G_s G_r \sigma_{s,r}{}^2+\lambda G_r G_d \sigma_{r,d}{}^2}{2G_s G_d \sigma_{s,d}{}^2 G_s G_r \sigma_{s,r}{}^2 \lambda G_r G_d \sigma_{r,d}{}^2}\right)(1+\lambda)^2 = \min F(\lambda) \quad (22)$$

Let $dF(\lambda)/d\lambda=0$, we get:

$$\lambda=\frac{\sqrt{G_s{}^2\sigma_{s,r}{}^4+8G_s G_d \sigma_{r,d}{}^2\sigma_{s,r}{}^2}-G_s\sigma_{s,r}{}^2}{4G_d\sigma_{r,d}{}^2} \quad (23)$$

So,

$$P_1=\frac{4G_d\sigma_{r,d}{}^2}{\sqrt{G_s{}^2\sigma_{s,r}{}^4+8G_s G_d\sigma_{r,d}{}^2\sigma_{s,r}{}^2}-G_s\sigma_{s,r}{}^2+4G_d\sigma_{r,d}{}^2}P \quad (24)$$

$$P_2=\frac{\sqrt{G_s{}^2\sigma_{s,r}{}^4+8G_s G_d\sigma_{r,d}{}^2\sigma_{s,r}{}^2}-G_s\sigma_{s,r}{}^2}{\sqrt{G_s{}^2\sigma_{s,r}{}^4+8G_s G_d\sigma_{r,d}{}^2\sigma_{s,r}{}^2}-G_s\sigma_{s,r}{}^2+4G_d\sigma_{r,d}{}^2}P \quad (25)$$

5 Simulation Analysis

The simulation in the paper supposes under the modulation of QPSK, the sending power of each joint is normalized as 1, then $N_{s,d}=N_{s,r}=N_{r,d}=1$, M=4, $b=1$. Suppose $\sigma_{s,r}{}^2=\sigma_{r,d}{}^2=1$, $\sigma_{s,d}{}^2=0.5$, R=1bps/Hz, the power allocated to relay node

and source node are consistent. When selecting SNR and antenna gain, considering the requirement of cooperative communication to SNR, so select directional antenna gain 15dB.

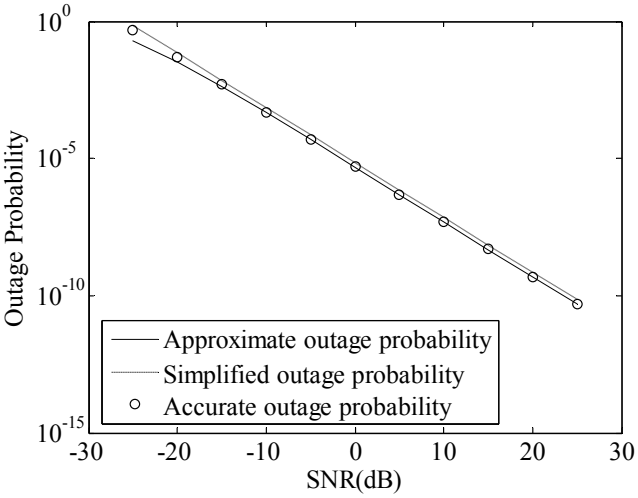


Fig.2 Relationship between outage probability and SNR

According to *Fig* 2, it can be seen that with the increasing of SNR, the accurate outage probability closes to approximate outage probability and simplified outage probability infinitely, so under general situation, the approximate outage probability or simplified outage probability can be used for related analysis. The analysis in the later part of the paper is made on the basis of simplified outage probability.

Keep parameters constant, through simulation contrast, the paper compares the outage probability of traditional fixed AF cooperative communication system and AF cooperative communication system based on directional antenna. The simulation result is shown as *Fig* 3:

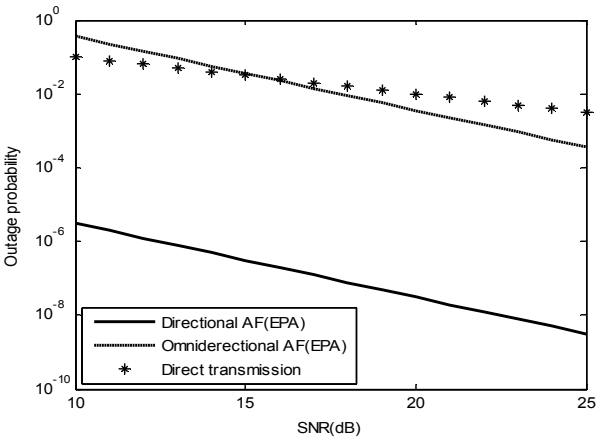


Fig.3 Relationship between outage probability and SNR

According to *Fig* 3, it can be seen, compared with omnidirectional decode-and-forward, the outage property of direct transmission becomes poor rapidly with the

increasing of SNR and each system gets 2 order diversity gains, but outage probability of directional fixed AF system is obviously superior to traditional fixed AF cooperative system, because directional fixed AF system has higher energy efficiency.

Keep other parameters constant, study the relationship between outage probability of directional fixed AF cooperative system and directional gain of each node. If changing direction gain of one node, the directional gain of other nodes is fixed as 14.5dB and fixed SNR is 15dB. Simulation result is as shown in *Fig* 4.

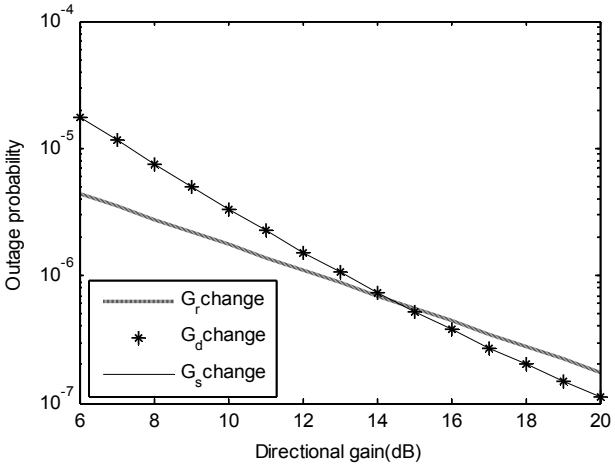


Fig.4 Relationship between outage probability and directional gain of each node

According to *Fig* 4, it can be know that outage probability of directional fixed AF system decreases with the increasing of gain of each node and the decreasing speed is consistent with the increasing of source node and destination node gain. The decreasing speed with the increasing of relay node gain is relatively slow, because relay node amplifies signal as well as noise, so its contribution to the enhancement of outage probability is inferior to the source node and destination node.

6. Conclusion

The paper takes outage probability as research focus to analyze the property of cooperative communication system under AF protocol after introducing directional antenna. Through the expression and approximate expression of outage probability after introducing directional antenna by means of theoretical deduction, the paper studies OPA problem further. It certifies that, after adopting fully directional antenna in cooperative communication system, it can provide communication gain in certain direction so as to enhance the system property remarkably; because applying relay node, signal transmission occupies one more communication time slot, which causes a certain spectrum waste. With the increasing of spectrum efficiency, the direction AF outage probability increasing speed is larger than traditional AF outage probability. It needs very big R to make AF effect inferior to that of traditional AF; in the situation of power limitation, OPA can enhance the outage property of system effectively. However, the relay node amplifies signal as well as its noise, because its contribution to the enhancement of system outage probability is inferior to source node and destination node.

References

1. Dhiya Al-Jumeily; Abir Hussain; Paul Fergus. Using adaptive neural networks to provide self-healing autonomic software[J]. International Journal of Space-Based and Situated Computing，2015 ， 5 (3): 122-129.
2. Yuki Arase, Takahiro Hara, Daijiro Komaki, Shojiro Nishio. Mobile search assistance from HCI aspect[J]. International Journal of Space-Based and Situated Computing，2011 ,1 (1): 18-29.
3. Felipe Barbosa Abreu; Anderson Morais; Ana Cavalli; Bachar Wehbi; Edgardo Montes de Oca; Wissam Mallouli. An effective attack detection approach in wireless mesh networks[J]. International Journal of Space-Based and Situated Computing，2015，5(3): 89-99.
4. Matija Puzar, Thomas Plagemann. You, Dwen-Ren Tsai. Data sharing in mobile ad-hoc networks – a study of replication and performance in the MIDAS data space[J]. International Journal of Space-Based and Situated Computing，2011 ,1 (1): 137-150.
5. Patricia Morreale; Allan Goncalves; Carlos Silva. Mobile ad hoc network communication for disaster recovery[J]. International Journal of Space-Based and Situated Computing，2015 , 5 (3): 178-186.
6. Slimane Hammoudi; Valérie Monfort; Olivier Camp.Model driven development of user-centred context aware services[J]. International Journal of Space-Based and Situated Computing，2015 , 5(2): 100-114.
7. Vander Meulen E C. Three terminal communication channels[J]. Adv. APPI. Probab. 1971: 120-154.
8. Laneman J N, Tse D N C, Wornell G W. Cooperative diversity in wireless networks: Efficient protocols and outage behavior[J]. Information Theory, IEEE Transactions on. 2004, 50(12): 3062-3080.
9. Hunter T E, Nosratinia A. Coded cooperation under slow fading, fast fading, and power control[C]//Signals, Systems and Computers, 2002. Conference Record of the Thirty-Sixth Asilomar Conference on. IEEE, 2002: 118-122.
10. Nasipuri A, Ye S, You J, et al. A MAC protocol for mobile ad hoc networks using directional antennas[C]//Wireless Communications and Networking Confernce, 2000. WCNC. IEEE, 2000: 1214-1219.

Research on Efficient Fibre-Channel-based Token-Routing Switch-Network Communication Protocol

Wen-lang LUO[1,2], Jing-xiang LV[1,2], Dong-sheng LIU[1,2], Xiao-hui ZENG[1,2,3]*
[1] School of Electronic and Information Engineering, Jinggangshan University, China
[2] Key laboratory of watershed ecology and geographical environment monitoring, NASG, China
[3] Department of Computer Science and Engineering, Northwestern Polytechnical University, China
*Corresponding author Email: zeng_xhui@163.com

Abstract. At present, the price of many network systems adopted for weather forecast in cluster system is too high, while the Ethernet is cheap, and Ethernet is the type of competition conflict network, but the Ethernet and TCP/IP can't meet the need of weather forecast computing which has the features of network traffics bursting and quasi real-time transmitting . To this end, we propose a Fibre-Channel-based Token-Routing Switch-Network (FC-Net, the technology has obtained a Chinese invention patent), and propose a new network communication protocol in which the efficient buffer management is the key to improve the data communication efficiency of. Experimental results show that our communication protocol, compared with other network systems, can effectively improve the data transmission rate .

1 Introduction

The price of many networks for cluster system to run weather forecast computing applications is too high nowadays. Ethernet is cheap, but Ethernet is competition conflict type. Ethernet and TCP/IP can't meet the requirements of large data exchange in the process of weather forecast computing. Thus, we designed and implemented the hardware of Fibre-Channel Token-Routing Switch-Network. Moreover, the suitable FCTRP (Fibre-Channel Token-Routing Switch-Network Protocol) is needed. In recent years, researchers have done a lot of meaningful work for the development network system [1, 2, 3, 4, 5, 6, 7].

2 Logic structure of FC-TR network Interface card

FC-TR Network Interface card is responsible for the network communication (including sending and receiving function), and the FCTRP communication protocol is based on it. The hardware logic structure of FC-TR network Interface card is as shown in figure 1.

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_48

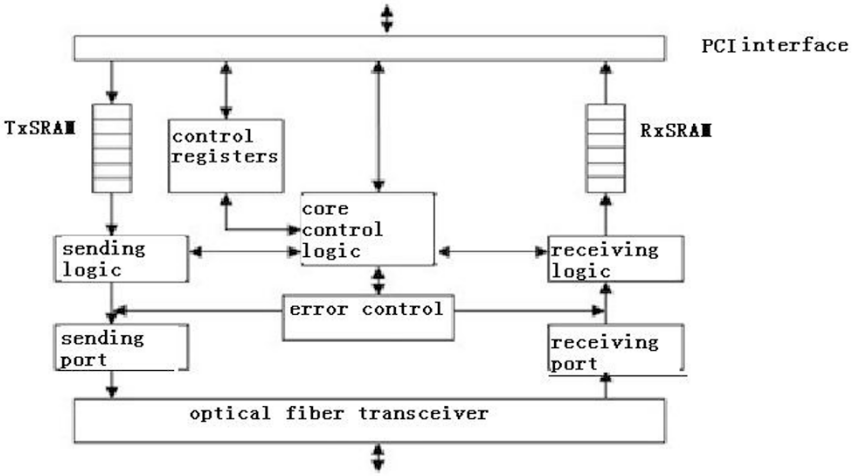


Fig. 1. Logic structure of FC-TR network Interface card.

FC-TR NIC logical structure mainly includes the following several parts:

- PCI interface

  Two data communication modes: master mode (DMA mode) can realize high speed data communication between the NIC and memory; slave mode can realize information interaction between the host and the NIC.

- core control logic

  Coordinating all of the NIC modules to finish processing data and related information, making them cooperate with each other smoothly.

- Control registers

  Including PCI configuration registers and status registers. The PCI configuration registers store initialization information, and the status registers store NIC working condition.

- error control

  By error control logic, the CRC check of data transmission can be directly implemented in the NIC hardware layer and achieve the purpose of error control; thereby the overheads of software layer for error control are reduced.

- Sending/receiving logic

  Being responsible for sending and receiving the NIC data. The sending logic will sent data frames or signaling frames from the NIC, and the receiving logic is utilized to receive data frames or signaling frames to the NIC.

- TxSRAM and RxSRAM

  On the FC-TR NIC, TxSRAM and RxSRAM (i.e. Sending and receiving SRAM) are 1 MB memory, being used to cache the sending and receiving data.

- optical fiber transceiver

  Being responsible for complete the photoelectric signal conversion and data transmission. It connects 2 fibre lines, one for outputting optical signals, another for inputting optical signals.

From the function of the NIC, optical signals are transmitted in the optical fibre network. According to physical properties of optical signals, optical fibre communication must comply with the standard of optical fibre communication. Since 1994, American National Standards Institute has issued the Fibre-Channel Standards, which are referred to as the FC standards[8]. The standards include FC-PH, FC-FS, and FC-SW, etc. . Each volume has multiple versions, and each volume content is hundreds of pages. FC standards communication protocol content is mainly about the hardware layer. This paper mainly studies the software platform of FC standard communication protocol.

The most basic feature of optical fiber communication is that the optical signals is being transmitted in a single optical fibre line with extremely high rate speed.To receive the code from the single fiber line, the clock signals must be extracted in the process of signal conversion.The 8b/10b encoding and 10b/8b decoding standards are firstly establish to solve the problem of extracting the clock signals.8-bit binary code is encoded for 10b code, and six continuous '0' or '1' are not permitted.As long as it is optical fiber communication, the FC standard must be complied, and our FCTRP is no exception.We find that the addressing mode of FC standards is only based on ID (identification), but the addressing mode of FCTRP protocol is based on the port number.So some modifications are made on the FC standard frame format, our FCTRP protocol can be realized.

3 FCTRP communication protocol design

3.1 FCTRP frame format

According to the Fibre-Channel Standards [8], an FC frame includes SOF (Start Of Frame), Frame Content, and EOF (End Of Frame).The Frame Content includes Frame_Header, Data_Field, and CRC. The standard FC frame format is shown in the figure 2.

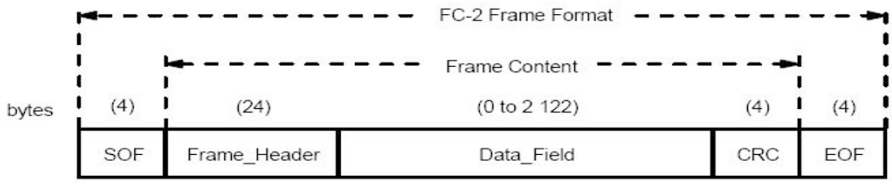


Fig. 2. Standard FC frame format

In the figure 2, SOF and EOF are four-byte "signal", regulated by the FC standards. The FC standard specifies the Class 1 service, Class 2 service, Class 3service, Class 6 service, and other services, and different services have different SOF or EOF.

Due to the FC-TR network is the end - to - end connection, belonging to the Class 1 service, "store - forward" and other services are not needed. Therefore, we simplifies the standard FC frame format for the FC-TR network. The simplified FC frame format is shown in the figure 3.

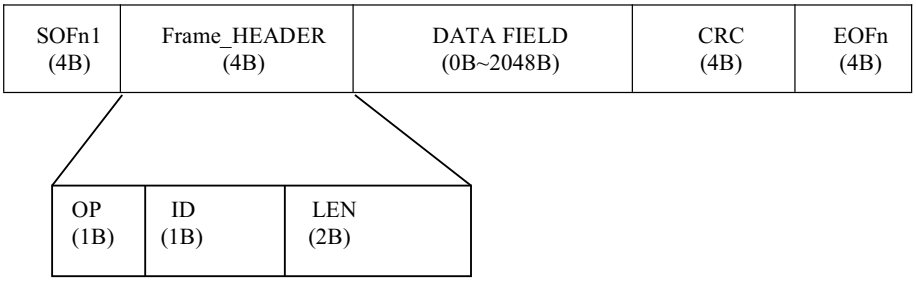


Fig. 3. FCTRP frame format

In the figure 3, "SOF" adopts the "SOFn1" of the FC standards (n is normal, and 1 represents the Class 1 service), and its hexadecimal code is 0xBCB53737. "EOF" uses the "EOFn" of the FC standards, and its hexadecimal code 0xBCB5D5D5. In addition, there is only Class 1 service in the FC-TR network, and the FCTRP Frame_HEADER is simplified to only 4 bytes length. OP is 1 byte (OP=0x70), and port (ID) is 1 byte; moreover, LEN (the frame data domain length) is 2 bytes, the FCTRP Frame_HEADER is decreased 20 bytes than that of the standard FC frame. For those dozens of bytes short frame, the transmission effect is improved obviously. In the FCTRP, in addition to the data frame, there is link control frame, which is divided into two subcategories: the ACK frame and the link command frame.

The ACK frame format is shown in the figure 4. In our ACK frame, SOFn1 code is 0XBCB53737, and OP is 11000000. The source port number (ID) is S_ID, and this kind of frame is mainly used to response sender, also being utilized to deal with flow control.

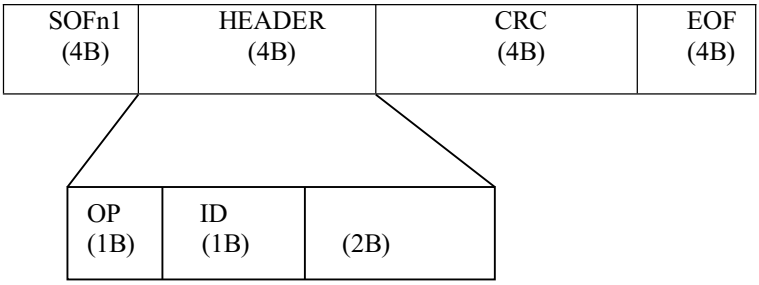


Fig. 4. FCTRP ACK frame format

The FC-TR network is based on our token routing technology, so the FC - TR NIC needs to the particular frame to deal with the work related to the switch. Thus, the link command frame is define, and the format is shown in figure 5.

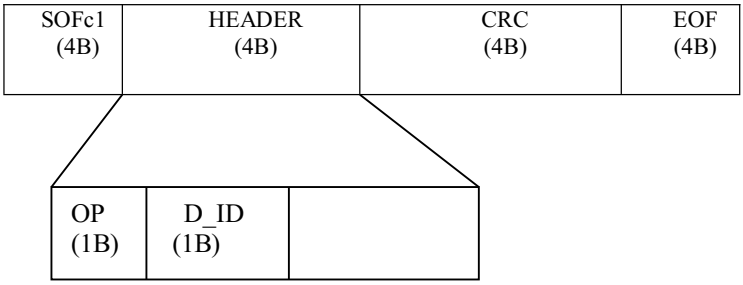


Fig. 5. FCTRP link command frame format

In the figure 5, the hexadecimal code of SOFc1 code (c representing connect, 1 meaning the Class 1 service) is 0XBCB51717. The link command frame shows that it is sent to the switch, and is related to the channel; and the OP code in the HEADER indicates a variety of different channel operation commands, which are as follows:

- command frame for channel, SHQ(D_ID)
  OP=0xA1 D_ID: destination port number
- command frame for broadcasting, BC ( )
  OP=0x22
- command frame for deleting channel, NDQ ( )
  OP=0x86

3.2 FCTRP frame-level communication protocol

The figure 6 shows a typical communication process of a set of “data frame”. In the figure 6, the source node and destination node finish communication with the

other side.by sending or receiving “frame”. Source node machine must apply to the switch to send “SHQ frame” through channels before sending the data frames; Switch sends back a “R_RDY” frame to the source node machine after connecting channel. And so on, until data frames are all sent out, source node sends an “NDQ frame” to cancel channel. Therefore, a set of sending “frame” or receiving “frame” commands is designed for software-level to control.

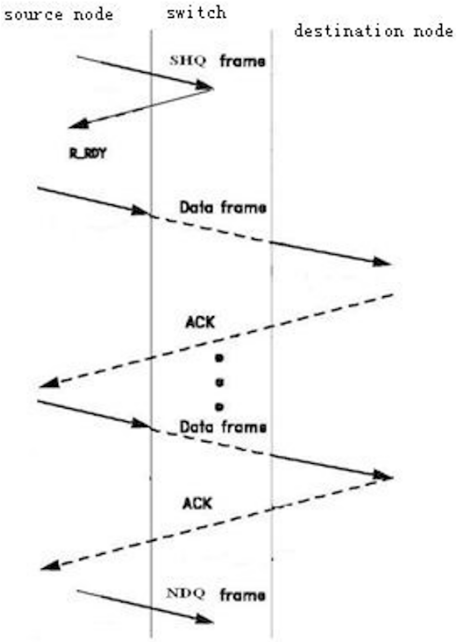


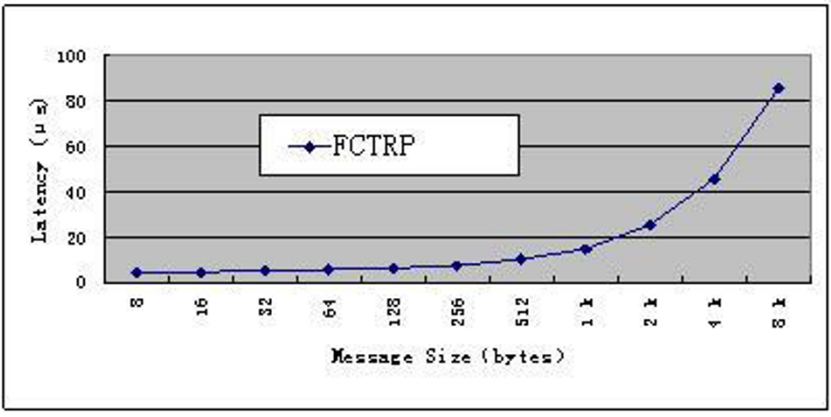
Fig. 6. FCTRP frame-level communication protocol

In general, the operating system does not allow applications to directly drive network card, and applications only can directly drive network card in kernel mode [9, 10, 11]. Thus, there is a communication bottleneck among user space, kernel space, and hardware layer. In the mesoscale numerical weather forecast parallel application, there are quite a few long data array to be transmitted, and the communication bottleneck has particularly effects on the communication efficiency. A special communication mechanism must be developed, this paper make full use of the VIA communication mechanism [12, 13, 14, 15].

4 Performance test and analysis

According to the above designed FCTRP communication protocol, combined with our existing research on VIA communication mechanism [16, 17, 18, 19], this paper implements a communication software based on the FC-TR network.

In order to make a further analysis on the performance of the FCTRP, this article tests FCTRP latency and bandwidth through the FCTR communication software. In this paper, the test environment is: two PC machines, each with an Intel Pentium 2.4G MHz CPU, 1GB memory, and FC-TR network Interface card, are installed Centos 5.5 operating system. A Ping-Pong benchmark is adopted between two processes, each executing blocking send and receive operations to wait for an incoming message and immediately responding once it has arrived. We have tested for 1000 times, and the bandwidth and latency are the mean values. The figure 7 and figure 8 show the communication latency and bandwidth.



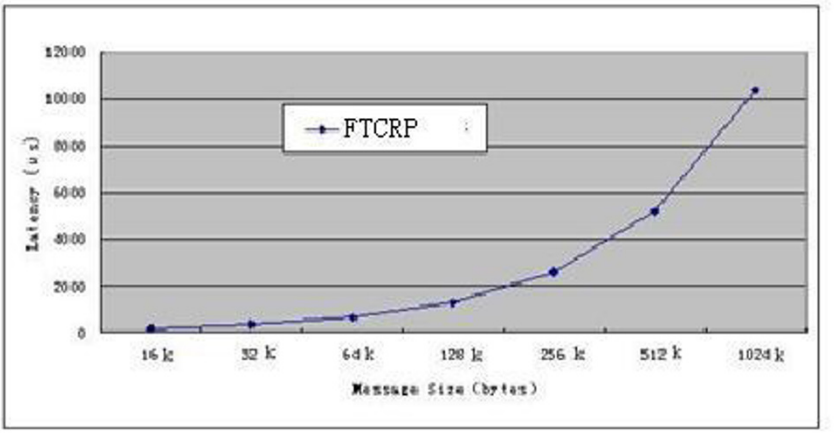


Fig. 7. FCTRP communication latency

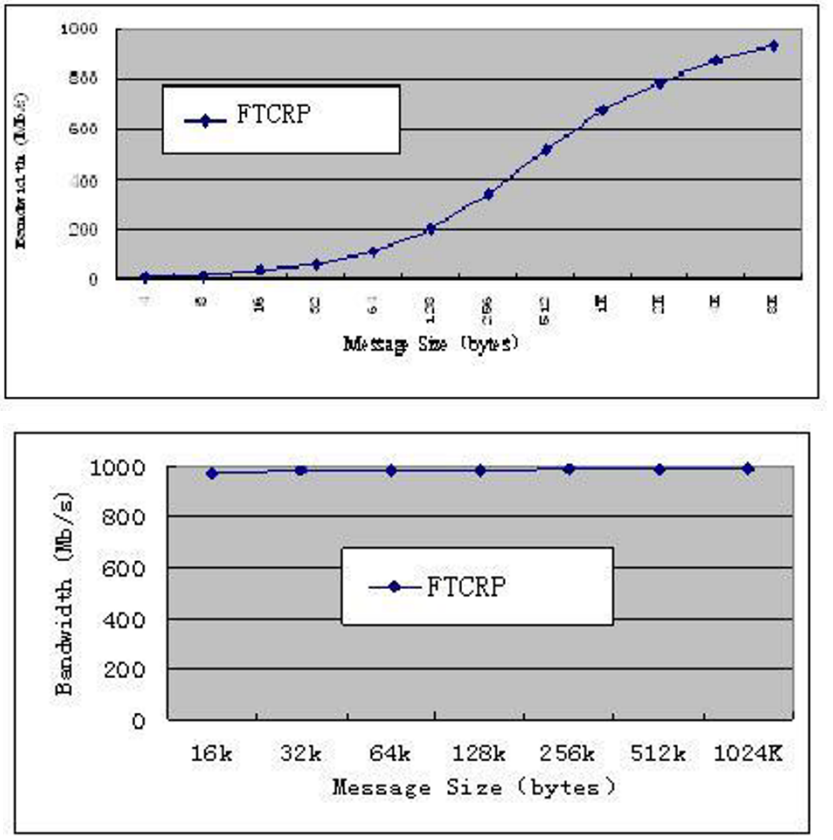


Fig. 8. FCTRP communication bandwidth

In the figure 7 and figure 8, the test results show the FCTRP minimum latency is about 4.2μs, and the FCTRP maximum bandwidth is about 987.9Mb/s. In general, when the sending frequency is around 1000 Mb/s, FCTRP communication performance is superior to Myrinet-GM [20, 21].

5 Conclusion

This paper introduces the logical structure of the FC-TR NIC and the hardware features; furthermore, the relevant issues about the FCTRP protocol design and implementation are emphatically discussed. In the process of designing FCTRP protocol and related communication software, a large number of long message communications in weather forecast computing applications is mainly considered. This paper makes full use of the VIA communication mechanism, and puts forward the parallel method of NIC DMA to reduce the overheads of memory copy, which significantly improves the communication efficiency.

Acknowledgment

This work is supported in part by the Key laboratory of watershed ecology and geographical environment monitoring (Grant No. WE2015013, WE2015002), the Opening Project of Key Laboratory of Embedded System and Service Computing of Ministry of Education (Grant No. 2010-09), the key technology R&D program of Jiangxi province (Grant No. 20151BBE50056, 2010BGA0090), the Science & Technology Transformation Program for Universities in Jiangxi Province (Grant No.KJLD13068, KJLD13066, KJLD14063).

References

[1]. S. Millich, A. George, and S. Oral, A Comparative Throughput Analysis of Scalable Coherent Interface and Myrinet, http://www.dolphinics.com/pdf/whitepapers/HSLNpaper2.pdf.
[2]. Mark A. Baker, Geoffrey C. Fox and Hon W. Yau, Cluster Computing Review, http://www.npac.syr.edu/techreports/hypertext/sccs-0748/cluster-review.html.
[3]. Shinko I, Fouquet Y, Nace D. Elastic routing for survivable networks[J]. International Journal of Grid and Utility Computing, 2015, 6(2): 121-129.
[4]. Niewiadomska-Szynkiewicz E, Sikora A. Simulation-based design of self-organising and cooperative networks[J]. International Journal of Space-Based and Situated Computing, 2011, 1(1): 68-75.
[5]. Houaidia C, Idoudi H, Van Den Bossche A, et al. Towards an optimised traffic-aware routing in wireless mesh networks[J]. International Journal of Space-Based and Situated Computing, 2014, 4(3-4): 217-232.
[6]. Ernst J B, Brown J A. Performance evaluation of mixed-bias scheduling schemes for wireless mesh networks[J]. International Journal of Space-Based and Situated Computing, 2013, 3(1): 22-34.

[7]. Damian A, Costan A, Cristea V, et al. A visualisation technique for network topology transformation within MonALISA monitoring framework[J]. International Journal of Grid and Utility Computing, 2011, 2(2): 119-129.
[8]. FIBRE CHANNEL FRAMING AND SIGNALING. ftp://ftp.t11.org/t11/pub/fc/fs/03-173v1.pdf.
[9]. Daniel P.Bovet, Marco Cesati. Understanding the Linux Kernel, second Edition (Chapter 7. Memory Management). December 2002, O'Reilly Press.
[10]. Robert Love. Linux Kernel Development, Second Edition (Chapter 11. Memory Management). January , 2005, Sams Publishing.
[11]. W.Richard Stevens. Advanced Programming in the UNIX Environment (Chapter 3, file I/O), Addison Wesley Publishing Company, 1992.
[12]. Compaq, Intel and Microsoft Corporations, "Virtual Interface Architecture Specification. Version 1.0", 1997.
[13]. NI, L. M., GUI, Y., AND MOORE, S. 1995. Performance evaluation of switch-based wormhole networks. In Proceedings of the International Conference on Parallel Processing
[14]. BOPPANA, R. V. AND CHALASANI, S. 1996. A framework for designing deadlock-free wormhole routing algorithms. IEEE Trans. Parallel Distrib. Syst. (Feb.), 169 – 183
[15]. Miao Kejian, Kang Jichang, Hong Yuanling. The research and design of network switcher in low degree parallel system. The Journal of northwest polytechnical unversity. 1999, 17(12):159-163
[16]. Lei Y, Junrui L, Chunyan M, et al. FC-TRSN: a new cluster-oriented high-speed communication network[C]//Proceedings of the 10th WSEAS international conference on Communications. World Scientific and Engineering Academy and Society (WSEAS), 2006: 557-561.
[17]. XU Xiaofei: Research on the Communication Protocols of the Cluster Inter-connection Based on the Software/hardware Co-design. [Ph.D. Thesis]. Northwestern Polytechnical University, Xi' an, China, 2004 (in Chinese with English abstract).
[18]. Zeng X, Yang S, Zhu B, et al. A Special Communication Model of Cluster for Multimedia Communications[C]//MultiMedia and Information Technology, 2008. MMIT'08. International Conference on. IEEE, 2008: 497-499.
[19]. Zeng X, Luo W, Yang C O, et al. A High Performance Computing Platform Based on a Fibre-Channel Token-Routing Switch-Network [M]//High Performance Networking, Computing, Communication Systems, and Mathematical Foundations. Springer Berlin Heidelberg, 2010: 83-87.
[20]. Yu W, Buntinas D, Panda D K. High performance and reliable NIC-based multicast over Myrinet/GM-2[C]//Parallel Processing, 2003. Proceedings. 2003 International Conference on. IEEE, 2003: 197-204.
[21]. Yu W, Panda D K, Buntinas D. Scalable, high-performance nic-based all-to-all broadcast over Myrinet/GM[C]//Cluster Computing, 2004 IEEE International Conference on. IEEE, 2004: 125-134.

Path Optimization of Coal Mining Machine Based on Virtual Coal-rock Interface

Dong Gang[1], Nie Zhen[1], Wang Peijun[2]

[1] School of Mechanical Engineering, Xi'an University of Science and Technology, Xi'an 710054, China.
dg5594@aliyun.com
[2] Border Defence Academy of PLA, Xi'an 710108, China.
pcbq@tom.com

Abstract. On the intelligent coal face, when coals fall down, the roller on the coal mining machine can recognize roof coal-rock interface and programme a motion path to avoid knifing the roof rock. It can not only effectively protect the rockerarm of coal mining machine but observably improve the quality of the raw coal produced on the working face. At present, the study on this subject focuses on two aspects: One is the study on technology for recognizing coal-rock interface based on all kinds of principles and the study on tech for projecting the slicing path of top roller; the other is the study on follow-up control technique which aims at the heightening system of coal mining machine. All the study on projecting the slicing path of top roller or the follow-up control of heightening system before did not take the size of the roller into account adequately, which caused the roof rock incised when the coals fell down. On the premise that the size of roller appearance was considered fully, this article puts forward the motion path optimizing of the coal mining machine's top roller based on virtual coal-rock interface, which solves the problem of roof rock's being sliced by top roller while working, and conducts a simulation experiment.

Key words: virtual coal-rock interface; coal mining machine roller; follow-up control; path optimization.

1 introduction:

The main equipment of the working face in the way of the longwall coal mining method's fully mechanized process includes coal mining machine, scraper conveyer and hydraulic support, which are abbreviated as 'three machines'. The coal cutter runs back and forth in the pathway of scraper conveyor. The upper roller makes coals fall along the roof coal-rock interface while the nether one picks up coals to the scraper conveyor. Meanwhile, the process of making coals fall from coal mining machine is very risky, so improving the automation level of the roller's falling coals in front of the coal mining machine becomes especially important[1, 2].

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_49

Now, there are two aspects of the study aiming at the coals falling automation on the coal mining machine: One is the study on technology for recognizing coal-rock interface based on all kinds of sensor devices and identification methods and the study on tech for projecting the slicing path of top roller; the other is the study on follow-up control technique which aims at the heightening system of roller on the coal mining machine.In the aspect of technology for recognizing coal-rock interface and projecting the slicing path of roller, some achievements have already been made.

For example, Zhi-peng believes that the roof coal-rock interface is a continuously process and comes up with the memorial cutting method based on artificial teaching, which helps to forecast the coal-rock interface[3,4].Aiming at the troditional memorial cutting way, Li Wei puts forward the further way on account of Hidden Markov Model (HMM) and validates that this method can provide a high-speed computation and a prediction of the errors between the reality and the result by emulating[5,6].

On the strength of the analyzing on the wavelet packet-energy spectrum Wang Bao Ping, Wang Bao Ping handles the rib image on the working face and gets the curve of the roof coal-rock interface. This method reduces the interference from equipment shaking[7,8].

In the way of inrfared imaging and image feat ure extraction, Ralston J.C achieves the identification of the coal-rock interface. This method reduces the interference of picture signal from coal dust[9]. Quan Guo Tong presents a method which can sleek the roller path by using certain cubic spline curve to optimize the roller cutting way and adopts the Particle Swarm Optimization(PSO) to get an accurate and fast solution[10,11].

All of these methods are able to identify the coal-rock interface successfully and determine the locate subject of the falling coals from the upper roller. However, none of them can solve the problem that the upper roller can not highten itself on the coal-rock interface.

The study on the heightening system follow-up control of coal mining machine is much less. FAN Qigao achieves the follow-up control of the coal mining machine heightening hydraulic system to the curve of the coal-rock interface's site[12]. Su Xiupin achieves the control of the follow-up system from the coal mining machine roller to the site of roof coal-rock interface[13,14].

Aiming at the element of the inadequate consideration to the roller's size in the past study of following coals from the coal mining machine's upper roller, this thesis puts forward a virtual coal-rock interface method in order to optimize the cutting path from the coal mining machine upper roller. This method firstly puts forward the virtual coal-rock interface obstacle curve based on the artificial potential field[15, 16]. According to it, we project the roller cutting path of equal step interval. Finally, we achieve the trace from roller to the coal-rock interface and avoid the cutting of the roof rock.

2 The strategy of optimizing coal mining machine roller

2.1 The traits of the coal mining machine cutting path

To try to sweep the coals under the coal mining machine and avoid cutting the roof rock, coal mining machine's upper roller tails after the roof coal-rock interface while working. In the normal condition, the coal-rock working face is smooth. When it appears the fault on the working face, the roof coal-rock interface will crack or fold. When this happens, the upper roller has to adjust the angle according to the changings of coal-rock interface, or it must cut overmuch roof rock.

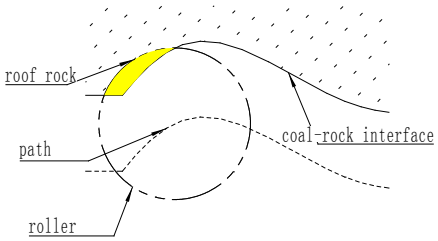


Fig.1. When the roller move in a path which is translational from coal-rock interface, it will cut the area of roof rock which is yellow in the Fig.

As shown in the figure 2.1, when the roof coal-rock interface bends, if the mining machine roller is still projecting path and cutting in the radius R of the next roller, it must cut the roof rock. The yellow area that roof and roller coincide is the cut roof rock.

Therefore, when the coal-rock working face bands, it is inappropriate to project the upper roller's cutting path by the translation of the coal-rock interface. On the contrary, we should consider the shape of coal-rock interface and the size of the roller adequately and adopt the method which can reduce or even avoid cutting rock.

2.2 The basic theory of the virtual coal-rock interface

The method of artificial potential field on projecting path of the coal mining machine's upper roller on the virtual coal-rock interface put forward by this thesis is an arithmetic based on virtual gravitation and repulsion.

The basic thought simplify the robot, barrier and moving target as dots. Barrier produces the repulsion; target produces gravitation to the robot in the whole situation; the robot avoids the barriers and move to the target with the combined action of gravitation and repulsion.

The definition of the gravitation and repulsion field is as follow:

$$U_{att}(x) = \frac{1}{2}\alpha\rho_g^2(x) \tag{1}$$

$$U_{rep}(x) = \begin{cases} \frac{1}{2}\eta\left(\frac{1}{\rho(x)} - \frac{1}{\rho_0}\right)^2 \rho_g(x)^m, \ \rho(x) \le \rho_0 \\ 0, \ \rho(x) > \rho_0 \end{cases} \quad (2)$$

α and η are the gain factors of the gravitation and repulsion. m is arithmetic number, ρ_0 is the sphere of influence, $\rho_g(x)$ is the Euclidean distance between the robot and target. The definition of the gravitation $F_{att}(x)$ and repulsion $F_{rep}(x)$ function is as follow:

$$F_{att}(x) = -grad[U_{att}(x)] = \alpha\rho_g(x)\vec{n}_g \quad (3)$$

$$F_{rep}(x) = \begin{cases} F_{rep1}(x) + F_{rep2}(x), \ \rho(x) \le \rho_0 \\ 0, \ \rho(x) > \rho_0 \end{cases} \quad (4)$$

$$F_{rep1}(x) = \eta\left(\frac{1}{\rho(x)} - \frac{1}{\rho_0}\right)\left(\frac{1}{\rho(x)}\right)^2 \rho_g(x)^m \vec{n}_0,$$

$$F_{rep2}(x) = \frac{m}{2}\eta\left(\frac{1}{\rho(x)} - \frac{1}{\rho_0}\right)\rho_g(x)^{m-1}\vec{n}_g$$

In the formula (3) and (4), $\vec{n}_g$ is the unit vector for the robot aiming at the target. We can infer that resultant STRESS to robot is $F(x) = F_{att}(x) + F_{rep}(x)$, which lead the pathway of the robot.

Traditional artificial potential is usually adapted in static projecting the path but is not quite much in applying in the dynamic aspect to solve the space overlay problem. This thesis improves the traditional artificial potential. And according to the traits of this problem, it is divided into some segments and is used in combination to add the whole situation's factors of the influence in the period of iteration so that the roller can get to the practical target as well as covering and avoiding the rock.

2.3 The realization of the virtual coal-rock interface.

To solve the problem of projecting the path of coal mining machine roller, this thesis puts forward the improved artificial potential, whose main thought is:

Arrange that the simplified curve of the roof coal-rock interface is $L_d : y = f(x)$. When the upper roller in the radius R moves from the left of the working face S to the right G, find dots on the curve L_d every interval divided by ∇d as the barriers in the process.When the edge of the roller enters the influence area, the barriers will reject the roller to avoid cutting overmuch rock; in order to ensure that the roller can finish cutting the roof coal, defined gravitational field on the working face can produce gravitation F_T; find curve L_d's every extreme point as a periodical target point. Let the roller arrive the target points one by one. Every point will be eliminated as soon as the roller arrives, and the forward one will be activated at the same time; Arrange a virtual target point outside the right of the working face ,

which produces gravitation to the roller and ensures that the roller can move to the right to get the mission finished.

Arrange n extreme points as the target points $G_1, G_2, \ldots, G_n$, and make the roller reach these points from far to near one by one, finally reach the other side of the working face. Arrange the target point G' over the roof outside the coal wall of working face, and define gravitation up the vertical roof, and the roller is pulled vertically, and the gravitation is constant F_T. Along the curve of the coal-rock interface, arrange the barrier points are $B_1, B_2, \ldots, B_m$, which will stop the roller entering into the roof. The initial position of the roller is on one of the endpoints, while the termination point is the other endpoint G. After the roller reaches the target point G_{i-1} and moves to target point G_i, STRESS function is as follow:

$$F(x) = F_{G_i}(x) + F_{G'}(x) + F_T + F_{B_j}(x) + \cdots + F_{B_{j+k}}(x) \quad (5)$$

$F_{G_i}(x)$ is the gravitation from the target point G_i to the roller; $F_{G'}(x)$ is the gravitation from virtual target point to the roller; F_T is the upright gravitational constant; $F_{B_j}(x), \ldots, F_{B_{j+k}}(x)$ are the repulsions from the barriers which enter the influence zone of the roller along the edge of the roller.

$$F_{G_i}(x) = \alpha\left(\rho_{G_i}(x) - R\right)\vec{n}_{G_i}$$

$$F_{G'}(x) = \beta\rho_{G'}(x)\vec{n}_{G'}$$

$$F_{B_{j+d}}(x) = \begin{cases} F_{rep1B_{j+d}}(x) + F_{rep2B_{j+d}}(x), & \rho_{B_{j+d}}(x) - r_2 \le \rho^0_{B_{j+d}} \\ 0, & \rho_{B_{j+d}}(x) - r_2 > \rho^0_{B_{j+d}} \end{cases}$$

$$F_{rep1B_{j+d}}(x) = \eta\left(\frac{1}{\rho_{B_{j+d}}(x) - R} - \frac{1}{\rho^0_{B_{j+d}}}\right)\left(\frac{1}{\rho_{B_{j+d}}(x) - R}\right)^2 \left(\rho_{G_i}(x) - R\right)^m \vec{n}^0_{B_{j+d}}$$

The improved multi-objection artificial potential combines the specific requirements from the production technology of fully-mechanized working face to the motion rule of coal mining machine, which means that in the period of the roller moves from one side to the other side, it can cling to the coal-rock interface facing to the coal wall, cover it and avoid overmuch cutting to roof rock at the same time. It expands the function of traditional artificial potential from point-to-point path projection to one can cover the working zone on the basis of certain rule.

3 Simulation experiment

According to the coal seam practical situation of the mine five in Pingdingshan, arrange the parameters to the simulation experiment. The working length face is 200m. The thickness of coal seam is 3m. There are three bending transformations in the middle of working face on account of the fault influence to the roof coal-rock inter-

face. The radius of the mining machine roller is 0.8m. Roller starts cutting from the left and ends up to the right. Project the coal mining machine upper roller pathway by MATLAB. In the simulation experiment, the curve of the coal-rock interface is set in the way of cubic polynomial interpolation. The main constants of the multi-objection artificial potential are set as:

Gravitation coefficient: 70 ; repulsion increment: 2; normal gravitation: 5
barrier influence distance: 0.05 ; virtual goal point: (300,50)

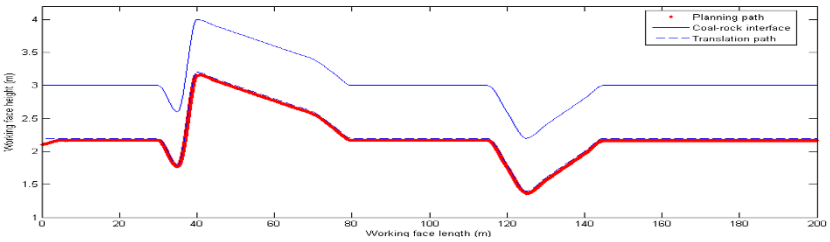


Fig.2. The planning pathway of coal mining machine's roller which is obtained with the artificial potential field method is showed in this Fig.

The red line in picture 2 is the improved pathway from the multi-objection artificial potential to the upper roller of coal mining machine. It is able to be close to the coal-rock interface and avoid cutting the rock. But if we set the roller moving path in the way of moving the curve, the rock will be cut the place where the curve is larger. In the profile of the roller, the center of the roller circle cuts the rock on the moving curve in the interface. Finding a center of roller circle in the profile two of the roller on the multi-objection pathway curve avoids the rock clearly. Proved by simulation, the roller path projected by the multi-objection can truly avoid the rock, which provides a solution for increasing the production efficiency on the intelligent working face.

4 Conclusion

Intelligent working face is the future of modern mine. As the central equipment of the working face, the intelligent level of coal mining machine is particularly important. The quality of the coal mining machine's working path directly influents the safety of roof, damage of equipment and quality of coal and so on. Aiming at that problem, this thesis puts forward multi-objective artificial potential to project the roller's working pathway. The result of simulation indicates that its ability in covering the rib and avoiding the rocks on fault. It can improve the safety working efficiency and provide new ideas and methods for further development of the intelligent working face.

References

1. Wang Jinhua, Huang Leting, Li Shoubing, et al. the development of fully mechanized working face technology and equipment [J]. Journal of Coal.
2. Ralston, Jonathon, Reid, et al. Sensing for advancing mining automation capability:A review of underground automation technology development[J]. International Journal of Mining Science & Technology. 2014, 24(3): 305-310.
3. Zhi-Peng X U, Wang Z B, Jin-Peng M I. Shearer self-adaptive memory cutting[J]. Journal of Chongqing University. 2011, 34(4): 134-140.
4. C. Houaidia, H. Idoudi, A. Van Den Bossche, T. Val, and L. A. Saidane, "Towards an Optimized Traffic-Aware Routing in Wireless Mesh Networks," International Journal of Space-Based and Situated Computing, vol. 4, p. 217-232, 2014.
5. Li W, Luo C, Yang H, et al. Memory cutting of adjacent coal seams based on a hidden Markov model[J]. Arabian Journal of Geosciences. 2014, 7(12): 5051-5060.
6.Lin, Z. (2014). Clock-controlled generators with large period output sequences. International Journal of Grid & Utility Computing, 5(4), 278-285.
7. Wang B P, Wang Z C, Li Y X. Application of Wavelet Packet Energy Spectrum in Coal-Rock Interface Recognition[J]. Key Engineering Materials. 2011, 474-476: 1103-1106.
8 Houaidia, C., Idoudi, H., Bossche, A. V. D., Val, T., & Saidane, L. A. (2014). Towards an optimized traffic-aware routing in wireless mesh networks. International Journal of Space-Based and Situated Computing,4(3/4), 217-232.
9. Ralston J C. Automated longwall shearer horizon control using thermal infrared-based seam tracking[C]. 2012.
10. M. Puzar and T. Plagemann, "Data sharing in mobile ad-hoc networks-a study of replication and performance in the MIDAS data space," International Journal of Space-Based and Situated Computing, vol. 1, p. 137-150, 2011.
11. Quan G T, Tan C, Hou H C, et al. Cutting Path Planning of Coal Shearer Base on Particle Swarm Triple Spline Optimization[J]. Coal Science & Technology. 2011.
12. Fan, Qigao, Li, et al. Control strategy for an intelligent shearer height adjusting system[J]. Mining Science & Technology. 2010, 20(6): 908-912.
13. Nesmachnow, S., & Iturriaga, S. (2013). Multiobjective grid scheduling using a domain decomposition based parallel micro evolutionary algorithm. International Journal of Grid & Utility Computing, 4(1).
14. Su X, Li W. SLIDING MODE ROBUSTNESS CONTROL STRATEGY OF SHEARER HEIGHT ADJUSTING SYSTEM UNDERGROUND COAL MINES[J]. Journal of Theoretical & Applied Information Technology. 2013, 12(2).
15. Khatib O. Real-time obstacle avoidance for manipu-lators and mobile robots[J]. International Journal of Robotics Research. 1985, 1(5): 500-505.
16. Li C, Jiang X, Wang W, et al. A Simplified Car-following Model Based on the Artificial Potential Field[J]. Procedia Engineering. 2016, 137: 13-20.

Test Sequencing Optimization for a Kind of Radar Receiver

LIU Xiaopan, CAI Jinyan, WU Shihao, LI Danyang
Department of Electronic and Optic Engineering
Mechanical Engineering College
050003, Shijiazhuang
China
Email: ldy200607@126.com

Abstract. Test scheme has a great influence on the test efficiency, for a certain type of radar receiving system, detailed analysis of the main components of the system has been made, signal flow graph has been established. Fault set, test set and the fault-test dependency matrix have been established. The ant colony algorithm is adopted to test sequence optimization modeling, the transfer rule of the ant state and feedback of the pheromone has been designed, and the optimized test sequence has been shown.

1 Introduction

Test sequence optimization is a key problem in the electronic circuit testing and diagnosis technology. It has great significance to improve the efficiency of testing and diagnosis. Test sequence optimization is to determine the scope and sequence of a set of tests, which meets a certain accuracy of fault isolation, and minimizes the testing cost. Test sequence optimization problem has been proved to be a NP-complete problem. Traditional global optimization algorithms (as simplex algorithm, implicit enumerative algorithm, et al) are difficult to solve this problem, because of the amount of calculation is too large to be applied in practice. Additionally, considering the cost of test time, there is no need to deliberately pursue a global optimization. Therefore approximate algorithms are usually used to solve the test sequence optimization[1-6].

The radar receiving system consists of many modules, and is generally very complicated, so the test scheme has an important impact on the efficiency of test. Firstly, the main structure and basic operating principle of radar receiving system are detailed analyzed, a corresponding test and fault signal flow graph and dependency matrix are established. Then mathematical model of the test sequence optimization is established, using the ant colony algorithm, corresponding state transition mechanism is designed, information pheromone updating mechanism and the mutation mechanism are proposed. The final optimization scheme is given.

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_50

2 Test Sequence Optimization

Test sequence optimization problem consists of the following elements:

1) Fault states of the system are represented by a finite set F. $F=\{f_0,f_1,\ldots,f_n\}$. f_0 indicates that the system has no fault, $f_i(0\le i\le m)$ represents the i-th fault in the set F.

2) The prior probability of each fault in the system is P. $P=[p(f_0),\ldots p(f_m)]^{\mathrm{T}}$. $p(f_i)$ $(0\le i\le m)$ is the priori probability of the occurrence of the fault state f_i.

3) In the system, $T=\{t_1,t_2,\ldots t_n\}$ is a set of measurable points, and the corresponding test cost set is $C=[c_1,c_2,\ldots c_n]$.

4) Fault-test related matrix is $D_{(m+1)\times n}=[d_{ij}]$. If the test point t_j detects fault f_i, $d_{ij}=1$, otherwise $d_{ij}=0$. As shown in table 1.

Table 1. Fault−Test Related Matrix

Fault state	t_1	t_2	…	t_n	Failure rate
	c_1	c_2	…	c_n	
f_0	d_{01}	d_{02}	…	d_{0n}	$p(f_0)$
f_1	d_{11}	d_{12}	…	d_{1n}	$p(f_1)$
⋮	⋮	⋮	⋱	⋮	⋮
f_m	d_{m1}	d_{m2}	…	d_{mn}	$p(f_m)$

5) Defined test cost J

$$J = P^T AC = \sum_{i=0}^{m}\sum_{j=1}^{n} a_{ij}\, p(f_i) c_j \tag{1}$$

Where $A=[a_{ij}]$ is a $(m+1)\times n$ matrix, when the selected test t_j is used to isolate the fault state f_i , $a_{ij}=1$, otherwise $a_{ij}=0$.

Test sequence optimization problem can be described as that, a set of tests are selected in the set T, and a test sequence scheme is given, which can isolate all the fault states in the set F and minimize the cost of J.

Definition 1: Vector $s=[s_i]^{\mathrm{T}}$ $(0\le i\le m, s_i=1 \text{ or } 0)$. s is used to represent the current state of the system, if f_i is concentrated in the current fault state, $s_i=1$ otherwise $s_i=0$. The complete system fault state is expressed as $s=[1,1,\ldots,1]^{\mathrm{T}}$, if the fault $f_k(0\le k\le m)$, is isolated from the set s, $s=[0,\ldots,0,1,0,\ldots 0]^{\mathrm{T}}$, that is, $s_k=1$.

Definition 2: Defines a test vector $test_j=D(:, j)$ $(1\le j\le n)$, where j is the corresponding test with the test vector, and $D(:, j)$ is the j-th column of the D matrix. So the test t_j is selected for the current state set s, can be expressed as $s\wedge test_j$, the isolated sub sets are denoted by s_1 and s_0, then $s_1=s\wedge test_j$, $s_0= s-s_1$.

Definition 3: When $s\wedge test_{j\ne s}$ and $s\wedge test_{j\ne 0}$, t_j is an effective test for s. The following inference is available from the definition 3:

Inference 1: For a fault set with m+1 faults, all faults can be separated by at most m effective tests.

By the above definition, the test sequence optimization problem can be described as the selection of a set of test, to isolate all the faults in s, and to minimize the test cost as shown in formula (1).

3 A Certain Type of Radar Receiving System

A certain type of radar receiving system mainly includes receiver front-end, receiving channel module, intermediate frequency module, 2nd local oscillator / reference signal module, transmitter excitation module, 1st local oscillator, A/D converter and stabilizer. As shown in Fig. 1.

RF (Radiofrequency, RF) echo signals come from the PIN switch and feed to the receiver front-end, the 1st LO signal is also applied to the receiver front-end. The RF echo signal is amplified and mixed with 1st LO signal. The output of Receiver Front-End is an intermediate frequency（IF1） signal. The IF1 signal is amplified and divided into A, B, C three channels of signals. The signals are mixed with the A, B, C three channel 2nd LO signal. The output feed to the A, B, C three channel I/Q phase detector. The reference signal also feed to A, B, C three channel I/Q phase detector. After phase detection, three sets of low-frequency echo signals are supplied to A, B, C three channel A/D converter. Three channels I/Q analog signals are converted into digital signals, and are sent to the signal processing system.

At the same time, the reference signal is applied to the digital waveform generator, under the control of Trigger and Waveform code, and produces a set of linear frequency modulation pulse.

Main functions of the receiving system include: 1. Weak echo signals are amplified, frequency converted, phase detected and A/D converted into a digital signal sent to signal processing system. 2. Emitting an excitation pulse and send it to the transmission system. 3. Producing synchronous reference clock and the reference signal.

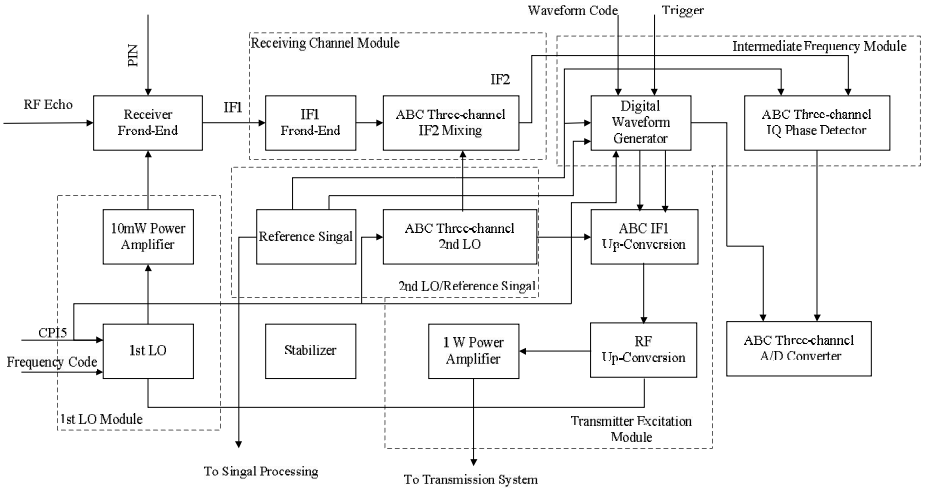


Fig. 1 Block diagram of the Receiving System

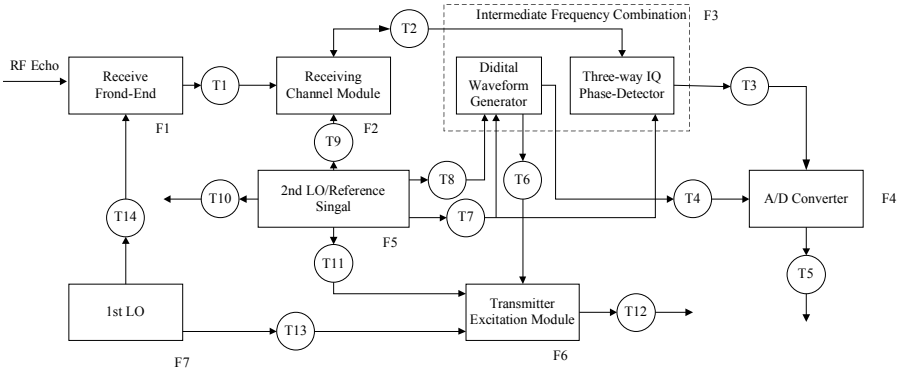


Fig. 2 Receiving System Signal Flow Diagram

Receiving system signal flow diagram is shown in Figure 2.

Table 2 Fault Set

Fault Number	Fault Description
F1	Receiver Frond-End Fault
F2	Receiving Channel Fault
F3	Intermediate Frequency Module Fault
F4	A/D Converter Fault
F5	2nd LO/Reference Fault
F6	Transmitter Excitation Module Fault
F7	1st LO Fault

T1~T14 represents a test set, as shown in the table 3：

Table 3 Test Set

Test Number	Test Description	Test Set Number	Test Description
T1	IF1 Signal Test	T8	Clock frequency Test
T2	IF2 Signal Test	T9	2nd LO Signal Test
T3	I/Q Output Signal Test	T10	Output Clock Test
T4	Clock frequency Signal Test	T11	2nd LO Signal Test
T5	A/D Output Signal Test	T12	RF Exciting Pulse Test
T6	Transmitted Pulse Signal Test	T13	1st LO Signal Test
T7	Reference Signal Test	T14	1st LO Signal Test

After the test node optimization[7], T1, T3, T5, T7, T12 are selected as the test nodes.

Firstly, the test cost of each test node is analyzed, and the test preparation and test time are mainly considered. As shown in table 4.

Table 4 Test Time of Node

Node Number	Test Time（/s）
T1	300
T3	30×3
T5	60×3
T7	30
T12	60

After the optimization, the dependence matrix as shown in table 5:

Table 5 Optimized Fault-Test dependency matrix

Fault	T1	T3	T5	T7	T12	Occurrence Probability
	300	90	180	30	60	
F0	0	0	0	0	0	0.93
F1	1	1	1	0	0	0.01
F2	0	1	1	0	0	0.01
F3	0	1	1	0	1	0.01
F4	0	0	1	0	0	0.01
F5	0	1	1	1	1	0.01
F6	0	0	0	0	1	0.01
F7	1	1	1	0	1	0.01

4 Ant colony algorithm

Ant colony algorithm is a kind of widely used heuristic search algorithm[8], just as genetic algorithm[9] and particle swarm algorithm[10], it is particularly suitable for solving programming problem problems. In this paper, test sequence optimization problem is a kind of dynamic programming problem. Traditional ant colony algorithm is difficult to solve this problem. In order to solve this problem, firstly, the full fault set $s=[1,1,\ldots,1]^T$ is defined as initial ant colony, each kind of fault state is an ant. For example, n faults in s, can be seen as n ants. The goal of ant colony is to find all fault states. When every ant finds the corresponding fault state, the ant colony search ends. When the search ends, walking distance of the ant colony is calculated as test cost J. The entire search process consists of many generations, and each generation is composed of many ant colonies. In every generation, after a certain number of ant colonies searching, the best test cost solution is saved, the search path of the solution is recorded. Pheromone vector of every node on the search path is updated according to the update rule. Until the update is complete, next generation searching starts. This process is repeated until the maximum generation.

In order to have a clear understanding of the algorithm, the algorithm flowchart is shown in Figure 3:

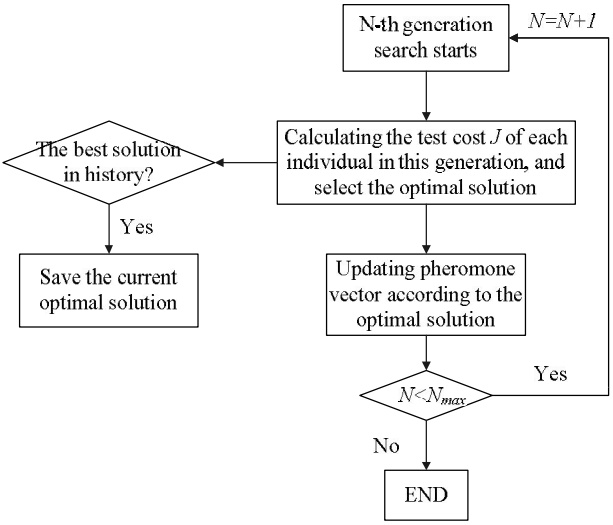


Fig. 3 Algorithm Flowchart

4.1 State transition rule

Pheromone vector is initially defined as $Tau=[1,1,\ldots,1]^{T}$, where n is the total number of tests. The transition probability of the current state s is shown as formula (2).

$$TP_k = \frac{Tau(allow(s)_k)^{\alpha} Eta(allow(s)_k)^{\beta}}{\sum_{k=1}^{length(allow(s))} Tau(allow(s)_k)^{\alpha} Eta(allow(s)_k)^{\beta}} \tag{2}$$

α represents the importance of the pheromone, and β represents the importance of the heuristic function. The value of α and β can be determined depending on the circumstances. All selectable tests are given by the *allow* function, according to the current state s. K is an index of the selectable tests. *Eta* is the heuristic function, and the expression is shown as formula 3:

$$Eta = \frac{p_{s_1}{}^2 + p_{s_0}{}^2}{p_{s_1} p_{s_0} cost_{test}} \tag{3}$$

$p_{s_1} = \sum_{f_i \in s_1} p(f_i)$ is the sum of the probability of failure in set S_1. $p_{s_0} = \sum_{f_i \in s_0} p(f_i)$ is the sum of the probability of failure in set S_0. $cost_{test}$ is the test cost of *test*.

4.2 Hierarchical weighted pheromone updating rule

In this paper, we propose a hierarchical weighted pheromone update mechanism, as shown in formula (3), (4), (5).

$$Tau(num+1)=\rho Tau(num)+\Delta Tau(num) \quad (0<\rho<1) \tag{3}$$

$$\Delta Tau(num)=[\Delta Tau_j(num)] \quad (1\le j\le n) \tag{4}$$

$$\Delta Tau_j(num)=\frac{Q}{cost}\sum_{l\in testlist}(\lambda^{l-1}\sum_{i\in testlist}p_{jl}(i)) \tag{5}$$

Initially, the value of every element in the pheromone vector is equal. In the following searches, the best pheromone vector in every generation will be updated. ρ in formula (3) is the retained coefficient. Q in formula (5) is a constant, *cost* is the best test cost in every generation, λ is the attenuation coefficient between layers, l is the number of tests used to isolate the fault state f_i, when the n-th test is performed, l=n. $\sum_{i\in testlist} p_{jl}(i)$ indicates the total probability of the test $test_j$ that is used to isolate the fault f_i in the l-th layer of the optimal test sequence.

4.3 Mutation

In order to avoid ant colony algorithm sinking into local optimization, the idea of variation in genetic algorithm is introduced. A threshold is set, and in every generation, if any element value of pheromone vector is lower than the threshold, the value of the element will be reset to the maximum. This process is called mutation. Mutation improves the ability of the algorithm to find the global optimal solution.

5 Results and conclusion

The optimal test sequence is shown in Fig. 4. 2~4 tests can locate any fault state. The optimal test cost is 251.4.

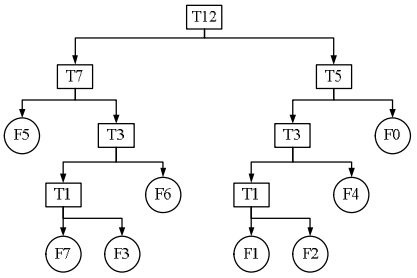


Fig. 4 The Result of Test Sequence Optimization

In this paper, the test sequence optimization problem of a certain type of radar receiving system is studied. Mathematical model of the test sequence optimization is established. The composition and signal flow diagram of the radar receiving system are established. Test set, fault set and fault-test dependency matrix are defined. To improve the ant colony algorithm, the flexible state transition rule and pheromone update mechanism are designed, hierarchical weighted and mutation are introduced into the algorithm, the result of test sequence optimization is shown. In summary, the proposed method provides a new attempt to solve the test sequence optimization problem of radar receiving system, which has engineering application value.

References

1. Krishna R. Pattipati, Mark G. Alexandridis. Application of heuristic search and information theory to sequential fault diagnosis[J]. IEEE Trans on Systems, Man and Cybernetics, 1990, 20(4):872-887.
2. Fang Tu, Krishna R. Pattipati. Rollout strategies for sequential fault diagnosis[J].IEEE Trans on Systems, Man and Cybernetics, 2003, 33(1):86-99.
3. HUANG Yifeng, JING Bo. Diagnosis strategy for multi-value attribute system based on Rollout algorithm[J]. Control and Decision, 2011, 26(8):1269-1272.
4. YANG Chenglin, TIAN Shulin, LONG Bing. Design of test points selection and test sequencing problem of multivalued fault dictionary[J]. Systems Engineering and Electronics, 2009, 31(9):2271-2275.
5. SHI Junyou, TIAN Zhong. Efficient algorithm for fault diagnosis strategy[J]. Acta Aeronautica Et Astronautica Sinica, 2003, 24(3):212-215.
6. YANG Peng. Optimization technology of design strategy based on dependency model. [D]. Changsha: National University of Defence Technology, 2008.
7. LI Danyang, MENG Yafeng, ZHU Sai, HAN Chunhui. Test node optimization with improved genetic algorithm[J]. Electronics Optics & Control，2013, 20(4):63-67.
8. P. Mathiyalagan, S. Suriya, S.N. Sivanandam. Hybrid enhanced ant colony algorithm and enhanced bee colony algorithm for grid scheduling[J]. International Journal of Grid and Utility Computing, 2011, 2(1): 45-58.
9. Amandeep Verma, Sakshi Kaushal. Deadline constraint heuristic-based genetic algorithm for workflow scheduling in cloud[J]. International Journal of Grid and Utility Computing, 2014, 5(2): 96-106.
10. K. Uchida, M. Takematsu, J.H. Lee, J. Honda. A particle swarm optimisation algorithm to generate inhomogeneous triangular cells for allocating base stations in urban and suburban areas[J]. International Journal of Space-Based and Situated Computing, 2013, 3(4): 207-214.

A Method for the Detection of Singularity in Pulse

Wei Qi [1], Hongwei Zhuang [2], Liqiong Zhang [3]
[1] Department of Information Engineering, Engineering University of the Chinese People's Armed Police, Xi'an 710086, China
wjv22@126.com
[2] College of Equipment Engineering, Engineering University of the Chinese People's Armed Police, Xi'an 710086, China
zhuanghw01@163.com
[3] Department of Electronic Technology, Engineering University of the Chinese People's Armed Police, Xi'an 710086, China
zhanglqwj@163.com

Abstract. The principle of wavelet transform (WT) about the singularity detection was used to reveal the rich cardiovascular physiological and pathological information of the pulse in this paper. The heart systole and diastole were determined accurately by the acquired WT modulus maxima figures and the detected breakpoints of the pulses. It is found that there are obvious differences in the rate of pulse wave and the ratio of the diastole and systole between the normal and the heart patient through a large number of the experimental analysis and statistics of signals, which provides a quantitative reference for some diagnosis of cardiac diseases.

1 Introduction

Human pulse signals have a great clinical diagnostic value in traditional Chinese medicine (TCM) and modern medicine (western medicine). Since the 1980s, in order to reveal the mechanism of the pulse objectively and quantitatively, many researchers have devoted themselves to the detection and analysis of human pulse signals [1-2]. From the perspective of TCM, the sphygmus is one of the important indications of illness diagnosis and analysis, because it represents the changing information of human pulse, such as frequency, rhythm, shape, depth and intensity. From a western point of view, the pulse signal is formed by heart, which is systolic and diastolic periodically, and aortic valves, which open and close instantaneously to spurt blood, causing the vibrations of the arteries and bloodstream to transmit to the periphery along the arterial system. The opening and closing features of aortic valves in pulse signals will make signals break, but these breakpoints in time-domain waveform are difficult to measure accurately.

With the development of modern signal processing method, the signal process based on the Fourier analysis is widely applied in the biomedical signal process,

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_51

obtaining a lot of scientific achievements and economic benefit. In [4] [5] [6], the author used the fast Fourier transform to obtain the power spectra of pulse signals, and extracted the features such as "spectral energy ratio ", "power spectrum picture", "the previous peak" and "harmonic number", getting the physiological and pathological information in human pulse signals effectively. The cepstrum theory was applied to analyze the subjects pulse signals of normal and abnormal heart rhythm in [7] [8], the features of it in cepstrum domain being obtained.

Since the spectrum analysis based on Fourier transform is integral in whole time-domain, it is a global transformation and can't be used in local analysis. Therefore, the time-frequency analysis is introduced for non-stationary signal in engineering practical applications, such as short-time Fourier transform, wavelet transform and Gabor expansion which are typical linear analyses of time-frequency. In the field of biomedical signal processing, the objects of the time-frequency analysis are mainly high frequency electrocardiogram, ventricular late potential, heart sound, cardiovascular sound, EEG and EMG signal [9-13]. Compared with the short-time Fourier analysis, the wavelet transform is a kind of method which the sizes of window is fixed but the shapes can be changed, namely an analysis method of local time-frequency with the time and frequency windows changing. It can well reflect the fast changing characters of signal, having great advantages in the analysis of non-stationary signal.

In this paper, we adopted the wavelet transform detecting break information effectively. After the pulse signals were processed by WT, the corresponding WT modulus maximum points of their breakpoints were very clear. We could make use of the wavelet singularity detection to obtain the systole and diastole accurately, evaluating and diagnosing the systolic and diastolic functions according to it. Compared to the electrocardiogram and electroencephalogram that we use commonly now, the modern signal process for pulse signal detection can be combined with pulse-taking in TCM so as to improve the accuracy of it, moreover, being more long-term, and used as an auxiliary mean in diagnosis and treatment.

2 The Composition of Pulse Signal Figure [3]

The recorded and plotted pulse curve with various ways and pressure is addressed as the pulse signal figure or pulse figure. The curve of pulse figure is made up of waves of which changes have certain physiological and pathological meaning. Fig.1 is a common pulse signal.

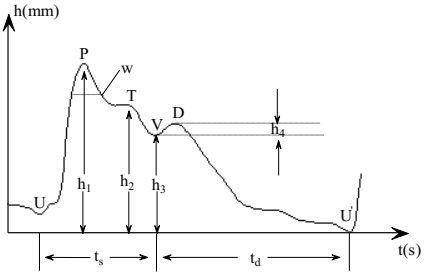


Fig. 1. Common pulse signal

U point (UP Stroke): the UP Stroke is the lowest point of the whole pulse signal figure, reflecting the end-diastolic intravascular pressure and volume. It is the initial point when the ventricle spurts blood fast, then flowing into the artery. It is also called "flowing point".

P wave (Percussion wave): the percussion wave is the main amplitude of the pulse signal figure. It is a rising curve from the baseline to the peak of the P wave. The vertex, that is the peak of the figure, reflects the maximum value of the pressure and volume in arteries. The rising limb of the percussion wave means the ventricle spurts blood fast, leading to a rapid increased arterial pressure and a sudden dilated vessel wall. The rising speed is in ratio of the height of P wave amplitude to the flowing time, and it is affected by the cardiac output, ventricular ejection velocity, arterial resistance and the elasticity of vessel wall.

P Angle (Angle of percussion wave $\angle$P): the angle of percussion wave is also called the peak angle or vertex angle, which is an angle of P wave peak. It reflects the action time of maximum pressure on the vessel wall in arteries. $\angle$P is affected by the cardiac output, arterial resistance and the elasticity of vessel wall.

T wave (Tidal wave): T wave located in recession limb is lower than P wave and higher than dicrotic wave, in delay of P wave. It reflects the left ventricle stops spurting blood, arteriectasia, anti-hypertension and antidromic reflected wave of flowing blood in artery. T wave is affected by the variety degree of the peripheral vascular resistance, the elasticity of vessel, the descent velocity of the recession limb and so on.

V wave (Valley wave): V wave is a downward notched wave which is formed by the rising limb of P wave and the recession limb of dicrotic wave. It is a cut-off point of systole and diastole, indicating the emptying time of the aortal static pressure, and affected by the peripheral vascular resistance and the descent velocity of the recession limb.

D wave (Dicrotic wave): D wave is a prominent one of recession limb waves. It is generally considered that the aortic valves are closed suddenly in early diastole phase of the ventricular, and the flowing back blood impacts on it, then being rebounded. The result is that the aortal pressure rises slightly again, and an aortal rebounded wave is formed because of slight anapetia. Therefore, D wave is affected greatly by the elasticity of vessel and the state of the closed aortic valves.

t_s (left ventricle systole): It mainly reflects the elasticity of vessel and bloodviscosity.

t_d (left ventricle diastole): It reflects the flowing state of the static pressure of blood in arteries, associated with the peripheral vascular resistance and bloodviscosity.

3 The Principle of the Signal Break Feature Extraction by WT [14]

Most of multi-scale edge detection smooth the original signal in different scales, and the upheaval point of the original signal is detected by the first or second order

derivative of smoothed signal. The smoothing function should satisfy $\int_{-\infty}^{\infty}\theta(t)dt=1$ and $\lim_{|t|\to\infty}\theta(t)\to 0$, such as the desirable Gaussian function and the canonical B spline function etc.

Generally, let $\theta(t)$ denote the Gaussian function,

$$\theta(t)=\frac{1}{\sqrt{2\pi}\sigma}\ e^{-t^2/2\sigma^2} \tag{1}$$

Let

$$\psi^{(1)}(t)=\frac{d\theta(t)}{dt},\quad \psi^{(2)}(t)=\frac{d^2\theta(t)}{dt^2} \tag{2}$$

Then,

$$\int_{-\infty}^{\infty}\psi^{(1)}(t)dt=\int_{-\infty}^{\infty}\frac{d\theta(t)}{dt}dt=\theta(t)\Big|_{-\infty}^{\infty}=0 \tag{3}$$

$$\int_{-\infty}^{\infty}\psi^{(2)}(t)dt=\int_{-\infty}^{\infty}\frac{d^2\theta(t)}{dt^2}dt=\frac{d\theta(t)}{dt}\Big|_{-\infty}^{\infty}=\frac{t^2}{\sqrt{2\pi}\sigma^2}e^{-t^2/2\sigma^2}\Big|_{-\infty}^{\infty}=0 \tag{4}$$

So we have chosen $\psi^{(1)}(t)$ and $\psi^{(2)}(t)$ as wavelet function. The function $\theta(t)$ with the introduced scale factor s is given by,

$$\theta_s(t)=\frac{1}{s}\theta\left(\frac{t}{s}\right) \tag{5}$$

$\psi^{(1)}(t)$ and $\psi^{(2)}(t)$ were taken as wavelet functions for the WT of function $f(t)$. Here the WT was defined by using convolution is given by,

$$\begin{aligned}W_s^1 f(t)&=f(t)*\psi_s^{(1)}(t)=f(t)*\frac{1}{s}\psi^{(1)}\left(\frac{t}{s}\right)=\frac{1}{s}f(t)*\left[d\theta\left(\frac{t}{s}\right)\Big/d(t/s)\right]\\&=sf(t)*\left[\frac{1}{s}\frac{d\theta(t/s)}{dt}\right]=sf(t)*\frac{d\theta_s(t)}{dt}=s\frac{d}{dt}\left[f(t)*\theta_s(t)\right]\end{aligned} \tag{6}$$

where it is not a common inner product form, without essential differences both, and similarly

$$W_s^2 f(t)=s^2\frac{d^2}{dt^2}[f(t)*\theta_s(t)] \tag{7}$$

The $W_s^1 f(t)$ and $W_s^2 f(t)$ in equation (6) and (7) denote the first and second order derivative of smoothed the function $f(t)$ by scale s respectively. So the $W_s^1 f(t)$ local extrema and the $W_s^2 f(t)$ zero-crossing point are all corresponding to the inflection points of $f(t) * \theta_s(t)$, as shown in Fig.2.

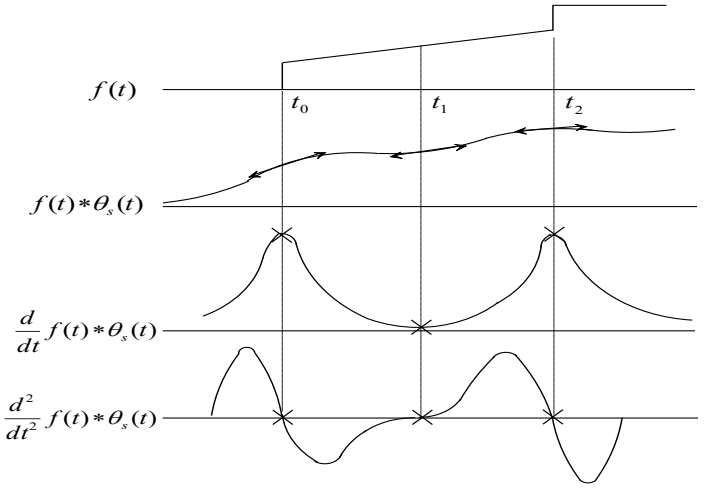


Fig. 2. The extrema of a signal local and detection of zero-crossing

The inflection point of $f(t) * \theta_s(t)$ may be the max or min of the first order derivative. In fact, only the sharp change points in $f(t)$ correspond to the maximum points in $W_s^1 f(t)$, the slow change points corresponding to the minimum points. However, the zero-crossing point in $W_s^2 f(t)$ is corresponding to all the inflection points in $f(t) * \theta_s(t)$, but it is not necessarily the breakpoint in $f(t)$. So a better result will be achieved by using the local modulus maxima points in $W_s^1 f(t)$ to detect the sharp change points in signals.

4 The Breakpoint Extraction of the Aortic Valves Opening and Closing

Human pulse signal is a weak signal, having a lower signal noise ratio (SNR). Owing to the effects of instruments and human body during the detection and acquisition of the pulse signal, there are often three kinds of noises mixed in it, which are baseline drift, interference caused by limbs tension and power line interference. In order to

improve the effect of computer analysis and diagnosis, the pulse signals must be denoised before the breakpoints extraction.

Detecting the diastole and systole is our purpose. In pulse signal diagram, the dicrotic notch is the cut-off point of the diastole and systole. The pulse signal in detection includes three key singularities that is the peak of P wave, the origin and destination point of period (O-D point), and the dicrotic notch. There are local modulus maximum because of other peaks and troughs of pulse signals. So we must find the main peaks, the corresponding WT modulus maximum of the P wave peak, the origin and destination point of period, and the dicrotic notch.

After the pulse signals are transformed by wavelet, the WT of P wave peak corresponds to the positive local modulus maximum, and the WT of period O-D point and dicrotic notch correspond to the negative modulus minimum. We should Observe the WT modulus maxima of different scales in pulse signals, find the three obvious singularities above, determine the local modulus maximum, and calculate the Lipschitz index [15]. The Lipschitz indexes α of P wave peak and period O-D point are positive, whereas the α of dicrotic notch is negative for the pulse signal. Then the WT modulus maxima of P wave peak and period O-D point will increase as the scales ascend, and the maximum of dicrotic notch will be opposite. The singularities we need can be found better on this law.

Finally, the first order derivative of Gaussian function was used as the wavelet function to extract the singularities of pulse signals by a lot of simulations on the principle of singularity detection, because the WT maxima were corresponding to the original signal turning points, determining the position of singularities more accurately. The positions of period O-D point and dicrotic notch in pulse signals were corresponding to the opened and closed points of aortic vaives, which time intervals were the systole and the diastole [16-20].

A pulse signal WT result obtained on and the local modulus maxima positions of it were given in Fig.3.

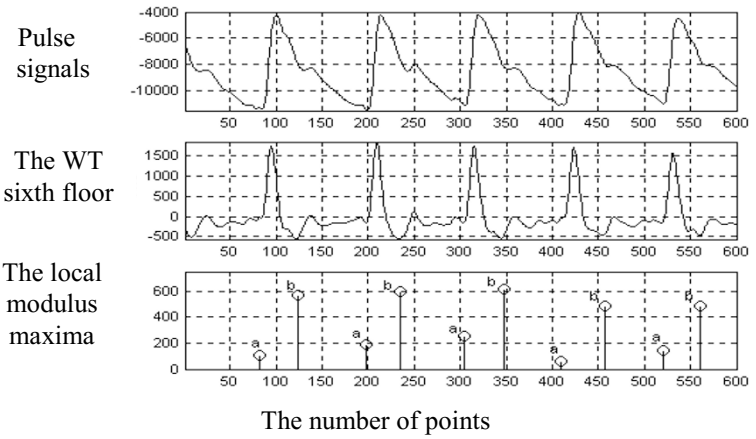


Fig. 3. The WT result of a pulse signal and its local modulus maxima positions

Compared with the change of the time-domain waveform, a point corresponds to the breakpoint when aortic vaives open, and b point corresponds to the closing moment in Fig.3. Then, the time between the two adjacent opened breakpoints (a-a) or closed breakpoints (b-b) is the basic cycle T of the heart beat; the time between the adjacent closed breakpoints and opened breakpoints (b-a) is the diastole t_d ; the time between the adjacent opened breakpoints and closed breakpoints (a-b) is the systole t_s .

The two kinds of pulse signals for the normal and the heart patients were collected respectively. Most of the patients with long-term hypertension and diabetes in clinical detecting record are abnormal in their cardiac function, many for coronary disease symptoms and a part of patients with heart failure. The pulse signals of these patients are mostly the irregularly pulse, the small and wiry pulse, the slippery and rapid pulse, the taut pulse, the retarded pulse, etc. According to the time-domain waveform and the pulse feeling and palpation in TCM, we grouped the collected pulse signals of heart patients into five: (1) the unsorted pulse of heart patients (code xz, 20 cases), the pulse like electrocardiograms (code xd, 17 cases), and (3) the irregularly pulse (code i, 73 cases), (4) the small and wiry pulse (code tw, 67 cases), (5) the slippery and rapid pulse (code z, 24 cases). In addition, there were 34 cases of the normal pulse code-named n.

We found the differences in the pulse wave rate, the diastole, the systole and the ratio of them between the heart patients and the healthy in these ways which were collected pulse signals (235 cases) discrete wavelet transform (DWT), extracting the singularities position by using modulus maximum, calculating the cycle T of the heart beat, the diastole t_d , the systole t_s and the ratio of them t_d / t_s , comparing and summarizing.

Because of the large amount of data, the average of all kinds of pulse cardiac function parameters here was only given in Table 1.

Table 1. The average of all kinds of pulse cardiac function parameters

	PW rate(n/m)	T (s)	t_d (s)	t_s (s)	t_d / t_s
n	68.8	0.8720	0.5401	0.3319	1.6251
xz	72.2	0.8311	0.4626	0.3685	1.2602
xd	66.8	0.8981	0.6308	0.2673	2.4616
i Ⅰ	36.7	1.6352	1.0182	0.6171	1.6509
i Ⅱ	70.5	0.8516	0.4659	0.3858	1.2102
i Ⅲ	21.8	2.7523	1.6267	1.1256	1.4751
tw Ⅰ	78.3	0.7665	0.4156	0.3509	1.1852
tw Ⅱ	57.8	1.0381	0.6594	0.3788	1.7445
tw Ⅲ	64.0	0.9381	0.5448	0.3933	1.3907
tw Ⅳ	68.5	0.8760	0.4857	0.3903	1.2492

tw Ⅴ	37.2	1.6146	0.9756	0.6389	1.5296
tw Ⅵ	106.6	0.5627	0.2861	0.2766	1.0368
tw Ⅶ	68.5	0.8761	0.5127	0.3634	1.4129
z Ⅰ	97.9	0.6129	0.3207	0.2922	1.0971
z Ⅱ	97.9	0.6128	0.3088	0.3040	1.0161

As you can see, the various pulse signals of the cardiac disease presented specificities in one or a few parameters of the cardiac function, and had a certain pathological significance.

5 Identification and Verification

From a medical point of view, a period of mechanical activity is composed of one systole and diastole. The time of systole is shorter than the diastole for a normal cardiomotility regardless of the atrial or ventricle. When the heart rate increases quickly, the duration of cardiac cycle is shortened, alike both the systole and diastole. But the shortened ratio of diastole is larger, even the diastole being shorter than the systole, and the ratio of them decreases obviously. Therefore, the heart rate increasing results a longer working time and shorter rest time of the myocardia relatively, which is detrimental to the enduring activities of heart. Conversely, the heart rate decreasing results a longer cardiac cycle and a larger increased ratio of diastole than systole, and the ratio of both will increase. So whether the ratio of systole and diastole is too small or too large is abnormal.

The test data we had got was conform to the above theory. Thus, we can draw the following preliminary conclusions: (1) the ratio of systole and diastole (t_d/t_s) decreased in the condition of normal heart rate for the heart patient; (2) t_d/t_s decreased significantly at the fast heart rate; (3) t_d/t_s was very high at the slow heart rate, out of the normal range.

In practical applications, the data T, t_s, t_d, t_d/t_s which have been obtained from the WT of massive normal pulse signals can be processed as a statistical way. The statistical indicators (mean and variance) will be calculated and stored into the computers for automatic diagnosis. When the unknown patients are diagnosed every time, the heart rate and t_d/t_s value will be compared with the normal values, and the degree of deviation will be observed to judge whether there has been a certain diseases of the heart or not. There are various pulse conditions such as wrong pulse rhythms, pulse positions and pulse shapes except normal pulse. We can also make an analysis and statistics according to different pulse signals so as to find out the laws between the pulse signals and symptoms, as the basis of diagnosis.

6 Conclusions

In this paper, we used the WT modulus maximum to detect the singularities of the pulse signal so as to obtain the systole and diastole accurately. We found there were obvious differences in the rate of pulse wave and the ratio of the diastole and systole between the normal and the heart patient. The normal pulse and the abnormal pulse could be preliminarily analyzed, judged and validated based on the criterion above.

References

1. M. Yoshizawa: Classical but effective techniques for estimating cardiovascular dynamics : IEEE Eng. in Med. & Bio, (1997) 106–112
2. Zhang jingren, Yang tianquan, Zheng xiuchun, et al.: Typical relative analysis of pulse signal frequency and time domain index: Liaoning Journal of traditional Chinese medicine, Vol. 22(9). (1995) 393–395
3. Huang shilin, Sun Mingyi: The Pulse in Chinese Traditional Medicine: People's Medical Publishing House, Beijing (1996)25
4. Wang Binghe, Yang Yong, Xiang Jinlin: Experiment Design and Power Spectral Characteristics of Pulse Signals: Chinese Journal of Physical Medicine and Rehabilitation, Vol. 20(3). (1998) 158–161
5. Wang Binghe, Luo Jian, Xiang Jinglin et al.: Power spectral analysis of human pulse and study of traditional Chinese medicine pulse-diagnosis mechanism: Journal of Northwest University (Natural Science Edition), Vol. 31(1). (2001) 21–25
6. WANG Binghe, WANG Haiyan, ZHI Liqin, et al.: Acquisition and spectral analysis of pulse signals of hypertension patients: Technical Acoustics, Vol. 18(3). (1999)135–138
7. Wang Binghe, Zhang Xiaomin, Xiang Jinglin, et al.: Application of Cepstrum and Homomorphic Deconvolution to Human Pulse System Analysis: Journal of Northwestern Polytechnical University, Vol. 8(4). (2000) 600–603
8. Song Jianqin, Dong Yanwu, Wu Shengju: Application of cepstral techniques to analysing pulse signals: Journal of Shaanxi Normal University (Natural Science Edition), Vol. 25(2) (1997) 35–39
9. Wang Binghe, Zangjianwei, Xiang Jinglin: Human Pulse Signal Time-Frequency Distribution and Heart Function Detection Based on Wavelet Transform: CCSP-99,551-554
10. Xu Jingling, Wu Yanjun: Generalized Time-Frequency Analysis of Biomedical Signals: Journal of Biomedical Engineering, Vol. 13(1). (1996) 67–74
11. Chai Xinyu, Wu Zhaoxia, Zhuang Tiange: Application of Time-Frequency Analysis on Medical Signal Processing: Journal of Biomedical Engineering, Vol. 18(1). (2001) 138–144
12. Zhang Feng, Chen Xiangli, Su Xiaogang: The wigner distribution application in biomedical signal analysis: Chinese Medical Equipment Journal, Vol.4. (1996) 19–23
13. Li Guonglin, Lu Weixue: Wavelet-Transform Theory and Biomedical Signals Processing: Shandong Journal of Biomedical Engineering Research, Vol. 14(2). (1995) 1–8.
14. Zhao Songnian, Xiong Xiaoyun: Wavelet Transform and Wavelet Analysis: Electronic Industry Press, Beijing (1996) 113–114
15. Yang Fusheng: The engineering analysis and application of wavelet transform: Science press, Beijing (1999)
16. Bahrepour M, Meratnia N, Poel M, et al.: Use of wireless sensor networks for distributed event detection in disaster management applications: International Journal of Space-Based and Situated Computing, Vol. 2. (2012) 58–69

17. Moore P, Thomas A, Tadros G, et al.: Detection of the onset of agitation in patients with dementia: real-time monitoring and the application of big-data solutions: International Journal of Space-Based and Situated Computing, Vol. 3(3). (2013) 136–154
18. Dai W, Liu J J, Korthaus A: Dynamic on-demand solution delivery based on a context-aware services management framework: International Journal of Grid & Utility Computing, Vol.5(1). (2014) 33–49
19. Zhao L, Wu C: Applying distribution of feature points to detect multiple plane: International Journal of Grid & Utility Computing, Vol. 3(3). (2012) 32–37
20. Indra G, Taneja R: A time stamp-based elliptic curve cryptosystem for wireless ad-hoc sensor networks: International Journal of Space-Based and Situated Computing, Vol. 4(1). (2014) 39–54

Design of non-lethal kinetic weapon impact method based on hybrid-III dummy

Wei Qi [1], Hongwei Zhuang [2], Fadong Zhao [2]
[1] Department of Information Engineering, Engineering University of the Chinese People's Armed Police, Xi'an 710086, China
wjv22@126.com
[2] College of Equipment Engineering, Engineering University of the Chinese People's Armed Police, Xi'an 710086, China
zhuanghw01@163.com, Zhaofd00@163.com

Abstract. A testing method based on the 50th percentile hybrid III dummy was designed to obtain the impacting response data of the non-lethal kinetic weapons and help to solve the difficult problems of getting target samples. Two types of non-lethal kinetic energy projectiles were used to impact the thorax of dummy in seven kinds of distances. The sensors in the 50th percentile hybrid III dummy which was applied to replace a person recorded all the impact response data of thorax. The results showed that the related response data can be obtained effectively by the designed method, and the dummy system can't be damaged with a better repeatability.

1 Introduction

Non-lethal kinetic weapons are a kind of non-lethal weapons which are mainly applied to stop people performing harmful actions and resolve the conflicts such as a single distraught individual or a large rioting crowd. They can rely on the flying kinetic energy of the projectiles to impact the body, so that the targets are too painful to resist, move or act. The projectiles, like rubber, plastic and bag bullets will cause severe blunt ballistic injury and even fatal when impacting the human body because of the higher kinetic energy and specific kinetic energy of them. Therefore, the research of the mechanical response data obtained from the non-lethal kinetic weapons impacting is an important way to improve the weapons fighting capacity and ensure the non-lethal uses of them.

The injuries caused by non-lethal kinetic weapons are blunt impact injuries [1]. Associated with the automotive crash, the dummy injuries are relatively large mass lower speed impacts [2-5]. The research conclusions can not be simply applied into the small mass, higher speed blunt ballistic injuries due to the nonhomogeneous, anisotropic, non-linear and viscoelastic characteristics of human tissue [6-7]. Moreover, the impact effect of the non-lethal kinetic energy is closely related to

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_52

impacted body parts. Clinical medical data shows that thorax region is one of the most vulnerable parts when non-lethal kinetic energy impact occurs [8-11]. Considering the less controllability, the quicker velocity decay, unstable performances and environmental influence of the non-lethal kinetic weapons [12], a new method using the hybrid III dummy to test the impacting response data of the weapons was designed. The method can measure the mechanical response data of dummy thorax effectively when weapons impacting, show the impact effect of the non-lethal kinetic weapons objectively, and prove the feasibility of making use of the dummy to test non-lethal weapons impact. The method also avoids the problems that may be from laws, ethic, and difficulties in obtaining the animal or cadaver samples during experiments.

2 Materials and Methods

2.1 Hybrid III Dummy

The hybrid III dummy has applied widely in different fields, especially in automotive crash. The skeletons and joints of its head, neck chest, pelvis and limbs all had been specially designed. It is about 178cm in height and 77.7kg in weight. The parameters of dummy such as mass, rigidity and damping are very close with human, having a similar impact biomechanical response. The test data can be recorded by the sensors and the data acquisition system placed inside the dummy body (Fig.1). Though they are accomplished without inflicting serious or lethal injuries as the name implies, the non-lethal kinetic energy munitions have been demonstrated that injuries and fatalities may still occur [13]. Therefore, using the hybrid III dummy to experiment can advoid the risk of impact injury.

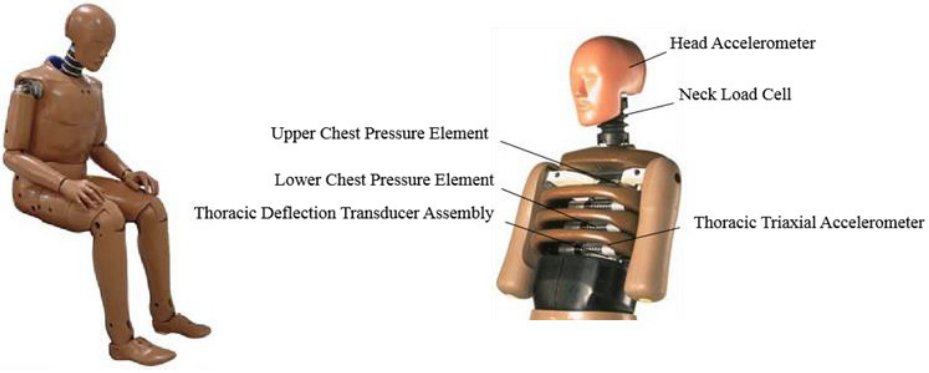


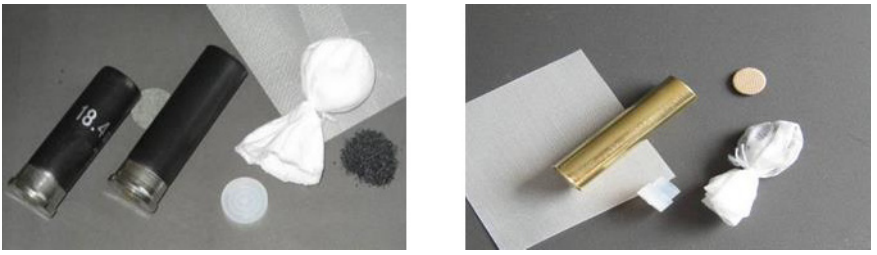
Fig. 1. The 50th percentile hybrid III dummy

2.2 Non-lethal Kinetic Energy Projectiles

There are a lot of non-lethal kinetic weapon types. According to the different fire modes, they can be divided into gas pressure, compressed gas, electrical energy, electromagnetic energy, etc [12]. 10mm and 18.4mm bag bullets were used as the projectiles of non-lethal kinetic weapon to impact in tests.

In Fig. 2(a), 18.4mm bag bullet adapts to a wide range with more flexibility. The weight of 18.4mm bag bullet is less than or equal to 50g, and the whole length is less than or equal to 65mm. Its effective range is greater than or equal to 30m. The bullet can be used not only to disperse a large rioting crowd within 5m to 30m, but also to bring suspects under control within 15m [14].

As shown in Fig. 2(b), 10mm bag bullet is a new police kinetic anti-riot munition that is designed to impact the body in a closer distance without producing severe or fatal injuries. It is both offensive and defensive strategies, but mainly self-defense. The weight of 10mm bag bullet is less than or equal to 25g, and the whole length is less than or equal to 50mm. Its effective range is greater than or equal to 15m, not resulting in fatal injuries of a living target within 5m [15].



(a) 18.4mm bag bullet (b) 10mm bag bullet

Fig. 2. 18.4mm and 10mm kinetic energy projectiles

2.3 Design of the Test Method

Referring to the practices of other impact tests, the impact mechanical response method for non-lethal kinetic weapons was designed to use the 50th percentile hybrid III dummy in this paper (Fig.3). The difficulty of this way was how to impact the frontal thorax and abdomen of the dummy target as accurately as possible when shooting. Combining with the technical and tactics indexes of the weapons, the

ballistic trajectories of the projectiles were found relatively straight and better results within 20m in the prepare experiments. And thus the safe shooting and test distances to ensure the test personnel and equipments safety were determined. At the same time, two excellent experienced shooters were chosen. Impact test cases is shown in Table 1 [16-20].

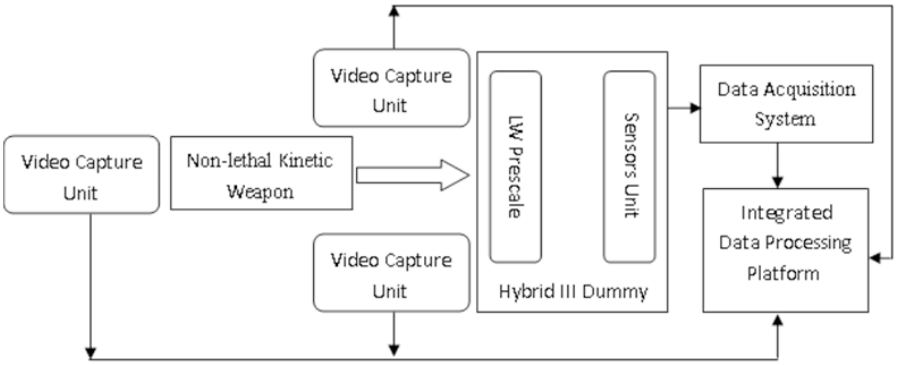


Fig. 3. Structure diagram of test method

Table 1. Impact test cases

Test Number	Non-lethal Kinetic Energy Projectiles	Impact Distance（m）
1	10mm bag bullet	3
2	10mm bag bullet	5
3	10mm bag bullet	10
4	18.4mm bag bullet	10
5	18.4mm bag bullet	15
6	18.4mm bag bullet	20
7	18.4mm bag bullet	30

3 Results

The triaxial accelerometer in dummy was used to record the acceleration parameters of the thorax impact in tests. And the chest deflection transducer assembly was used to record the displacement parameters of impact. Moreover, the impact pressure parameters depended on the LW prescale on the surface of the dummy chest to record.

3.1 The Acceleration Data of 10mm Bag Bullet Impact

The sitting dummy was impacted by the 10mm bag bullets in the distance of 3m, 5m and 10m respectively. The impacted acceleration data of the dummy thorax in three different directions was measured. The acceleration response curves and the synthetic acceleration response curves impacted in every distance also were plotted according to the data, as shown in Fig.4.

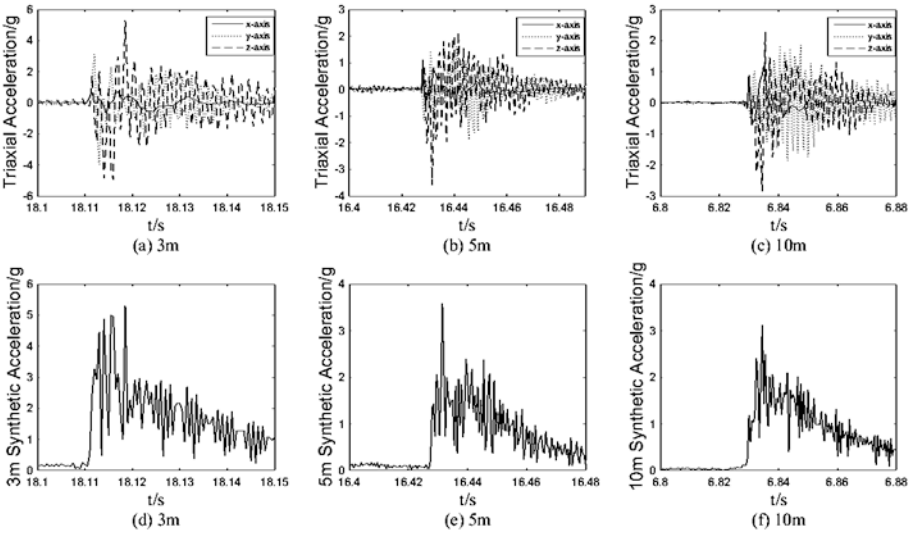


Fig. 4. Dummy acceleration curves of 10mm bag bullet impacting

3.2 The Acceleration Data of 18.4mm Bag Bullet Impact

The sitting dummy was impacted by the 18.4mm bag bullets in the distance of 10m, 15m, 20m and 30m respectively. The impacted acceleration data of the dummy thorax in three different directions was measured. The acceleration response curves and the synthetic acceleration response curves impacted in every distance also were plotted according to the data, as shown in Fig.5.

3.3 The Displacement Data of Two Types Bag Bullet Impact

The sitting dummy was impacted by the 18.4mm and the 10mm bag bullets in the distance of table.1 respectively. The impacted displacement data of the dummy thorax was measured. The deflection response curves impacted in every distance also were plotted according to the data, as shown in Fig.6.

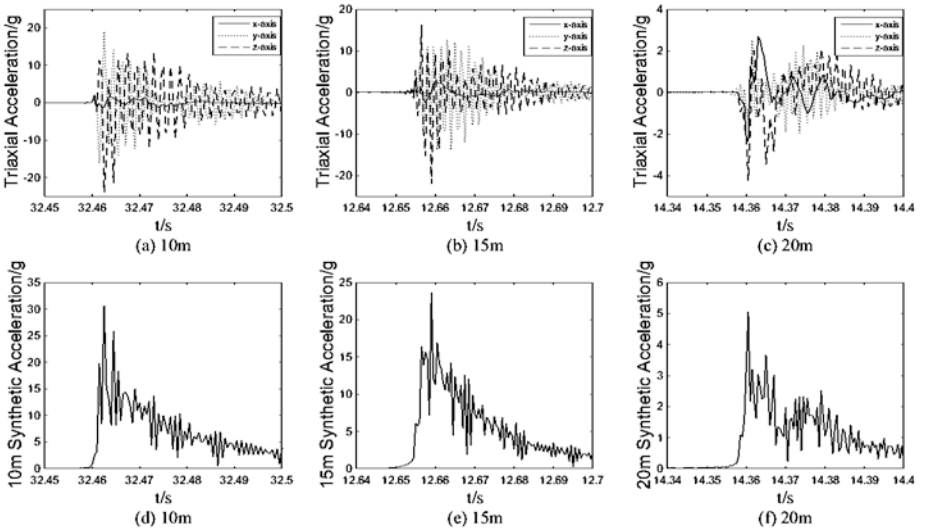


Fig. 5. Dummy acceleration curves of 18.4mm bag bullet impacting

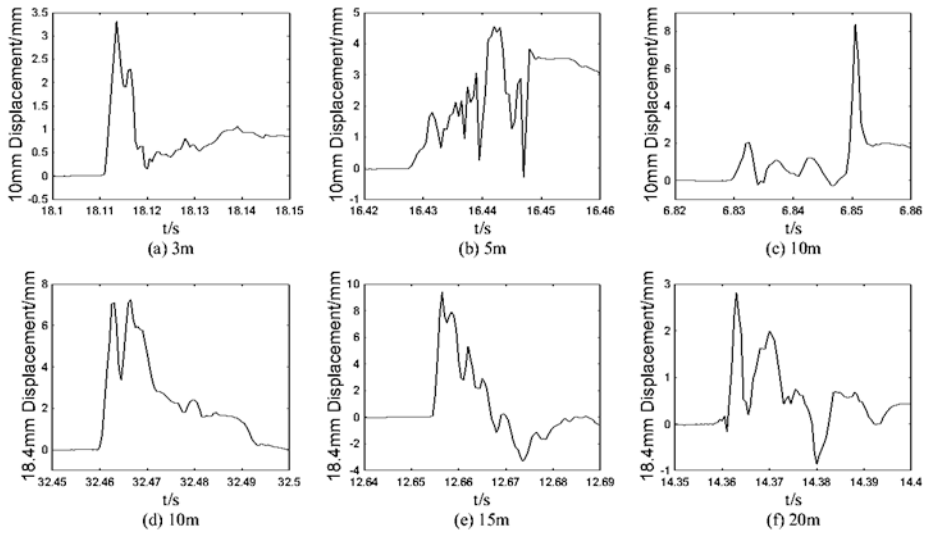


Fig. 6. Dummy impact deflection curves

3.4 The Pressure Data of Two Types Bag Bullet Impact

The sitting dummy was impacted by the 18.4mm and the 10mm bag bullets in the distance of table.1 respectively. The impacted pressure data of the dummy thorax was measured and recorded by the LW prescale, as shown in Fig.7.

Fig. 7. LW prescale impact results

4 Discussion

Due to the little use of the hybrid III dummy in the impact experiments of the non-lethal kinetic weapons, the thorax injury of the dummy is still evaluated and analyzed according to the Abbreviated Injury Scale (AIS) set by American Association for Automotive Medicine and the CMVDR294 criterion , as shown in Table 2.

Table 2. Injury criteria of acceleration and displacement [21-22]

Chest Acceleration (≤3ms)	Chest Injury	Chest Compression Deformation
17~37g	Single rib fracture	≤75mm
38~54g	2 to 3 rib fractures; sternum fracture	
55~68g	4 or more rib fractures; 2 to 3 rib fractures with hemothorax or pneumo-thorax	
69~79g	Greater than 4 rib fractures with hemo-thorax or pneumo-thorax; flail chest	
80~90g	Aorta laceration (partial transection)	
>90g	Non-survivable	

In accordance with the provisions of Table 2, the peak value of synthetic acceleration measured during impacting the dummy thorax at less than or equal to 3m

is selected to analyze in this paper. And as shown in Table 3, the maximum deformations produced when the two projectiles impacted the chest in various distances are listed on the basis of the Fig.6.

Table 3. Acceleration and displacement peak values

Projectile	Impact Distance(m)	Duration of Time(s)	Synthetic Acceleration peak(g)	Time(s)	Displacement peak (mm)
10mm bag bullet	3	18.1185-18.1215	5.3014	18.1135	3.3232
	5	16.4315-16.4345	3.5880	16.4420	4.5518
	10	6.8345-6.8375	3.1108	6.8505	8.3865
18.4mm bag bullet	10	32.4625-32.4655	30.6213	32.4665	7.2433
	15	12.6590-12.6620	23.5893	12.6565	9.3760
	20	14.3605-14.3635	5.0446	14.3630	2.8194

4.1 The Acceleration Data Analysis

Because of the trigger of the data acquisition system in dummy and the shooting preparation time, the acceleration data of the 10mm bag bullets impacting is recorded at 18.1s, 16.4s and 6.8s in distances of 3m, 5m and 10m respectively. Similarly, the acceleration data of the 18.4mm bag bullets impacting is recorded at 32.4s, 12.6s and 14.3s in distances of 10m, 15m and 20m (Fig.4 and Fig.5). The trends of acceleration curves in Fig.4 and Fig.5 accord with the impacting general rules of the non-lethal kinetic weapons. The data is also available.

The maximum of the thoracic synthetic acceleration peak values is 5.3g (Table 3). This value appears at 3m distance from the dummy under the conditions of 10mm bag bullets and less than or equal to 3ms, which is below 17g, the minimum acceleration value of thoracic injury level 1 in Table 2, and will not cause injuries. Under the conditions of 18.4mm bag bullets and less than or equal to 3ms, the maximum of the thoracic synthetic acceleration peak values is 30.6g in Table 3, which appears at 10m distance from the dummy.

The value is above the minimum acceleration value of thoracic injury level 1 in Table 2, but below the maximum. This will cause mild injuries likely. The six groups of test results are no more than the value in FMVSS208 criterion, that when the synthetic acceleration is less than or equal to 60g, the impact will not cause injury for human.

4.2 The Displacement Data Analysis

As shown in Fig.6 above, the displacement data of the 10mm bag bullets impacting is recorded at 18.1s, 16.4s and 6.8s in distances of 3m, 5m and 10m respectively after the data acquisition system in dummy is triggered. And the displacement data of the 18.4mm bag bullets impacting is recorded at 32.4s, 12.6s and 14.3s in distances of 10m, 15m and 20m respectively

The curve trend of the deflection in Fig.6 is consistent with the acceleration curve trend at the same distance, and the time also keeps consistent with the peak values of the acceleration at the same distance in Fig.4 and Fig.5. That is to say the deformation of the impacted dummy thorax is the maximum when the acceleration is the maximum.

The maximum of the displacement peaks impacted by both of projectiles in different distance is 9.4mm in Table 3. It appears at 10m distance from the dummy when 18.4mm bag bullets impacting. The value of 9.4 mm is far lower than what the CMVDR294 criterion has provided in Table 2, which will not cause injury for human when the deformation of the chest is less than or equal to 75mm.

The displacement of the 10mm bag bullet impacting increases with increase of the impact distance in Table 3. The results show the projectile is accelerating. The displacement of the 18.4mm bag bullet impacting at 20m distance from the dummy is less than the value at 15m, which shows that the projectile is slowing down after 15m flight. The dummy was impacted from the distance of 30m in test, but the kinetic energy of the projectile lost so much that it can not hit the dummy at all. Therefore, the data at the distance is not shown in any figure and table, which also proves that the data obtained when the 18.4mm bag bullet slowing down after 15m flight is reasonable. The data in Table 3 and the technical and tactics indexes of the weapons described in previous parts of this paper are basically the same.

4.3 The LW Prescale Data Analysis

As shown in Fig.7, the LW prescale, that is a kind of pressure sensitive paper, is stuck on the skin surface of the dummy thorax. After the bag bullets impacting, the microcapsules in color layers fracture and the chromogenic material is absorbed to form obvious red spots. Neither the pressure sensitive paper nor the dummy skin is broken. The pressure sensitive paper still need to be processed by the pressure measurement system according to the degree of its discoloration in order to get the impacting pressure distribution, which can prove the rationality of the acceleration data indirectly.

5 Conclusions

A new testing method that uses the 50th percentile hybrid III dummy to obtain the response data of non-lethal kinetic weapons impacting is designed in this paper. The method can be applied to most of non-lethal kinetic weapons, and solve the difficult problems of getting experiment samples. The non-lethal kinetic energy projectiles

were impacted on the dummy during experiments. The results of experiments have proved the method universal, repeatable and valuable in military application.

The 10mm bag bullets in range of 3-10m, and the 18.4mm bag bullets in range of 10-30m will not cause serious kinetic energy injuries of the 50th percentile hybrid III dummy system. The sensors inside the dummy system do not fall off. Furthermore, the skin and skeleton of the dummy are not penetrated and injured. The data such as acceleration, displacement and pressure could be obtained effectively to evaluate the mechanical response of the dummy thorax. The velocities of the projectiles can be measured with the help of the high-speed camera or other technical means, and the thoracic pressure distribution also can be analyzed after impacting by pressure sensitive paper and its analysis system.

The experimental research, testing and evaluating method in this paper are important guide to develop, design and evaluate the new non-lethal kinetic weapons, which may provide the reference values for the research of police equipments and other impacting protective equipments. The incomparable biomechanical experiment provided by the dummy will further open up its application prospects. It is of great importance for studying intensively the mechanical response of the non-lethal kinetic weapons impacting the body that making use of the obtained impact results in experiments.

References

1. Bir C: Ballistic injury biomechanics. In: Yoganandan N, Nahum A M, Melvin J W, et al. Accidental Injury-Biomechanics and Prevention, Springer-Verlag, New York (2015) 829–839
2. Lan Fengchong, Cai Zhihua, Chen Jiqing, et al.: Biomechanical responses and injury evaluation of human thorax and abdomen during vehicle collision: Journal of South China University of Technology (Natural Science Edition), Vol. 40(12). (2012) 70–78
3. Nahum A M, Melvin J W.: Accidental Injury-Biomechanics and Prevention. Second Edition, Springer-Verlag, New York (2001) 51–65
4. Ma Xiaolin, Yang Zhihuan: Study of biomechanical mechanism about blunt impact injury of the liver: Chinese Journal of Trauma, Vol. 16(8). (2000) 508–509
5. Cai Zhihua: Study on biomechanical responses and injury assessment of human thorax in vehicle collision: South China University of Technology, GuangZhou (2013)
6. Fung Y. C: The application of biomechanics to the understanding of injury and healing. In: Nahum A M, Melvin J W. Accidental Injury-Biomechanics and Prevention, Springer, New York (1993) 1–11
7. Eck J V: Biomechanical response and abdominal injury due to blunt ballistic impacts :Wayne State University, Detroit (2006)
8. Bir C, Eck J: Preliminary analysis of blunt ballistic impacts to the abdomen. In: Gilchrist M D. IUTAM Symposium on Impact Biomechanics: From Fundamental Insights to Applications, Springer, Netherlands (2005) 25–32
9. Steele J A, McBride S J, Kelly J, et al.: Plastic bullet injuries in Northern Ireland: experiences during a week of civil disturbance: Journal of Trauma, Vol. 46(4). (1999) 711–714
10. Mahajna A, Aboud N, Harbaji I, et al.: Blunt and penetrating injuries caused by rubber bullets during the Israeli-Arab conflict in October, 2000: a retrospective study, Lancet, Vol. 359(9320): (2002) 1795–1800

11. Hubbs K, Klinger D.: Impact munitions data base of use and effects [EB/OL]. https://www.ncjrs.gov/App/Publications/abstract.aspx?ID=204433. 2002
12. Xiang Hongjun, Li Zhiyuan, Lei Bin, et al.: Research on developing situation and trend of nonfatal kinetic weapon: Journal of Sichuan Ordnance, Vol. 32(9). (2011) 36–38
13. Bir C, Viano DC.: Design and injury assessment criteria for blunt ballistic impacts: Journal of Trauma, 2004, Vol. 57(6). 1218–23
14. He Haijun, Zhang Liguo: Domestic new police kinetic anti-riot munition-18.4mm bag bullet: Small Arms, (14). (2005) 22–23
15. Zou Jijing: Analysis on the new police weapons-2010 type 10mm anti-riot revolver system: Police Education and Training, (5). (2011) 40–42
16. Ahmad N, Imran M, Khursheed K, et al.: Model, placement optimisation and verification of a sky surveillance visual sensor network: International Journal of Space-Based and Situated Computing, Vol. 3(3). (2013) 125–135
17. Xia J, Yun R, Yu K, et al.: A coordinated mechanism for multimode user equipment accessing wireless sensor network: International Journal of Grid & Utility Computing, Vol. 5(1). (2014) 1–10
18. Yuriyama M, Kushida T: Integrated cloud computing environment with IT resources and sensor devices: International Journal of Space-Based and Situated Computing, (2011) 1
19. Bahrepour M, Meratnia N, Poel M, et al.: Use of wireless sensor networks for distributed event detection in disaster management applications: International Journal of Space-Based and Situated Computing, Vol. 2. (2012) 58–69
20. Dutu C, Apostol E, Leordeanu C, et al.: A solution for the management of multimedia sessions in hybrid clouds: International Journal of Space-Based and Situated Computing (IJSSC), Vol. 4(2). (2014) 77–87
21. Yu Zhigang: Research on the difference of frontal impact response between Chinese people and the standard test dummy : Hunan University, ChangSha (2008)
22. Wei Liang, Zhang Lele, Cui Jin, et al.: Response simulation and injury prediction of standing dummy in a subway collision: Journal of the China Railway Society, Vol. 37(1). (2015) 16–23

Review and Research Development of Optimization Algorithms Based on Wolf Intelligence

Sun Yixiao[1,2], Zhan Renjun[1], Wu Husheng[1], Han Zhexin[2], Ma Yanbin[3]
[1] Equipment Engineering College, Engineering University of APF, Xi'an, China
32692660@qq.com, zhanrenjun@aliyun.com, wuhusheng0421@163.com
[2] Information Engineering Department, Officers College of APF, Chengdu, China
328693098@qq.com
[3] The Second Team of The First Battalion, Guangzhou Command College of APF, Guangzhou, China
2422343997@qq.com

Abstract. The article firstly introduces two respective theories on research and evolvement about Wolf Pack Intelligent Optimization Algorithm, and compares with the differences and similarities between the two theories. It also illustrates the differences in performance on optimization solution through the experimental result. Then, it concludes the improvement research on wolf pack algorithm, focused on the improvement on parameter setting and hybrid algorithm. Thirdly, the article elaborates the typical applications about wolf pack algorithm on function optimization, combination optimization and engineering optimization. At last, the article summarizes the deficiency in the research and proposes the research goal of next step.

1 Introduction

Optimization algorithm is an effective tool[1] that used to reach the best index on performance for the model or system. It mainly divides into two categories, classic optimization algorithm and intelligent optimization algorithm. Classic optimization algorithm, for example, the steepest gradient descent method, Newton method, simplex method and so on, which though has strict mathematical analysis model, has its special scope of application. Intelligent optimization algorithm stems from in 1960s, people regards the intelligence from the collective behavior among social animals and insects as swarm intelligence, featuring simple behavior and limited intelligence. But under the encouragement of the same goal, group cooperation and self-organization can spring up the group behavior which is high complex and intelligence. People puts forward the intelligent optimization algorithm through the research of compound animals' exchange mechanism, simulation of behaviors and mapping relation in optimization. Intelligent optimization algorithm which combined with many solutions through some or whole individuals' information exchange is an

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_53

algorithm[2] to find a solution of optimization problems. The typical intelligent optimization algorithm, ant colony optimization[3], put forward in 1992 inspired by ant colony sharing mechanism which Dorigo found it. Kennedy, America's social psychologist and Eberhart, America's electrical engineer put forward particle swarm optimization algorithm[4] in 1995 according to the migration and group behavior's simulations of the bird flock's frying process in foraging. Karaboga put forward artificial bee colony algorithm[5] according to the concept of queen, colony reproductive capacity and foraging pattern. And then according to the observation of lots of organisms, a lot of scholars put forward many swarm intelligent algorithm such as fish swarm algorithm[6], shuffled frog leaping algorithm[7], dolphin swarm algorithm[8] and so on.

Wolf pack algorithm is a bionic swarm intelligent optimization algorithm which is recently proposed. Although it lacks of complete theories of mathematics, it has a good optimal performance from confirmation of previous studies, and it shows a good application prospect in many fields. Therefore, it is of great significance to study the theory and application of wolf pack algorithm.

2 The Basic Principle of Wolf Pack Algorithm

Wolf locates in the second top of the biologic chains, it's a highly social species and often be a group in 5 to 12 wolfs. At the same time, they have extremely strict group hierarchy, different hierarchy wolfs work together to hunt, live together and suckle their young. In 1970, Mech, the famous expert who published a book The Wolf. The Ecology and Behavior of Endangered Species[9] studied ecosystems and behaviors of the wolf pack. The biology backgrounds of two wolf pack intelligent algorithms in this article are both come from the book. The two algorithms which one is wolf pack algorithm[10] put forward by Wu shenghu in 2013, and the other is grey wolf optimizer put forward by Seyedali Mirjalili[11].

2.1 The Basic Theory of WPA

Individual Level Division of Wolf Pack. According to the theory of system, divided responsibilities and cooperation principles, wolf pack can divided into the head wolf, exploring wolf and violent wolf. The head wolf makes decisions and be responsible for the whole wolf pack's command according to the hunting information they feel. Exploring wolves belong to the acumen, they can explore the environment by themselves in a adjustable range, and move to the place which is the concentration of strongest flavors. Violent wolves catch their prey according to the head wolf's call. In the practical algorithm design, all the wolves are artificial, exploring wolves and violent wolves can swap their roles according to each iterative value, and can enhance the searching capability of the algorithm.

The Relationship between Artificial Wolves and Optimization Problem. The mission of head wolves is to precisely locate their prey, that is the optimal solution x. The objective function value of the optimal solution, $Y = f(x)$, is the strength of the odors perceived by the head wolf. The position of artificial wolves is correspond to the candidate solution x_i. The objective function value of the candidate solution, $Y_i = f(x_i)$, is the strength of the odors perceived by the artificial wolf. N represents the number of artificial wolves. The searching and hunting process of wolf pack are the process which finds the optimal solution.

The Main Body of Algorithms. It's composed of two intelligent rules (the generation rule of head wolves called "winner-takes-all", strong survived and eventually prospered) and three behaviors (migratory behavior, summon behavior and siege behavior). Head wolf's producing is the position of artificial wolves which is an optimal function value after each iteration. The strong survived, the weak R went to the ground. Artificial wolves can be regarded as exploring wolves in migratory behavior except for the head wolf, they move a step to h direction in the next iteration, and record the value of the objective function obtained in each step, then return to the original place, and move to the p direction which is the best objective function value, the formula is: $x_{id}^{p} = x_{id} + \sin(2\pi \times p / h) \times step_a^d$. In the summon behavior, the left wolves are violent wolf when eliminates head wolves and exploring wolves which are produced by each iteration, they move quickly to the direction of head wolves. The formula is : $x_{id}^{k+1} = x_{id}^{k} + step_b^d \cdot (g_d^k - x_{id}^k) / \left| g_d^k - x_{id}^k \right|$. In the siege behavior, the wolf will together move to the grey step by step and hunt them, the formula is : $x_{id}^{k+1} = x_{id}^{k} + \lambda \cdot step_c^d \cdot \left| G_d^k - x_{id}^k \right|$. The relationship of three step length is: $step_a^d = step_b^d / 2 = 2 \cdot step_c^d$.

The Process of Algorithm. Initialization, migratory behavior (who can be the head wolf if it has the best objective value according to movement comparison between exploring wolves and head wolves), summon behavior (the position of violent wolves can be recorded in the movement, and compared to head wolves, then the better will be the head wolf) find the distance between head wolves and other wolves in the optimal iteration, and start the siege behavior if the distance is small enough (In the siege behavior, all the artificial wolf would record all objective function value except head wolves, the best would be the next iteration's head wolf) until output the optimal solution.

2.2 The Basic Principle of GWO

Individual Level Division of Wolf Pack. According to the cooperation mode and categories in the nature, wolf pack can be divided into α wolf, β wolf, δ wolf and ω wolf. α wolf is the headmaster in the wolf pack, commanding wolf pack's behavior, it would also improve the decision according to the other wolves' information. β wolf is the second master compared to α wolf, and it is the candidate of the next headmaster. β wolves can command the lower-tiered wolves, and offer or feedback information so as to help α wolf make decisions. ω wolf is the lowest grades, they have to obey to other wolves. The rest of wolves are δ, although they can't compare with α and β wolves, they can command behaviors of ω wolves. In the practical algorithm design, all the wolves are artificial, they can swap roles according to each iterative value, and can enhance the searching capability of the algorithm.

The Relationship between Artificial Wolves and Optimization Problem. The position of the grey is decided by the common position of α wolf, β wolf and δ wolf, that is the optimal solution $\vec{X}$. The objective function value of the optimal solution, $Y = f(X)$, is the strength of the odors. The position of artificial wolves is corresponded to the candidate solution X_i . The objective function value of the candidate solution, $Y_i = f(X_i)$, is the strength of the odors perceived by the artificial wolf. N represents the number of artificial wolves. The searching and hunting process of wolf pack are the process which finds the optimal solution.

The Main Body of Algorithms. It's composed of two intelligent rules (the establishment of social classes and hunting rules) and three behaviors (tracking behavior, bounding behavior and attacking behavior). The optimal function value of the position for artificial wolves obtained by each iteration respectively is α wolf, β wolf and δ wolf. The rest is defined as ω wolf or other hunters. The action rule is the next step which determined by the optimal location of three artificial wolves. In tracking behaviors, the distance between each wolf and the prey is defined as $\vec{D} = \left| C \cdot \vec{X}_p(t) - \vec{X}(t) \right|$, the next iteration position of wolves is $\vec{X}(t+1) = \vec{X}_p(t) - \vec{A} \cdot \vec{D}$. In bounding behaviors, the optimal solution of the three artificial wolf can form the next iteration of the solution space, for example, a circle can be formed in the 2D vector, a cube can be formed in the 3D vector, in the next iteration, ω wolves and other hunting wolves can be randomly moved into any location in the solution space, the formula is $\vec{X} = \frac{\vec{X}_1 + \vec{X}_2 + \vec{X}_3}{3}$. In attacking behaviors, wolves reach the position of prey by setting the vector $|A| < 1$.

The Process of Algorithm. The interaction number of initialization, the establishment of the wolves' level, tracking behaviors, bounding behaviors and attacking behaviors reaches the predetermined value, the optimal solution will be output.

3 The Similarities and Differences between Two Algorithms

3.1 Social Class of Wolf Pack

Similarities: *1. The gradation of individuals.* The two algorithms for candidate solution are both came from leadership rules for social grade in nature, and the excellence of the wolves, at the same time, in the hunting process which is the way to find the optimal solution, rely on the strength of the wolf groups to update information so as to get results. *2. Individual role exchange.* The position of leader wolves is usually the position of the grey. It calls to other wolves to move to its position so as to capture the prey. In algorithm modeling, the two algorithms define the gradation of individuals by the proximity of the candidate solution to the optimal value. After each iteration, individual role exchanges are redefined reasonably according to the change value of candidate solution.
Differences: The number of gradation and each grade's artificial wolves are different in the two algorithms. In WPA, the artificial wolves are divided into head wolves, exploring wolves and violent wolves. The head wolf has optimal function value after each iteration, the exploring wolf with a better function value, the rest of artificial wolves are violent wolves. In GWO, the artificial wolves are divided into α wolf, β wolf, δ wolf and ω wolf. α wolf, β wolf , δ wolf is the artificial wolf with the first three optimal function value after each iteration. ω wolf is the artificial wolf with the worst function value, the rest of wolves are alternative in each iteration .

3.2 The Function of Artificial Wolves

Although the divisions of two algorithms are both from their social gradation, and have similar importance in the artificial wolves, it has different function in hunting process which is the way to find solutions.

In WPA, the optimal value is the head wolf, and does not take part in position transformation in the next iteration. Exploring wolves take part in the next iteration, and the better value will be the head in the next iteration. Violent wolves begin iteration when the position between the artificial and the grey is less than a certain range. At the same time, it compares with the position of other artificial wolves in the process of iteration, and be the head wolf if has optimal value. In the solving process, different grade has different function.

In GWO, although α wolf, β wolf and δ wolf are in a different artificial grade because of different function value, they are composed to a reference value for candidate solution in each iteration. ω wolves and other alternative wolves randomly move to a position of solution space so that they can choose the role in the all artificial wolves again before next iteration. The solving process of α wolf , β wolf and δ wolf have the same functions, making sure the solution space for ω wolf and other candidate wolves .

3.3 The Way of Avoiding to Fall into Local Solution

Many optimal value function which need to solve are multimodal functions with multiple local minima. Some algorithms are easy to falling into local solution or the shocking in the local solution[12], it's hard to find the global optimal value. A good algorithm should avoid falling into local solution. In the algorithm design of WPA and GWO, they find a way to avoid falling into local solution by making use of the random evaluation, and the rules are completely different.

1. WPA uses population renewal mechanism. Wolves have a very cruel population renewal mechanism in nature, a few wolves will be eliminated through the inside battle and food distribution, but this rule can make sure that the competent wolves are capable of strong bodies so that to maintain the continuation and development of the main body of wolves. Survivor Series in WPA algorithm can eliminate the worst objective function value R in each process of algorithm's iteration, at the same, it can randomly produce a new value R on behalf of artificial wolves. The evaluation of R are the random integer from $\left[n/(2\times\beta), n/\beta\right]$, which keeps the diversity of individuals. It can also develop a new solution space, avoiding effectively the algorithm to fall into a local solution.

2. GWO uses the way of adjusting parameter value, and behavior of hunting wolves are decided by the position of α wolf, β wolf and δ wolf. The range of solution space will be bigger when they disperse to several directions, which is mainly used for attacking the prey. It can adjust the size of solution space by changing the value of $\vec{A}$, according to the formula $\vec{X}(t+1) = \vec{X}_p(t) - \vec{A}\cdot\vec{D}$, choosing the random parameter when $|A| > 1$, making the range of optional position bigger than ever, avoiding effectively the algorithm to fall into a local solution.

3.4 The Operation Time and Termination Condition

In the process of WPA and GWO, they assume that the artificial wolf is aware of the location of the prey, but it still need other termination conditions according to complex processes of function which need to be solved .

The solving process in WPA can be divided into circulations of three big modules, the circulation process of migratory behavior, the circulation process of summon behavior and the circulation process of integral solving solution. In the process of migratory behavior, it usually seeks for a better solution through exchanging the role of the head wolf and exploring wolf. The termination condition is to find an elegant solution or reach to iteration times. In the process of summon behavior, it seeks for a better solution through exchanging the role of the head wolf and violent wolf. The termination condition is to find an elegant solution or reach to a smaller value through the better solution and elegant solution. The termination condition of integral circulation is to find the optimal solution or reach to the circulation time.

The solving process in GWO is the calculation of all candidate solutions in each iteration, consequently the termination condition is to reach the circulation time, and the termination condition of each iteration is that ω wolves and other candidate wolves whether find the movement position, so the number of cycles depends on the number of initial values.

3.5 The Comparison of Calculation's Result

To compare the performance of two algorithms, this article chooses Sphere function, Rastrigin function in multimodal separable function, and Six Hump Camel Back function in multimodal inseparable function to have the test on the two algorithms. The formula shows in table 1, and the result of calculation shows in table 2.

Table 1. Test function.

Function	Expression	Dimension	Data range	Theoretical optimal solution
Sphere	$f(X) = \sum_{i=1}^{D} x_i^2$	30	[-1.5，1.5]	0
Rastrigin	$f(X) = \sum_{i=1}^{D} [x_i^2 - 10\cos(2\pi x_i) + 10]$	60	[-10，10]	0
Six Hump Camel Back	$f(X) = 4x_1^2 - 2.1x_1^4 + \frac{1}{3}x_1^6$	2	[-5，5]	-1.0316

Table 2. Experimental results.

Function	Example		GWO algorithm	
	Ave	Std	Ave	Std
Sphere	0	0	6.59E-28	6.34E-0.5
Rastrigin	0	0	0.310521	47.35612
Six Hump Camel Back	-1.0316	0	-1.0316	0

From table 2, it shows that in the solving process of Low dimensional functions, 2 dimensional function of Six Hump Camel Back, although it is the multimodal inseparable complex function, the two algorithms both accurately work out the optimal solution and the variance is a zero. In solving two high-dimensional functions, WPA successfully works out the value, and the standard deviation is zero, but there is a little difference between average value and theoretical optimal solution in GWO. The results of above algorithms are from calculation processes in different references, and the adopted iteration number are not the same, so it can't prove that the computational accuracy of GWO is not as good as WPA algorithm. Compared with other classical algorithms, under the same iteration times, GWO with better accuracy can effectively avoid falling into local solution in the literature [11].

4 The Improvement of Wolf Pack Algorithm

4.1 The Improvement of Parameter Settings

How to set parameter which is one of difficult problems faced by many algorithms, and how to effectively solve the problem in parameter settings is an important direction of algorithms improvement. The main parameter of GWO are just A vector which take part in the next calculation of position movement in artificial wolfs and C vector which take part in the position judgment between artificial wolves and grey. In order to effectively solve the practical problems, it usually change the range of two values. However, there exists several problems in parameter settings in WPA algorithm, including exploring the scaling factors α for wolves, migration times Tmax, distance judgment factor ω, step length factor S, update scale factor β.The scholars' researches and experiments showed that: 1. 2-dimensional multi peak non separable function in Bridge function for measuring determinant factor ω and step length factor S according to reference [10] showed that ω is an important parameter which control algorithm execution changes from summon behavior to attacking behavior, its value has an important influence on the accuracy of the algorithm. The value of S can be obtained only need to limit a relatively large range when the value of other parameter is fixed. Its tiny change has little effect on the property, so the algorithm is not sensitive to step length S; 2. In reference [13], scholars found that in the process of optimizing, the greater the amount of artificial exploring wolves and wolfs are, the higher accuracy can be obtained. With the improvement of properties in computers, the number of initial artificial wolves will not affect the calculation speed, so the scale factor α of exploring wolves can be deleted.

4.2 The Improvement of Complex algorithm

According to "no free lunch" theories[14], None of the swarm intelligent optimization algorithms are perfect so far. However, different algorithms have different optimization mechanism in the optimization process, which has better performance in solving some problems. There are some scholars who have combined wolf pack algorithm with other classic algorithms so as to get better optimization performance. Reference [15] embedded searching methods into the colony algorithm for solving complex function optimization and TSP problems. Reference [16] used the distribution theory of wolf pack as reference to update ant colony algorithms. Reference [17] also makes use of the distribution of wolf pack to update population.

5 Wolf Pack Algorithm's Application and Research

5.1 Function Optimizing

Swarm intelligent optimization algorithm was put forward to settle the optimization application at first. In practical application, function optimization problems can divided into continuous space optimization and discrete space optimization according to the difference in solution space. Continuous space optimization applied Improved parameter setting pack IWPA[18] in high-dimensional complex function solution based on WPA. The experiments proved that it has a good solving precision compared to lots of classic algorithms. Catastrophe wolves algorithm CWPA[18] and OWPA[19] algorithm was used to reverse the wolves in continuous function optimization, the experimental results show that the two algorithms can effectively jump out of local solution and accelerate the speed of global convergence. The reverse algorithm OCWPA[19] with better population diversity and better accuracy are introduced chaos packing design to set the initial value. For the discrete optimization problem, binary wolves algorithm[20, 21, 22] due to its good ability to develop and mining capability, it can better solve the TSP problems.

5.2 Combination Optimizing

Combination optimizing, together with route planning and parameter settings are the main aspects of application in optimization. The improved wolf pack algorithm has better advantages in solving combination optimizing problems. Reference [13] adopted improved binary wolf pack algorithm to solve classic combination optimizing problems, 0-1 knapsack problems and multi-dimensional knapsack problems, finding that have better performance in exploitation and production. Compared with the binary particle swarm optimization algorithm and the binary genetic algorithm which adopted greedy decisions, it has a better property in optimal solution, the worst solution, average value and a smaller standard deviation. Reference[23] uses the improved algorithm of binary wolves to solve Web service combination problems based on Qos property, the experimental result shows that it has some improvements on efficiency and property compared to the traditional optimization algorithm.

5.3 The Optimization of Engineering Field

The vitality of swarm intelligent algorithm is from the widely optimal application of engineering fields. The wolf pack algorithm has better convergence performance and rapid optimizing ability, which has been used in the controlling of power ratings, sensor arrangement and so on. To solve the redundancy problems of sensor placements, reference[24 , 25] uses wolf pack algorithm which has the function of grade partition to optimize the sensor arrangement' s project, and uses a bridge criterion model as an example to prove that wolf pack algorithm can solve this problem effectively. Reference[26] uses wolf pack algorithm to choose in hyper spectral band, the experiments proved that it has better property than particle swarm algorithm, difference algorithm and differential evolution algorithm .

6 The Development Tendency and Prospect

The results showed that wolf pack intelligent optimization algorithm with merits of converging speed, good global search capability, strong robustness and easy to operate, and its thought began to receive widespread academic attention . But compare to the swarm intelligence algorithm of classic ant colony algorithm, particle swarm optimization, bee colony algorithm, the related researches still in an initial stage, a lot of problems still need a further study, it mainly displays in the following respects:

1. The theory research on wolf pack algorithm. The research of using wolf pack algorithm is rather little, especially the research for its public published references, leading to the demonstrations of wolf pack's convergence are still not perfect. Further study is needed on the research of the algorithm and its improvement, the optimization mechanism and convergence rate of the derivative algorithm.

2. The theory research on swarm intelligence algorithm. In the references which were already published that concerned wolf pack algorithm' s research, lots of references mentioned that making use of random ways to evaluate parameters, updating individual rules, initializing solution settings, which fully proved that they can effectively avoid to fall into local extremum, at the same time can protect the individual diversity better. This is 'oriented random thought' that put forward by WPA author, it can afford improved mechanism of universality for swarm intelligence algorithm, but now the thought still in a budding stage. Its theoretical framework, theoretical basis and theoretical identification need a further study.

3. The research of multiple target wolf pack algorithms. The issue of multiple target optimizations is one of the most difficult problems in optimizing research area, the trait is that it doesn't exist the global optimal solution, but searching for the collection of solution. Wolf pack algorithm has a good property in solving high dimension function and multimodal function, and has better advantages in theory in solving multiple target optimal problems, but there is few related research references about this, this also the challenge when improving the algorithm.

Acknowledgments. This work is financially supported by National Natural Science Foundation of China under Grant No.61502534. Based on my tutor and his students' several previous papers about wolf pack algorithms, this paper is proposed to compare those algorithms and their modified version. And co-authors help me to complete this paper.

References

1. Xing, L.: Research on Evolutionary Learning Intelligent Optimization Algorithm and Its Application. National University of Defense Technology, Changsha (2009)
2. Liang, Y. (ed.): Theory and Application of Swarm Intelligence Optimization Algorithm. In: Abadi, M., Ito, T. (eds.): Theoretical Aspects of Computer Software. Science Press, Beijing (2009)
3. Dorigo, M.: Optimization，learning and natural algorithms. Dipartimento di Elettronica, Politecnico di Milano, Italy (1992) 140
4. Kennedy, J., Eberhart, C.: Particle swarm optimization. Proceedings of IEEE International Conference on Neural Networks, Washington, USA (1995) 1942-1948
5. Karaboga, D.: An idea based on honey bee swarm for numerical optimization. Erciyes University, Engineering Faculty, Computer Engineering Department (2005)
6. Li, X. Shao, Z. Qian, J.: An optimization model based on animal autonomy: Fish swarm algorithm. System engineering theory and Practice. Vol. 22. (2009) 32-38
7. Eusuff, M. Lansey, K.: Optimization of water distribution network design using the shuffled frog leaping algorithm. Journal of Water Resources Planning and Management. Vol. 129. (2003) 201-225
8. Jorge, A., Octlán, D., Felipi, C.: Meta-heuristics algorithms based on the grouping of animals by social behavior for the traveling salesman problem. International Journal of Combinatorial Optimization Problems and Informatics. Vol. 3. (2012) 104-123
9. Mech, L. (ed.): The wolf: the ecology and behavior of an endangered species. Natural History Press, New York (1970)
10. Wu, H., Zhang, F., Wu, L.: New Swarm Intelligence Algorithm – Wolf Pack Algorithm. System Engineering and Electronics. Vol. 35. (2013) 2430-2438
11. Seyedali, M., Seyed, M., Andrew, L.: Grey Wolf Optimizer. Advances in Engineering Software (2014) 46-61
12. Tang, Q., Shen, Y., Hu, C.: Swarm Intelligence：Based Cooperation Optimization of Multi-modal Functions. Cognitive Computation. Vol. 5. (2013) 48-55
13. Wu, H.: Wolf Pack Algorithm and Its Application. Air Force Engineering University, Xi'an (2013)
14. Wolpert, D. Macready, W.: No free lunch theorems for optimization. Evolut Comput. IEEE Trans (1997) 67-82
15. Yang, C. Tu, X., Chen, J.: Algorithm of marriage in honey bees optimization based on the wolf pack. Proceeding of 2007 International Conference on Intelligent Pervasive Computing, Jeju Island, Korea (2007) 462-467
16. Liu, C., Yan, X., Liu, C.: A Mobile Robot Dynamic Path Planning Method Based on the Improved Ant Colony Algorithm. Electronic. Vol. 39. (2011) 1220-1224

17. Yu, M. Li, G., Li, D.: Application of Improved Quantum Genetic Algorithm in UAV Flight Path Planning. Computer Simulation. Vol. 32. (2015) 106-109, 131
18. Wu, H., Zhang, F.: An uncultivated wolf pack algorithm for highdimensional functions and its application in parameters optimization of PID controller. Proceedings of the IEEE Congress on Evolutionary Computation, Beijing, China (2014) 1477-1482
19. Wu, H., Zhang, F.: Opposition-based chaotic wolf pack algorithm. Communications in Nonlinear Science and Numerical Simulation (2015)
20. Wu, H., Zhang, F., Zhan, R.: A Binary Wolf Pack Algorithm for Solving 0-1 Knapsack Problem. System Engineering and Electronics. Vol. 36. (2014) 1660-1667
21. Wu, H., Zhang, F., Zhan, R.: Improved Binary Wolf Pack Algorithm for Solving Multidimensional Knapsack Problem. System Engineering and Electronics. Vol. 37. (2015) 1084-1091
22. Dong, Y., Du, J., Li, B.: Research on Discrete Wolf Pack Algorithm of Multiple Choice Knapsack Problem. Transducer and Microsystem Technologies. Vol. 34. (2015) 21-23, 26
23. He, J.: WCA – based Web Service Composition Optimization Research. Computer Knowledge and Technology. Vol. 11. (2015) 70-73
24. Yi, T., Wang, C., Li, H.: Hierarchic Wolf Algorithm for Optimal Triaxial Sensor Placement. Journal of Building Structure. Vol. 35. (2014) 223-229
25. Yi, T., Wang, C., Li, H.: Optimal Triaxial Sensor Placement Based on Distributed Wolf Algorithm for. Journal of Vibration Engineering. Vol. 27. (2014) 223-229
26. Medjahed, S., Saadi, T., Benyotto, A.: Gray Wolf Optimizer for Hyperspectral Band Selection. Applied Soft Computing. http://dx.doi.org/10.1016/j.asoc.2015.09045.
27. Joanna, K. Fatos, X.: Supporting Situated Computing with Intelligent Multi-agent Systems. International Journal of Space-Based and Situated Computing. DOI: 10.1504/IJSSC. 2011.039105:43-58
28. Tomasz, W., Han, Y., Xiao, Y.: Evaluation of Some Optimisation Techniques for Semantic Query Answering on Shared-nothing Architecture. International Journal of Space-Based and Situated Computing. DOI: 10.1504/IJSSC.2012.045567:23-30
29. Ivan, D., Mychailo, K., Natalia, K.: Enterprise Distributed Service Platform – network Architecture and Topology Optimisation. International Journal of Space-Based and Situated Computing. DOI: 10.1504/IJSSC.2012.045565:31-44
30. Uchida, K., Takematsu, M., Lee, J.: A Particle Swarm Optimisation Algorithm to Generate Inhomogeneous Triangular Cells for Allocating Base Stations in Urban and Suburban Areas. International Journal of Space-Based and Situated Computing. DOI: 10.1504/IJSSC. 2013.058372:215-226
31. Mathiyalagan, P., Suriya, S., Sivanandam, S.: Hybrid Enhanced Ant Colony Algorithm and Enhanced Bee Colony Algorithm for Grid Scheduling. International Journal of Grid and Utility Computing http://dx.doi.org/10.1504/IJGUC.2011.039980:59-75

SFCS-Broker: Supporting Feedback Credibility Brokering Service for Multiple Clouds Environments

Heba Kurdi, Sara Alfaraj, Asma Alatawi
Computer Science Department, King Saud University, SA
heba.kurdi@gmail.com
saranaser128@gmail.com
asmasuliaman@outlook.com

Abstract. Due to the widespread of multiple clouds computing, the demand has increased for trust and credibility systems. In this paper, we present a SFCS-Broker (Supporting Feedback Credibility Brokering Service for Multiple Clouds Environments) to make an efficient matching between users and providers by monitoring the real-time behavior, collecting users' feedback and ensuring credibility of the provided feedback. The proposed SFCS-Broker scheme allows balancing between delay and ratio of the calculated credibility. Additionally, the experimental results show that SFCS-Broker successfully detects various models of malicious users.

Keywords: cloud computing; multiple clouds; cloud broker; Trust management.

1 Introduction

Multiple clouds environments represent a hot direction in cloud computing industry. However, open and anonymous nature of these environments make trust management a significant challenge; it leads to lack of credibility [1, 2] and leaves many open doors for malicious users [3]. According to [4], trust is one of the most critical obstacles to the growth of multiple cloud computing. A lack of trust between cloud users and providers has hindered the universal acceptance of clouds. Thus, the development of trust awareness schemes for multiple cloud environments has become a key and urgent research direction [5, 6].

Recently, as a promising concept to offer enhanced service of cloud environment, cloud-brokering systems have been emerged. Brokers can provide intermediation and aggregation capabilities to enable users to deploy their virtual infrastructures across cloud systems, and to negotiate about allocating resources among multiple sites [7]. To increase the adoption of cloud services, cloud providers must first establish trust to limit the worries of a large number of users [8]. In order to manage and schedule resources

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_54

with high trustworthy, we need an accurate way of measuring and predicting usage patterns of computing resources whose patterns are changing dynamically overtime. As observed in previous study of brokers, such RESERVIOR[9], PCMONS[10] and OPTIMIS[11] (discussed in section 2) we inspired that these brokers do not provide trust management capabilities for multiple cloud collaborative computing, such as how to select the optimal cloud resources to deploy a service or how to optimally distribute the different components of a service among different clouds. In addition, as observed in [7], which introduced approach (called T-Broker) for efficient matching computing resources that involved trust management capabilities, to satisfy various user requests in the multi-cloud with lightweight feedback to meet service level agreements (SLA).

In fact, SLA are not sufficient to introduced trust among peers in cloud environments because of its unclear and inconsistent clauses [12]. Even more, several researchers have recognized the significance of trust management and proposed solutions to assess and manage trust based on feedbacks collected from participants. In spite of users' lightweight feedback was a good source to assess the overall trustworthiness of cloud services with fast convergence and slight delay, it may affected by malicious users (e.g. collective, isolated and naïve attackers users) and give misleading results which may preclude inaccurate evaluation of credibility [2].

This work present an overview design and implementation of approach called SCFS-Broker (Supporting Feedback Credibility Service in brokering scheme for multiple cloud). Concisely, the salient features of SCFS-Broker are:

-Implementing trust-aware brokering architecture for management and resources scheduling.

-Monitoring the real-time behavior of resource providers (direct trust) and feedback from users (indirect trust).

-Using hybrid and adaptive trust model to compute the overall trust degree of service resources, and uses a maximizing deviation method to compute the direct trust of service resource.

-Collecting users' feedback and aggregating them to evaluate providers based on their reputations. SFCS-Broker ensures feedback credibility by using three different levels of trustworthiness in order to make balance between delay and credibility based on users' needs.

The experimental results show that SFCS-Broker allows balancing between delay and ratio of the calculated credibility. Additionally, the experimental results show that SFCS-Broker successfully detects various models of malicious users.

The remaining of this paper is organized as follow; section 2 gives an overview of related work, section 3 discusses the SFCS-Broker, the experimental results are shown in section 4. Finally, section 5 concludes the work and suggests future work.

2 Related Work

A cloud broker is an entity that manages use, performance and delivery of cloud services and intermediates the relationships between cloud providers and cloud consumers. In recent years, there are many cloud service brokers and monitoring systems emerged as a promising concept to offer enhanced services over multiple cloud environments.

In [7], a Trust-aware service Brokering Scheme (T-Broker) was proposed to construct a trust-aware service brokering system for efficient matching of computing resources to satisfy various user requests. T-Broker used hybrid and adaptive trust model to compute the overall trust degree of service resources. Furthermore, T-Broker uses a maximizing deviation method to compute the direct trust (first hand trust) of service resources, which can overcome the limitations of traditional trust models in which the trusted attributes weighted manually or subjectively.

An Optimized Infrastructure Services (OPTIMIS) scheme was introduced in [11], with a vision for hybrid clouds to become a commonplace, realized by private clouds interacting with a rich ecosystem of public and other cloud providers. OPTIMIS enables clouds to be composed from multiple services and resources. It support service brokerage via interoperability, and is architecture-independent.

Considering the lack of generic and open-source solutions for management and monitoring of private clouds, PCMONS [13] was developed. It is intended to be an extensible and modular monitoring system for private clouds .In its first release, PCMONS acted principally retrieving, gathering and preparing relevant information for monitoring data visualization and is specially focused on virtual machines. It was compatible with Eucalyptus (IaaS platform) and Nagios (used specially to visualize monitoring data). However, its development considers easily integration with other toolkits and cloud frameworks, such as OpenNebula, through the development of extensions [13]. PCMONS can be considered as a private cloud that uses open source solutions and integrating with traditional tools like Nagios [10].

According to [14], cloud service consumers and providers interact via trusted relations between them (denoted as Tr). The values of Tr can be either 0 (the trusted relationship does not exist) or 1 (the trusted relationship exists). An unrecognized relation occurs when a cloud service consumer x approaches a cloud service provider y for the first time. Several approaches can be used to compute the trust. For instance, the monitoring and auditing approach calculates the Service Level Agreement (SLA) violations in cloud services. The entities credibility approach specifies a set of parameters to measure the credibility of parties while the feedback credibility approach considers a set of factors to measure the credibility of feedbacks.

As stated in [12, 14-16], reputation as a trust management technique (Rep T) is important because the feedback of the various cloud service consumers can dramatically influence the reputation of a particular cloud service either positively or negatively. Reputation can have direct or indirect influence on the trustworthiness of a particular entity (e.g., cloud service). In RepT, cloud service consumers do not know the source of the trust feedback (i.e., there is no trusted relations). For instance, as mentioned in [14], if the cloud service consumer x has a certain minimum trust threshold Tx and the cloud service provider y has a set of trusted relations

T r(y) = [r1,r2… ri], (i.e., other cloud service consumers), which give trust feedbacks on the cloud service provider T f(y) = [f1, f2, … fn]. These feedbacks used to calculate the reputation of y, denoted as Rep(y), as shown in Equation 1. The cloud service consumer x determines whether to proceed with the transaction based on the reputation result of y, as shown in Equation 2. The more positive feedbacks that y receives is the more likely x will trust the cloud service provider y.

$$\text{Rep (y)}= \frac{\sum_{x=1}^{|T\,f(y)|} T\,f(x;y)}{|T\,f(y)|} \tag{1}$$

$$T\,r(x,y) = \begin{cases} 1\ if\ Rep(y) \geq Tx \\ 0\ otherwise. \end{cases} \tag{2}$$

This model focuses on aggregating the reputation of a particular cloud service based on feedback using QoS and other attributes (e.g., elasticity, geographical location).

3 System Architecture

As mentioned in section 1, most cloud brokers do not provide trust management services. However, [7] introduced a system with trust management capabilities based on time attenuation function and lightweight feedback mechanism to avoid delay among huge number of machines. The main drawback of this approach is that trust may cause misleading feedback (by malicious) and increase the number of transaction that can harm the network. In this work, we present SFCS-Broker to collect users' feedback and aggregate them to evaluate providers based on their reputations and ensure feedback credibility.

The system architecture of SFCS-Broker is shown in Fig.1. Each cloud provider offers several virtual machine (VMs) configurations of hardware such as CPU, memory space and hard disk capacity. In a similar approach to [7], we used OpenNepula as an infrastructure to collect and index all resources information from providers. Infrastructure layer responsible for managing the deployment of each VM in multi-cloud environment, and make resource matching with expected level of service between users and providers by defining a service level agreement (SLA) contract.

SFCS-Broker trust computation layer, monitors the real-time service behavior (Direct Trust) to guarantee the service level agreement (SLA) requested by users, this is the same approach followed by [7]. Moreover, SFCS-Broker trust computation layer compute the (indirect trust) by collecting users' feedback and calculate credibility ratio then, stores the feedback in a database to make a prior notion to the next transactions. As shown in Fig2, the process starts by a request from an end user and ends with the user giving his feedback. Before starting a transaction, the user requests the reputation of the cloud service provider. Similar to [7][2], while calculating the reputation SFCS-Broker considers values larger than 0.8 as high, below this but larger than 0.4 as

medium reputation, and below 0.4 but larger than 0.01 means low reputation. After determining which level is needed and requesting the service, the user leaves his feedback regarding the service.

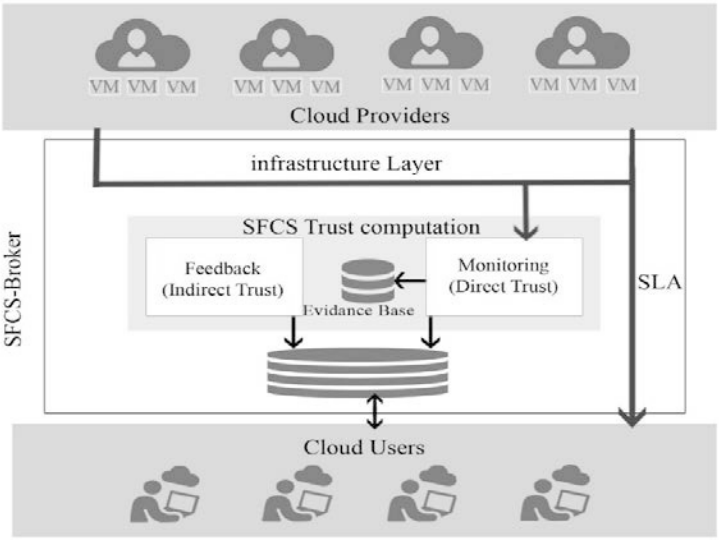


Fig 1. SFCS-Broker Architecture

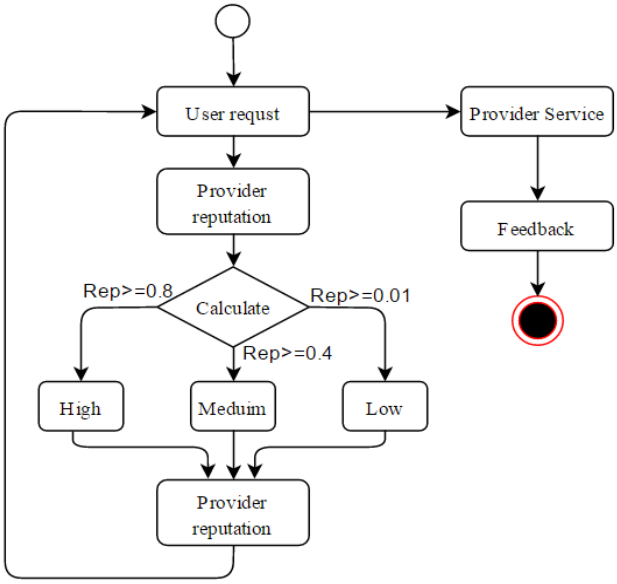


Fig 2. Improved credibility model of SFCS-Broker

4 Evaluation Framework

The hardware and software specifications of the computer system used to implement and run SFCS-Broker are shown in Table1.

Table 1. Material list used to implement SFCS-Broker.

Hardware	
Operating System	windows 8 x64
RAM	4GB
Programing Language	QTM : P2p source code (in JAVA)
Software	
infrastructure	NetBeans IDE 8.1
Programming Language	QTM: P2p source code (in JAVA)

In a similar approach to [1, 17], experimental sittings consisted of 10000 transactions, 5000 files, 25 users, and variant malicious behavior with different percentage of presence in the cloud environment (from 10% to 70%). Furthermore, we tested our approach with three type of malicious strategies: (1) Collective: group of malicious users who provide good resource but give positive feedback for bad resources to increase their reputation, (2) Isolated: malicious users who provide positive feedback to bad users and (3) Naïve: malicious users are simple and clear; they provide invalid resources [17]. The evaluation focused on (A) Accuracy, to check whether the proposed algorithm can accurately provide trust measurement. (B) Credibility which represent level of feedback credibility (high, medium, low or none). Table2 lists the key symbols with their brief descriptions.

Table 2. Symbols and their meanings.

Symbols	**Meaning**
$Rep(y)$	Reputation of provider y
$\overline{Rep(y)}$	Mean for FB reputation
n	Number of FB reputation
p	Number of positive feedback
N	Number of negative feedback

We considered two performance indicators:

- *Accuracy*: calculated as the Mean Absolute Deviation (MAD) to reflect the system accuracy:

$$\text{MAD} = \frac{\sum Rep(y) - \overline{Rep(y)}}{n}. \tag{3}$$

If MAD value close to zero then it has higher accuracy. Similarly, we applied the following equation to the Light Wight Trust Algorithm [17], which we used as a benchmark:

$$\text{Lightwight} = \frac{p+1}{P+N+2}. \quad (4)$$

If the result is larger than 0.5, this means positive ratings, if the result smaller than 0.5 it means negative ratings and if the result equal 0.5 it means new provider in the cloud market [7].

- *Credibility*: credibility can be high, medium, low or no trust based on:
  - If $Rep(y) \geq 0.8$ ∴ high credibility.
  - If $Rep(y) \geq 0.4$ ∴ medium credibility.
  - If $Rep(y) \geq 0.01$ ∴ low credibility.
  - If $Rep(y) = 0.0$ ∴ no trust (new cloud provider).

5 Result

5.1 Accuracy

The system accuracy (MAD) when malicious collectives, isolated and Naïve users are introduced to the system are shown in Fig 3, Fig 4 and Fig 5 respectively. We started with a small percentage of 10% malicious users and increased it gradually up to 70%. It is clear from all the figure that SFCS-Broker has considerably better performance when compared to the benchmark, the Light weight trust algorithm [7].

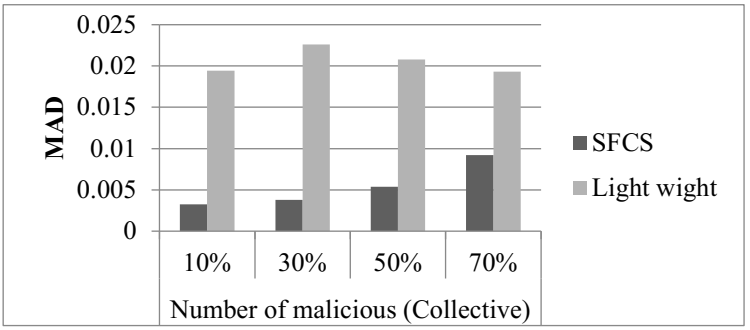


Fig 3. MAD result with malicious (Collective)

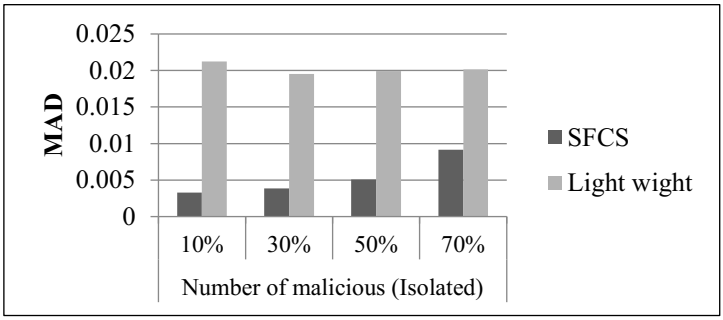


Fig 4. MAD result with malicious (Isolated)

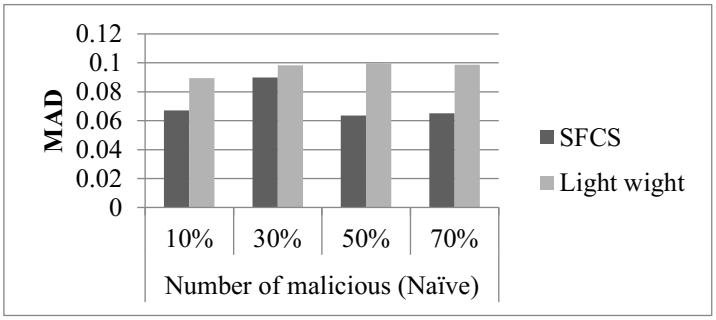


Fig 5. MAD results with malicious (Naïve)

5.2 Credibility

The credibility results when malicious collectives, isolated and Naïve users are introduced to the system are shown in Fig 6, Fig 7 and Fig 8 respectively. We started with a small percentage of 10% malicious users and increased it gradually up to 70%. It is clear that SFCS-Broker was able to identify the credibility of the feedback with high accuracy at different levels.

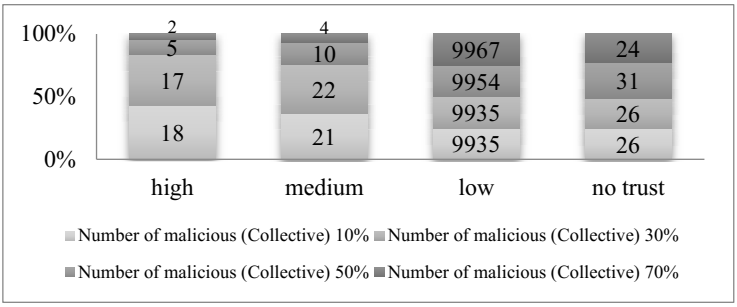


Fig 6. Credibility with malicious (Collectives)

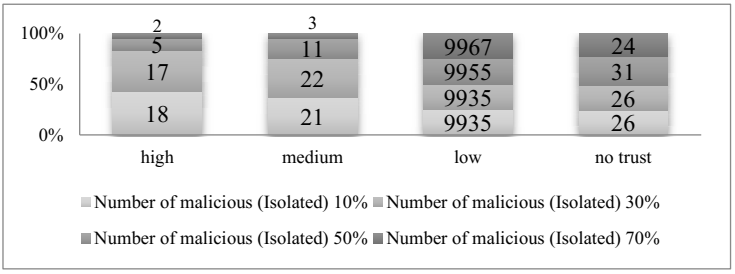


Fig 7. Credibility with malicious (Isolated)

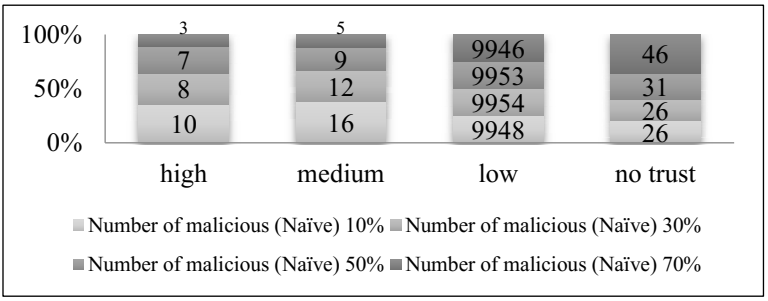


Fig 8. Credibility with malicious (Naïve)

6 Conclusion

Trust management is a demanding issue in multiple clouds environments. In this paper we introduced a Supporting Feedback Credibility Brokering Service for multiple clouds environments (SFCS-Broker). The proposed scheme showed promising results in all experimental scenarios. In the future, we plan to combine different trust management techniques, such as reputation and recommendation, to our scheme to increase the trust accuracy.

References

[1] S. D. Kamvar, M. T. Schlosser, and H. Garcia-Molina, "The eigentrust algorithm for reputation management in p2p networks," in Proceedings of the 12th international conference on World Wide Web, 2003, pp. 640-651.
[2] T. Noor, Q. Sheng, L. Yao, S. Dustdar, and A. Ngu, "CloudArmor: Supporting reputation-based trust management for cloud services," 2015.
[3] A. Alexa and A. Theobald, "Reputation management in P2P networks: the eigentrust algorithm," ed: Thesis, Supervised by Anja Theobald.
[4] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, et al., "A view of cloud computing," Communications of the ACM, vol. 53, pp. 50-58, 2010.
[5] F. Paraiso, N. Haderer, P. Merle, R. Rouvoy, and L. Seinturier, "A federated multi-cloud PaaS infrastructure," in Cloud Computing (CLOUD), 2012 IEEE 5th International Conference on, 2012, pp. 392-399.
[6] P. Jain, D. Rane, and S. Patidar, "A novel cloud bursting brokerage and aggregation (CBBA) algorithm for multi cloud environment," in Advanced Computing & Communication Technologies (ACCT), 2012 Second International Conference on, 2012, pp. 383-387.
[7] X. Li, H. Ma, F. Zhou, and W. Yao, "T-Broker: A Trust-Aware Service Brokering Scheme for Multiple Cloud Collaborative Services," Information Forensics and Security, IEEE Transactions on, vol. 10, pp. 1402-1415, 2015.
[8] K. Hwang, S. Kulkareni, and Y. Hu, "Cloud security with virtualized defense and reputation-based trust mangement," in Dependable, Autonomic and Secure Computing, 2009. DASC'09. Eighth IEEE International Conference on, 2009, pp. 717-722.
[9] B. Rochwerger, D. Breitgand, E. Levy, A. Galis, K. Nagin, I. M. Llorente, et al., "The reservoir model and architecture for open federated cloud computing," IBM Journal of Research and Development, vol. 53, pp. 4: 1-4: 11, 2009.
[10] D. Chaves, S. Aparecida, R. B. Uriarte, and C. B. Westphall, "Toward an architecture for monitoring private clouds," Communications Magazine, IEEE, vol. 49, pp. 130-137, 2011.
[11] (2013). OPTIMIS. Available: optimis-project.eu/project
[12] S. M. Habib, S. Ries, and M. Mühlhäuser, "Towards a trust management system for cloud computing," in Trust, Security and Privacy in Computing and Communications (TrustCom), 2011 IEEE 10th International Conference on, 2011, pp. 933-939.
[13] PCMONS. Available: findbestopensource.com/product/pcmons
[14] T. H. Noor and Q. Z. Sheng, "Trust as a service: a framework for trust management in cloud environments," in Web Information System Engineering–WISE 2011, ed: Springer, 2011, pp. 314-321.
[15] F. J. Krautheim, D. S. Phatak, and A. T. Sherman, "Introducing the trusted virtual environment module: a new mechanism for rooting trust in cloud computing," in Trust and Trustworthy Computing, ed: Springer, 2010, pp. 211-227.
[16] P. D. Manuel, S. T. Selvi, and M.-E. Barr, "Trust management system for grid and cloud resources," in Advanced Computing, 2009. ICAC 2009. First International Conference on, 2009, pp. 176-181.
[17] A. G. West, S. Kannan, I. Lee, and O. Sokolsky, "An evaluation framework for reputation management systems," Trust Modeling and Management in Digital Environments: From Social Concept to System Development, pp. 282-308, 2010.

Part III
Workshop SMDMS-2016:The 7th International Workshop on Streaming Media Delivery and Management Systems

QoE Assessment of Multi-View Video and Audio Transmission with a Packet Discarding Method over Bandwidth Guaranteed IP Networks

Keita Furukawa and Toshiro Nunome

Abstract This paper proposes a packet discarding method for multi-view video and audio (MVV-A) transmission over bandwidth guaranteed IP networks. In general, a network equipment in the IP networks decides to drop a packet on packet-by-packet basis. A video frame is transmitted by several packets. The proposed method discards several packets according to average video frame size when the queue length becomes long. We investigate the effect of the discarding method on QoE by a subjective experiment. The assessors are asked to watch video by selecting a viewpoint from eight viewpoints according to instructions. As a result, we confirm that the packet discarding method can enhance QoE.

1 Introduction

Multimedia communication services have been popularized owing to high-speed and broadband IP networks. Such network services require high *QoS (Quality of Service)*. However, best-effort IP networks cannot guarantee QoS and then cause packet losses and delays. On the other hand, bandwidth guaranteed IP networks can provide QoS for flows on the networks. The guaranteed networks are suitable for delay critical services.

QoS can be defined at each layer of the IP-based networks. In addition, QoS of each layer (namely, level) is closely related to the adjacent lower-level QoS [1]. *QoE (Quality of Experience)* [2], which is the most important quality for network users,

Keita Furukawa
Department of Computer Science and Engineering, Graduate School of Engineering, Nagoya Institute of Technology, Nagoya 466–8555, Japan, e-mail: furukawa@inl.nitech.ac.jp

Toshiro Nunome
Department of Computer Science and Engineering, Graduate School of Engineering, Nagoya Institute of Technology, Nagoya 466–8555, Japan, e-mail: nunome@nitech.ac.jp

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_55

is largely affected by the application-level QoS. We can enhance QoE by means of enhancement of application-level QoS.

As multimedia services become popular in the networks, there is high expectation for the users to feel sense of really being there; we call the sense as "presence". For giving higher presence to the users, *MVV (Multi-View Video)*, in which the users can change the viewpoint, has been studied [3].

Reference [4] assesses QoE of *MVV-A (Multi-View Video and Audio)*, which is MVV accompanied by audio. The paper evaluates the effect of video encoding bitrate and playout buffering time at the receiver terminal in a experimental system with eight cameras and a microphone. It discusses that the viewpoint change response affects QoE largely in MVV-A. The response can improve as the playout buffering time decreases, while short buffering time causes degradation of output quality. Thus, there is a tradeoff relationship between the viewpoint change response and the output quality. The study is carried out for the best-effort IP network.

The bandwidth guaranteed network can ensure network resource. However, it takes costs for ensuring the resource. Thus, rich multimedia services such as MVV-A cannot always achieve enough bandwidth. In the case, exceeded packets enqueues on network routers. If the number of queueing packets reaches a limit of queueing length, the packets will drop on packet-by-packet basis.

Since a video frame is transmitted by several packets, it is efficient that the router discards several packets from a video frame once. FIPD (Frame-Induced Packet Discarding) [5] employs this kind of mechanism. However, it is difficult for the router to recognize which packet belongs to which video frame. In addition, Reference [5] does not evaluate QoE. The ultimate goal of network services is provision of high QoE. Thus, QoE-level discussion is very important.

In this study, we propose a packet discarding method based on WRED (Weighted Random Early Detection) [6]. For each class which handles video data, the router discards packets for the class when the number of queueing packets exceeds the average number of packets for a video frame + 1. With the mechanism, the router acts like FIPD. In addition, the proposed method does not add any overhead in the data transmission phase; it just tell the average number of packets for a video frame in the session establishment phase. We assess QoE of MVV-A transmission with the packet discarding method over a bandwidth guaranteed IP network.

The rest of the paper is structured as follows. Section 2 defines the packet discarding method in this paper. Sections 3 and 4 outline methods of the experiment and assessment methods, respectively. We present results of the experiment in Section 5, and Section 6 concludes this paper.

2 Packet discarding method

In this paper, we compare QoE of MVV-A transmission with the packet discarding method and that without the method. This section explains overview of the packet

discarding method. The method exploits the average number of packets for a video frame and WRED.

2.1 WRED

WRED enhances RED (Random Early Detection) [7], which is a congestion avoidance algorithm, to manage multiple traffic classes on the network. It applies RED for each traffic class. RED randomly discards packets according to average queue length of the observed queue. It avoids congestion by discarding packets intentionally before the queue length reaches the maximum queue length.

Average queue length

As Eq. (1), the average queue length q_{avg} is calculated with Exponentially Weighted Moving Average (EWMA).

$$q_{avg} = w_q q + (1 - w_q) q_{avg} \tag{1}$$

Here, w_q is an weight factor, and q is current queue length.

Packet discarding probability

In the router which we employ, there are two thresholds T_{min} and T_{max} for the average queue length. When the average queue length exceeds T_{min}, the router discards packets randomly with probability P. If the average queue length exceeds T_{max}, the router performs tail-drop.

2.2 Calculation of average number of packets for a video frame

When we transmit video over IP networks, a video frame is divided into several IP packets according to MTU (Maximum Transfer Unit). We refer to the transmission unit at the application-level as an *MU (Media Unit)*. A video MU is a video frame. The average number of packets for an MU can be calculated from the average encoding bit rate of video, the video MU rate, and the MTU size. For example, when the average encoding bitrate, the MU rate, and the MTU size are 3000 kbps, 30 MU/s, 1500 bytes, respectively, the average number of packets is nine.

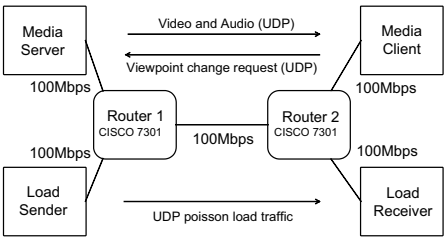


Fig. 1 Network configuration

2.3 Discarding method

In this paper, we employ WRED for each class in CBWFQ (Class-Based Weighted Fair Queueing). We set the minimum threshold T_{min} to the average number of packets for a video frame + 1 and the discarding probability P to 1. By discarding several packets once, the router can recover from a congested state quickly.

3 Experimental methods

Figure 1 shows the configuration of the experimental system. The system consists of Media Server, Media Client, Load Sender, Load Receiver, and two routers. All the links are 100 Mbps Ethernet. Eight video cameras are connected to Media Server, which is equipped with real-time H.264 encoding boards. Media Server captures the video of each camera. The audio is captured by a microphone.

Router 1 performs bandwidth control by LLQ (Low Latency Queueing); it employs PQ (Priority Queueing) for audio and applies CBWFQ for video and load traffic. Router 1 also carries out the packet discarding control explained in Section 2.

The specifications of audio and video are shown in Table 1. An audio MU consists of 320 audio samples. Each MU is transmitted as a UDP packet. We employ frame skipping as the output method of video. That is, when some packets consisting of an MU is lost, output of the MU is skipped.

We employ three methods differentiated by the number of simultaneously transmitted streams [4]. We consider three streams, five streams, and eight streams. The average bit rate for each video stream with the three streams method is 4.0 Mbps, the five streams method is 3.0 Mbps, 2.5 Mbps, 2.3 Mbps, 2.15 Mbps and 2.0 Mbps, and the eight streams method is 1.5 Mbps.

The three streams method selects the current viewpoint and the nearest viewpoints from the current one for transmission.

In the five streams method, the server selects several viewpoints; one is the current viewpoint, and the others are not selected recently and near the current viewpoint. It assigns the highest bitrate for the current viewpoint. As far from the current

Table 1 Specifications of video and audio

	video	audio
coding scheme	H.264 (704×480 pixels)	G.711 μ-law
GOP	I	-
average MU rate	30 [MU/s]	25 [MU/s]

Table 2 Bandwidth assignment

	stream	bitrate [kbps]	bandwidth (insufficient) [kbps]	bandwidth (minimum) [kbps]
video	3	4000	4000	4122
	5	3000	3000	3095
		2500	2500	2585
		2300	2300	2375
		2150	2150	2225
		2000	2000	2065
	8	1500	1500	1555
audio	-	64	90	

viewpoint, the server reduces encoding bitrate. The viewpoint which is more likely to be selected is transmitted with higher bitrate.

Since we use eight cameras for our research, all the streams will be transmitted in the eight streams method. So, we can see the response of the viewpoint change being faster in the eight streams method. At the same time, the three streams method can suffer more delay of responding the viewpoint change than the five streams method owing to less number of streams being transmitted.

When we send application data through the IP networks, header information is added in each layer of the Internet protocol stack. The bandwidth control is performed on the data-link layer. Thus, when we assign bandwidth for the average media encoding bitrate and the header size of each layer, the minimum bandwidth will be guaranteed; however, starvation can occur because of jitter of bitrate. On the other hand, we assign only the average media encoding bitrate for the bandwidth, it is obviously insufficient. In this study, we consider the above two patterns: *minimum* and *insufficient*. We show the assigned bandwidth in Table 2.

For load traffic, we assign the residual bandwidth (100 Mbps − audio and video bandwidth). Load Sender transmits load traffic (1480 bytes UDP packets generated with the exponential distributed intervals) as the average bitrate equals assigned bandwidth.

Media Receiver uses playout buffering control for absorbing delay jitter. In order to investigate the effect of playout buffering time, we employ two values of the playout buffering time: 100 ms and 500 ms.

Figure 2 shows the camera arrangement in this paper. In the assessment, we ask the users to watch moving toy insects. The number of insects are four, and each insect has different color (red, blue, green, and black). The eight cameras surround the course in which the insects move. The microphone is placed on the center of the course. The user follows one of the four insects by changing viewpoint with GUI in Fig. 3. Figure 4 depicts the screen capture in the experimental run.

In this paper, we employ 24 stimuli from the combinations of the two buffering time, three number of transmission streams, two bandwidth assignment types, and with or without the packet discarding method. The assessors are 20 male students in their twenties.

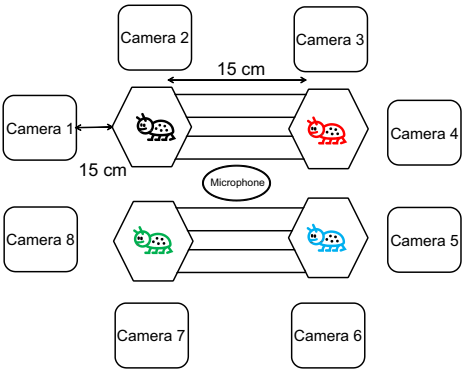


Fig. 2 Camera arrangement

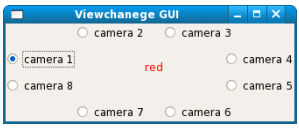


Fig. 3 Viewpoint change GUI

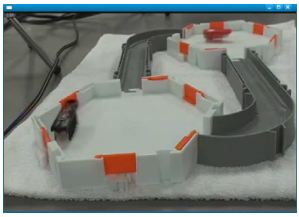


Fig. 4 Screen shot

4 Assessment methods

4.1 Application-level QoS assessment

In the experiment, we employ the following parameters for the application-level QoS assessment.

The *MU loss ratio* is the ratio of the number of MUs not output at the recipient to the number of MUs transmitted by the sender for the displayed video streams.

The *average viewpoint change delay* is defined as the time in seconds from the moment the user inputs a request for viewpoint change by the GUI until the instant a new viewpoint is output at the client.

4.2 QoE assessment

In the experiment, we perform multidimensional QoE assessment with seven adjective pairs of polar terms. The pairs of polar terms in the subjective experiment are shown in Table 3. The pairs can be classified into five categories; there are three pairs for video, a pair for audio, a pair for inter-stream synchronization, a pair for response, and a pair for overall satisfaction.

Note that the experiment was performed with the Japanese language. This paper has translated the used Japanese terms into English. Therefore, the meanings of adjectives or verbs written in English here may slightly differ from those of Japanese ones.

For each criterion, a subjective score is measured by the *rating scale method*. In the method, an assessor classifies the stimuli into a certain number of categories; here, each criterion is evaluated to be one of five grades. The best grade (score 5)

Table 3 Adjective pairs

category	pair of polar terms
video	video is rough - smooth
	video is blurred - sharp
	video is difficult to grasp - easy to grasp
audio	audio is artificial - natural
synchronization	audio and video are out of synchronization - in synchronization
response	viewpoint change is slow - fast
overall satisfaction	bad - excellent

represents the positive adjective (the right-side one of each pair), while the worst grade (score 1) means the negative adjective (the left-side one).

The numbers assigned to the categories only have a greater-than-less-than relation between them; that is, the assigned number is nothing but an ordinal scale. When we assess the subjectivity quantitatively, it is desirable to use at least an interval scale. In order to obtain an interval scale from the result of the rating scale method, we first measure the frequency of each category with which the stimulus is placed in the category. With the law of categorical judgment [8], we can translate the frequency obtained by the rating scale method into an interval scale. Since the law of categorical judgment is a suite of assumptions, we must test goodness of fit between the obtained interval scale and the measurement result. This paper uses Mosteller's method [9] to test the goodness of fit. Once the goodness of fit has been confirmed, we refer to the interval scale as the psychological scale; it is a QoE metric.

5 Experimental results

5.1 Application-level QoS

Figure 5 depicts the MU loss ratio of video. Figure 6 shows the average viewpoint change delay. The abscissa of these figures show the combinations of the packet discarding method and the number of transmitted video streams. Each figure plots the results for four combinations of the playout buffering time (100 ms, 500 ms) and bandwidth assignment (insufficient, minimum). The figures also show the 95 % confidence interval.

We find in Fig. 5 that the packet discarding method can decrease the video MU loss ratio when the playout buffering time is 100 ms. On the other hand, the loss ratio for the playout buffering time 500 ms with the packet discarding method slightly larger than that without the method. This is because the method drops the packets which can be arrived at the receiver by the output scheduled time in the large playout buffering time.

We have also assessed the MU loss ratio of audio although we do not show the result because of space limitations. We then notice that the audio MUs scarcely loss.

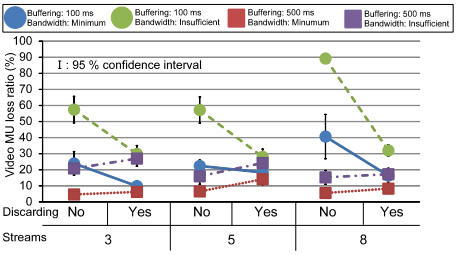


Fig. 5 Video MU loss ratio

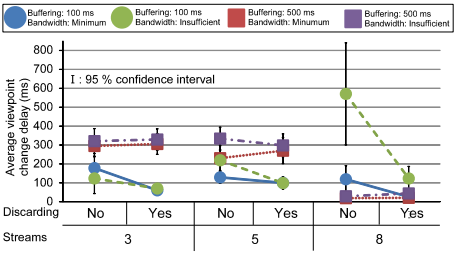


Fig. 6 Viewpoint change delay

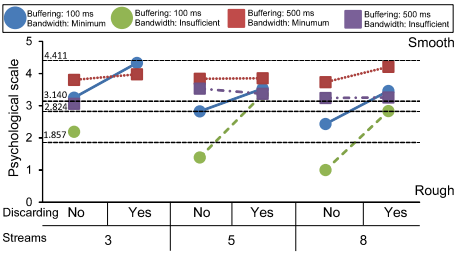


Fig. 7 Video is rough - smooth

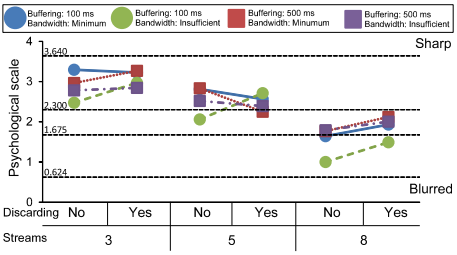


Fig. 8 Video is blurred - sharp

This is because we employ PQ for audio; it transmits audio traffic with the highest priority.

We see in Fig. 6 that the packet discarding method can decrease the viewpoint change delay in most of the cases. For the playout buffering time 100 ms in the eight streams transmission with the minimum bandwidth assignment, the packet discarding method has larger effect. This is because the method can decrease the MU loss of backup viewpoint (not a main viewpoint), and then the users feel quick viewpoint change response.

5.2 QoE assessment

We calculated the interval scale for each criterion. We then carried out the Mosteller's test. As a result, we have found that the hypothesis that the observed value equals the calculated one can be rejected with significance level of 0.05 in some criteria. Therefore, we removed the stimuli which have large errors until the hypothesis cannot be rejected. In this paper, we use obtained values by these processes as the psychological scale.

Since we can select an arbitrary origin in an interval scale, for each criterion, we set the minimum value of the psychological scale to unity.

Figures 7 and 8 show the assessment results of video. Figure 9 depicts the assessment result of audio. Figures 10, 11, and 12 are the results of synchronization,

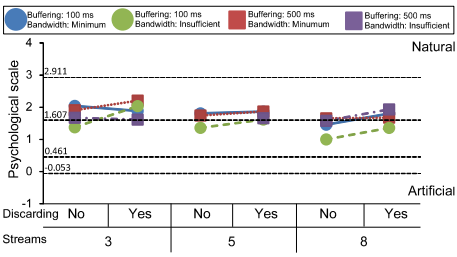


Fig. 9 Audio is artificial - natural

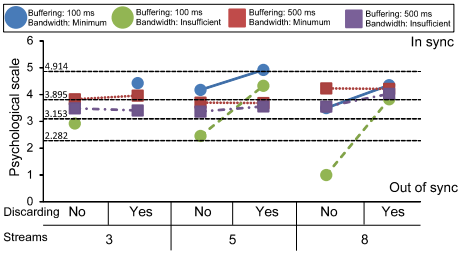


Fig. 10 Audio and video are out of synchronization - in synchronization

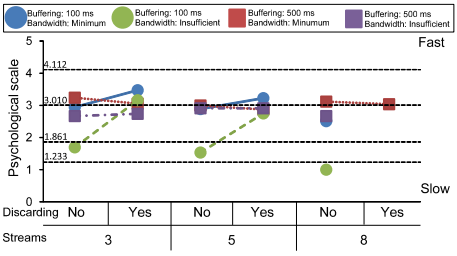


Fig. 11 Viewpoint change is slow - fast

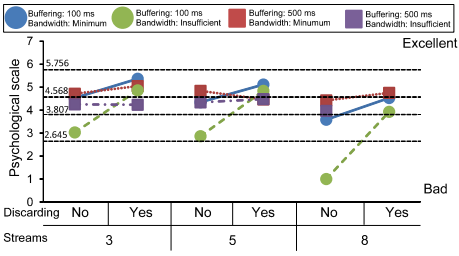


Fig. 12

response, and overall satisfaction, respectively. The abscissa means the packet discarding method for the number of streams. The black dotted lines parallel to the abscissa mean category boundaries.

We see in Fig. 7 that for the playout buffering time 100 ms, the packet discarding method enhances video temporal quality (video is rough - smooth). This is because the discarding method can decrease the lost MUs under the short playout buffering condition. On the other hand, when the playout buffering time is large, the discarding method is not so effective for enhancement of the quality.

In Fig. 8, we observe that the psychological scale values for video spatial quality (video is blurred - sharp) increase as the number of streams decreases. This is because each stream can use higher encoding bitrate as the number of transmitted streams decreases.

We do not show the assessment result for video overall quality (video is difficult to grasp - easy to grasp) because it is closely related to the results in Figs. 7 and 8.

In Fig. 9, we find that the psychological scale values for audio (audio is artificial - natural) decrease as the video quality degrades in Fig. 8. Although the audio MU loss ratio is almost 0 for all the experimental conditions here, the audio quality is affected by the video quality, i.e., the cross-modal effect between audio and video exists.

As for the media synchronization quality (audio and video are out of synchronization - in synchronization) in Fig. 10, the psychological scale values decrease according to the video quality degradation.

In Fig. 11, we see that when the playout buffering time is 100 ms and the bandwidth is insufficient, the psychological scale values without packet discarding method are lower than that with the method. This is because the MU loss ratio is high, and then the delay until the arrival of new viewpoint becomes large.

We notice in Fig. 12 that the overall satisfaction (bad - excellent) with the packet discarding method is equal to or higher than that without the method. In this study, the overall satisfaction is the highest when the number of streams is 3, the playout buffering time is 100 ms, and the minimum bandwidth is guaranteed with the packet discarding method.

6 Conclusions

In this paper, we investigated the effect of the packet discarding method on QoE and application-level QoS of MVV-A transmission over bandwidth guaranteed IP networks. From the assessment results, we found that for the short playout buffering time, the packet discarding method can enhance QoE. This is because the method can discards packets which cannot arrive at the destination by their scheduled output time. However, the method increases MU losses when we can use large playout buffering time. In addition, as the number of transmitted streams decreases, the overall satisfaction of the users enhances.

This paper just evaluates the effectiveness the proposed method with the very simple situation. In future work, we need to assess the effectiveness under various conditions.

References

1. S. Tasaka and T. Nunome, "QoE enhancement of multimedia communications by user-assistance," *Proc. World Telecommunications Congress 2014*, SA.5.3, June 2014.
2. ITU-T Rec. P.10/G.100 Amendment 2, "Amendment 2: New definitions for inclusion in Recommendation ITU-T P.10/G.100, " *International Telecommunication Union*, July 2008.
3. I. Ahmad, "Multiview video: Get ready for next-generation television," *Proc. IEEE Distributed Systems Online*, vol. 8, no. 3, art. no. 0703-o3006, Mar. 2007.
4. M. Yamamoto, T. Nunome and S. Tasaka, "QoE assessment of simultaneous transmission methods of multi-view video and audio IP transmission," *Tech. Rep. IEICE*, CQ2012-15, Apr. 2012 (in Japanese).
5. S. Ramanathan, P. Rangan and H. Vin, "Frame-Induced Packet Discarding: An Efficient Strategy for Video Networking," *4th International Workshop on Network and Operating System Support for Digital Audio and Video*, 1993.
6. S. Vegesna, "IP Quality of Service, Cisco Press," 2001.
7. S. Floyd and V. Jacobson, "Random early detection gateways for congestion avoidance," *IEEE/ACM Trans. Networking*, vol. 1, no. 4, pp. 397-413, Aug. 1993.
8. J. P. Guilford, *Psychometric methods*, McGraw-Hill, N. Y., 1954.
9. F. Mosteller, "Remarks on the method of paired comparisons: III a test of significance for paired comparisons when equal standard deviations and equal correlations are assumed," *Psychometrika*, vol. 16, no. 2, pp. 207-218, June 1951.

Evaluation of Division-based Broadcasting System over Wireless LAN

Yusuke Gotoh and Yuki Takagi

Graduate School of Natural Science and Technology, Okayama University, JAPAN

Summary. Due to the recent popularization of the Internet, continuous media data broadcasting such as audio or video have attracted great attention. Although continuous media data broadcasting can reduce the more necessary bandwidth than conventional on-demand delivery, waiting time from requesting data to playing it increases. In division-based broadcasting, many researches have proposed scheduling methods to reduce the waiting time. Although a division-based broadcasting system has also been proposed to evaluate these methods, this system does not consider wireless LAN environment. In wireless LAN environment, we have to consider the such problems as a communication protocol and a packet loss. In this paper, we implement and evaluate a division-based broadcasting system to solve the problem in the conventional system. In our proposed system, we confirm the effectiveness for reducing the waiting time using the scheduling method.

1 Introduction

Due to the recent popularization of the Internet, continuous media data broadcasting such as audio or video have attracted great attention [1]. In on-demand delivery, available bandwidth increases in proportion to the number of clients. On the other hand, in broadcasting, the server can deliver the data to many clients with a constant available bandwidth. However, waiting time from requesting data to playing it increases.

In division-based broadcasting, many studies employ the scheduling methods, which reduces waiting time by dividing the data into several segments and frequently broadcasting the precedent segments. These scheduling methods make the broadcast schedule considering the situation in actual network environments. Since most scheduling methods in broadcasting evaluate the waiting time in simulation environments, we need to evaluate the waiting time using the scheduling method in network environments.

We have been proposed a division-based broadcasting system [2], which can evaluate scheduling methods in network environments. Conventional division-based broadcasting systems have implemented the telecommunication protocols based on the wired LAN, and evaluate the scheduling methods. However, since the number of users using a smart phone and a computer increases, many clients connect to the Internet over wireless LAN. Since the bandwidth variation in wireless LAN is larger than that in wired LAN, when clients watch the

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_56

movie in actual network, the packet loss occurs. Therefore, we need to make the broadcasting system considering the case where clients watch the movie over wireless LAN.

In this paper, we propose the solution for this problem and implement the division-based broadcasting system in wireless LAN. In addition, we evaluate our proposed system and confirm the effectiveness for reducing the waiting time using conventional scheduling method.

The remainder of the paper is organized as follows. We explain the division-based broadcasting system in Section 2. Related works are introduced in Section 3. Our proposed system is designed in Section 4. Next, we implement and evaluate our proposed system in Sections 5 and 6. Finally, we conclude the paper in Section 7.

2 Division-based Broadcasting System

2.1 VoD and Broadcasting Systems

In webcasts, there are mainly two types of delivery systems: VoD and broadcasting. In the VoD system, the server starts delivering data sequentially based on client requests. Waiting times under VoD systems are roughly equal to receiving times. When the server repetitively broadcasts continuous media data, clients have to wait until the first portion of the data is broadcast.

When the server broadcasts the continuous media data repetitively, clients have to wait until the first portion of the data is broadcast. To reduce the waiting time, many methods employ the division-based broadcasting technique, which reduces the waiting time by dividing the data into several segments and broadcasting precedent segments frequently.

2.2 Waiting Time in Division-based Broadcasting

In the division-based broadcasting technique, since the data is divided into several segments, it is important to schedule segments without interrupting clients' continuous play. An example of division-based broadcasting technique is shown in Figure 1. The example uses the Fast Broadcasting (FB) [3] method to explain the problem caused by heterogeneous clients easily. In the FB method, the broadcast bandwidth is divided into several channels. Bandwidth for each channel is equivalent to the consumption rate. In this case, the server uses three channels. Also, the data is divided into three segments, S_1, S_2, and S_3. When the total playing time is 7 min., the playing time of S_1 is calculated to be 1 min., S_2 is 2 min., and S_3 is 4 min. The consumption rate is 5.0 Mbps and the available bandwidths for clients are 15 Mbps. Bandwidth for each channel is $15/3 = 5.0$ Mbps, which is the same as the data consumption rate. The server broadcasts S_i $(i = 1,2,3)$ via broadcast channel C_i repetitively. Clients can store broadcasted segments into their buffers while playing the data and can play each segment after receiving them. In this case, since clients can receive broadcasted segments from their midstream, the waiting time is the same as the time needed to receive only S_1 and the average waiting time is 1 min.

3 Related Works

In broadcasting systems, several scheduling methods to reduce waiting time have been proposed. In these methods, by dividing the data into several segments and producing a proper broadcast schedule, waiting time can be reduced.

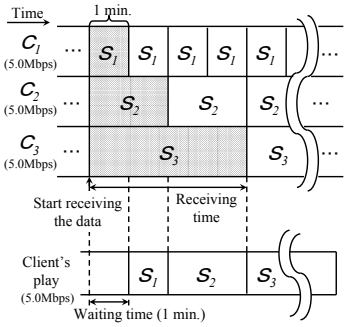


Fig. 1. Example of division-based broadcasting situation without interruption.

Fig. 2. Screenshot in *TeleCaS*.

In the Optimized Periodic Broadcast (OPB) [4], each data is separated into two parts. The server uses several broadcast channels and broadcasts each segment on each channel. When clients complete receiving the precedent parts of the content, they start receiving the rest portions of the data. Since clients can get the sub-segment data in advance, waiting time can be reduced. However, the bandwidth increases as the number of contents increases.

In the Hierarchical Stream Merging (HSM) [5], when the server broadcasts separated data via each channel repetitively, if clients complete receiving the data, the server merges the channel which the client receives. Since the method can reduce the necessary bandwidth, waiting time is reduced under the same bandwidth.

In BroadCatch [6], the server divides the data into 2^{K-1} segments of equal sizes. The server broadcasts them periodically using K channels. The bandwidth for each channel is the same as the data consumption rate. By adjusting K according to the available bandwidth for clients, waiting time is reduced effectively.

In the Layered Internet Video Engineering (LIVE) [7], clients feed back virtual congestion information to the server to assist both bandwidth sharing and transient loss protection. LIVE can minimize the total distortion of all participating video streams and maximize their overall quality. Also, several researches study the reliable broadcasting in wireless network with packet loss probability [8, 9].

In the Heterogeneous Receiver-Oriented Broadcasting (HeRO) [10], the server divides the data into K segments. Let J be the data size for the first segment. The data sizes for segments are J, $2J$, 2^2J, ..., $2^{K-1}J$. The server broadcasts them using K channels. However, the data size of the K^{th} channel becomes a half of the data, clients may have waiting time with interruptions.

In the BE-AHB method [11], since the data are divided into several segments, segments must be scheduled without interrupting clients' continuous play.

4 Design

4.1 *TeleCaS*

As explained in Section 3, many scheduling methods have been proposed to reduce waiting time in division-based broadcasting. However, most scheduling methods calculate the waiting

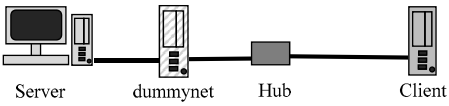


Fig. 3. Network environment for wired LAN in conventional *TeleCaS*.

time by computational simulation. To evaluate the scheduling method in actual network, we have proposed a division-based broadcasting system Telecommunication and BroadCasting System (*TeleCaS*) [2].

The screenshot is shown in Figure 2. We designed our proposed system using the C programming language and Java. Our proposed system can operate with several types of operating systems including the Windows system. In addition, we implemented the function that the server can operate the process by Graphical User Interface (GUI) to set the delivery of several segments easily. This system can make the cooperation of C language and Java by Java Native Interface (JNI). Clients play the data using the media player Totem [12].

4.2 Application to Wireless LAN Environment

Since the scheduling methods described in Section 2 suppose a situation where processing load and packet loss are not considered, they fail to consider the effect on division-based broadcasting systems. To evaluate scheduling methods in actual networks, we proposed a division-based broadcasting system called *TeleCaS*, which introduces several types of conventional broadcasting methods and designs Internet broadcasting systems based on a network environment with a server and clients.

Conventional *TeleCaS* constructs the wired LAN environment by connecting the server and clients. In Figure 3, the connection between the server and the client is routed through Dummynet that controls the bandwidth. With Dummynet, we can construct an evaluation environment to reproduce various network environments.

Due to the recent popularization of smart phones and compact notebooks, many clients can connect to the wireless LAN environment. Therefore, we need to modify the mechanism for wireless LAN environment in *TeleCaS*.

4.3 Packet Loss in Wireless LAN Environment

In division-based broadcasting system, the rate of packet loss in the wireless LAN is more increased than that in the wired LAN. Conventional *TeleCaS* has evaluated scheduling methods in the case of where the packet loss is not occurred. However, since the packet loss in the wireless LAN effects the waiting and interruption times, we modify the protocol considering the packet loss and implement the *TeleCaS* in the wireless LAN environment.

There are two types of packet loss, one is the packet loss for delivering the data by not delivering data to clients and damaging to the packet due to noise. When the client can not receive the packet from the server, or the packet that is received from the server is damaged, clients require a retransmission of the packet to the server. In the case of multicast with User Datagram Protocol (UDP), clients can not require a retransmission of the packet to the server. In wireless LAN environment, the rate of packet loss that is opportunity occurred is less than 1.0%. The packet loss is occurred by three reasons, which are the failure of the communication

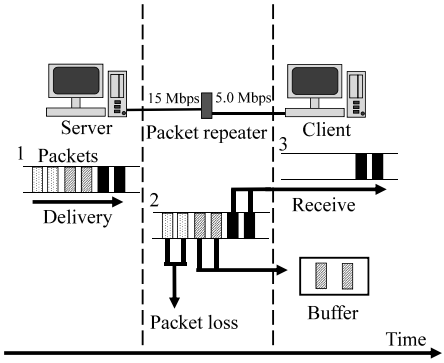


Fig. 4. Packet loss for computing environment.

path in a wired LAN environment, the interference from other radio waves in a wireless LAN environment, and a lack of data due to not reach the client.

The other is the packet loss for computing environment, which is the available bandwidth and the availability of the computer between the server and clients. In broadcasting system with wireless repeater, when the difference between the available bandwidth of the server and that of the client is large, the packet that can not be stored in the buffer area is discarded.

The flow of discarded packets that is not stored in the buffer area is shown in Figure 4. In Figure 4, the server delivers six packets, and the client makes the processing of playing data using two packets. When the rest of packets is four and the buffer size of clients is only two packets, although they can store two packets in four packets in their buffer, the rest of two packets is discarded. In on-demand delivery, clients can receive these packets again by requiring a retransmission of the packet to the server. On the other hand, in broadcasting, the server does not retransmit the packet based on the UDP. Therefore, in the communication protocol for *TeleCaS*, we need to add the mechanism about requiring the retransmission of the packet.

4.4 Rate of Packet Loss

In communication with the client and the radio base station, when the distance from the base station to the client becomes long, the packet loss may occur. In addition, when the client performs a wireless communication by Wi-Fi standard, the packet loss occurs due to the radio interference. Therefore, in the wireless LAN, we need to evaluate the rate of packet loss considering the system environments.

5 Implementation

5.1 *TeleCaS* over Wireless LAN

We implement the *TeleCaS* over the wireless LAN environment. To evaluate the *TeleCaS*, we uses the ethernet converter connected to the client machine via LAN cable. In addition, we use

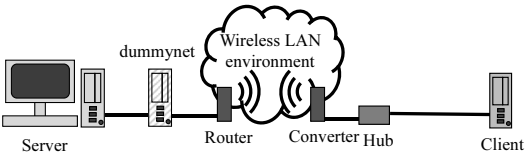


Fig. 5. Wireless LAN environment in *TeleCaS* (dummynet on server side).

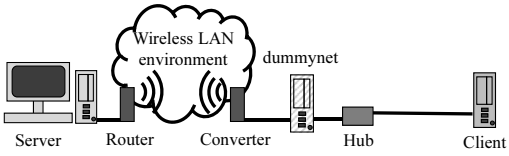


Fig. 6. Wireless LAN environment in *TeleCaS* (dummynet on client side).

the router for the wireless LAN environment The data format of *TeleCaS* in wireless LAN is the same as that in wired LAN.

5.2 Evaluation Items in *TeleCaS*

In *TeleCaS*, the server uses the IP multicast, the communication protocol uses UDP/IP. the server delivers data using UDP that is a connectionless protocol. Since clients have to wait until the data that their packets are lost are broadcast again, the interruption time occurs during playing the data.

We plan three types of evaluation. Firstly, to evaluate the influence of the available bandwidth, we calculate the rate of packet loss under different available bandwidths. Secondly, we evaluate the influence of data size based on the rate of packet loss. Finally, to evaluate the influence of the number of channels, we calculate the rate of packet loss under different available bandwidths.

5.3 Wireless LAN Environment in *TeleCaS*

We explain the structure of wireless LAN environment in *TeleCaS*. To control the available bandwidth, there are two cases in the location of dummynet that has a connection between the server and clients. One is a case of connecting the dummynet on the server side, which is shown in Figure 5. The other is a case of connecting the dummynet on the client side, which is shown in Figure 6. In division-based broadcasting system, the server can reduce waiting time by setting the broadcast scheduling based on the available bandwidth. When the clients receive data in *TeleCaS*, the server sets the available bandwidth. Therefore, we connect the dummynet on the server side.

In this paper, we evaluate the rate of packet loss in *TeleCaS*. The server delivers data to one client in wireless LAN environment.

Table 1. Measurement environment.

Server	CPU	Intel®Core 2 Duo E7500 (2.93 GHz)
	Memory	2.0 GBytes
	OS	Ubuntu 12.10
	NIC	RTL810E / RTL8102E
Client	CPU	Intel®Core 2 Duo E7500 (2.93 GHz)
	Memory	2.0 GBytes
	OS	Ubuntu 12.10
	NIC	RTL810E / RTL8102E
dummynet	CPU	Intel®Core 2 Duo E7500 (2.93 GHz)
	Memory	2.0 GBytes
	OS	FreeBSD 8.2-RELEASE
	NIC1	RTL8169SC
	NIC2	RTL8169SC
Router	Model	CG-WLBARGPXW
	Model	CG-WLBARGPXW
	Standard	IEEE802.11g
	Speed	Max 54 Mbps
Converter	Model	Aterm WG1200HP
	Model	PA-WG1200HP
	Standard	IEEE802.11g
	Speed	Max 867 Mbps

6 Evaluation

6.1 Evaluation Environment

We evaluate the performance in both actual and simulation environments. We use the FB method to evaluate the effectiveness of packet loss in wireless LAN environment, as explained in Subsection 4.3. The performances of machines, wireless LAN router, and Ethernet converter are shown in Table 1. We construct the evaluation environment based on IEEE 802.11g. To evaluate the packet loss in wireless LAN environment, we use the wireless LAN router that has a low spec. In addition, we use a converter for setting the desktop computer as a client in wireless LAN environment.

6.2 Effect of Bandwidth

In wireless LAN environment, we evaluate the interruption time and the rate of packet loss under different available bandwidth. The number of channels is 4, the playing time is 2 min., the data size is 21.4 Mbytes, and the consumption rate is 1.5 Mbps. The interruption time is the average value measured five times. The rate of packet loss in wireless LAN environment is about 1%, and we use a 2.4 Ghz frequently band.

The result of interruption time is shown in Figure 7. The vertical axis is the average interruption time. The horizontal axis is the available bandwidth. In Figure 7, in both wired and wireless LAN environments, interruption time is lengthened according to the available

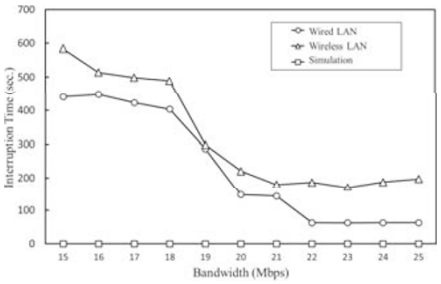


Fig. 7. Bandwidth and interruption time.

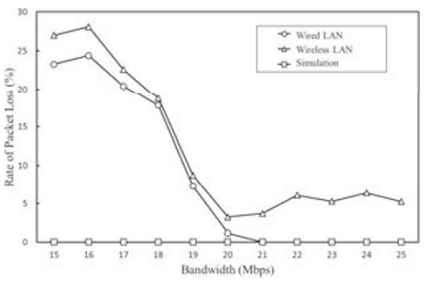


Fig. 8. Bandwidth and rate of packet loss.

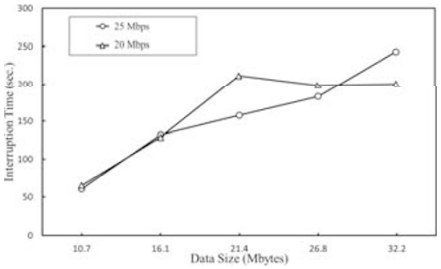


Fig. 9. Data size and interruption time.

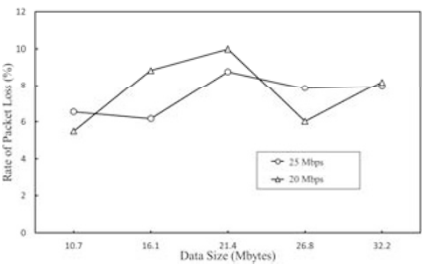


Fig. 10. Data size and rate of packet loss.

bandwidth. When the available bandwidth is less than the necessary bandwidth of server, interruption time occurs by the packet loss while delivering data.

The result of rate of packet loss is shown in Figure 8. The vertical axis is the rate of packet loss. The horizontal axis is the available bandwidth. In Figure 8, when the available bandwidth is more than 22 Mbps, the rate of packet loss is 0% in wired LAN environment, and that is about 5% in wireless LAN environment. In wireless LAN environment, the packet loss is occurred in only the case of where the server delivers data to clients.

6.3 Effect of Data Size

We evaluate the interruption time and the rate of packet loss under different data size. The number of channels is 4 and the consumption rate is 1.5 Mbps. There are five types of data size: 10.7 MBytes, 16.1 MBytes, 21.4 MBytes, 26.8 MBytes, and 32.2 MBytes. Therefore, the playing time of each data size is 60 sec., 90 sec., 120 sec., 150 sec., and 180 sec., respectively. The available bandwidths of server are 25 Mbps and 20 Mbps. The interruption time is the average value measured five times. The rate of packet loss in wireless LAN environment is about 1%.

The result of interruption time is shown in Figure 9. The vertical axis is the average interruption time. The horizontal axis is the available bandwidth. When the data size is 10.7 MBytes or 16.1 Mbytes, since the packet loss is not occurred, interruption time is the same. When the available bandwidth is 20 Mbps and the data size is 21.4 MBytes, interruption time

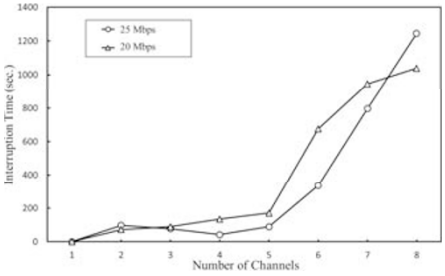


Fig. 11. Number of channels and interruption time.

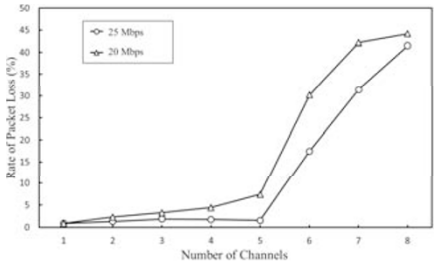


Fig. 12. Number of channels and rate of packet loss.

is greatly lengthened by occurring the packet loss. On the other hand, when the available bandwidth is 20 Mbps and the data size is 26.8 MBytes or 32.2 MBytes, since the rate of packet loss is not changed while delivering data, interruption time is not lengthened. In actual network, since the rate of packet loss is changed according to the data size, the interruption time is also lengthened. Next, when the data size is more than 32.2 MBytes, since the rate of packet loss is changed while delivering data, the interruption time is lengthened as increasing of the available bandwidth.

The result of rate of packet loss is shown in Figure 10. The vertical axis is the rate of packet loss. The horizontal axis is the available bandwidth. The rate of packet loss is from 5 to 10 % in every cases. On the other hand, since the sum of packet loss increases as increasing of the data size, the interruption time is lengthened.

6.4 Effect of Number of Channels

We evaluate the interruption time and the rate of packet loss under different number of channels. The playing time is 120 sec., the data size is 21.4 Mbytes, and the consumption rate is 1.5 Mbps. The available bandwidth of the server is 25 Mbps or 20 Mbps. The interruption time is the average value measured five times. The rate of packet loss in wireless LAN environment is about 1%, and we use a 2.4 Ghz frequently band.

The result of interruption time is shown in Figure 11. The vertical axis is the average interruption time. The horizontal axis is the available bandwidth. In addition, the result of rate of packet loss is shown in Figure 12. The vertical axis is the rate of packet loss. The horizontal axis is the available bandwidth. In Figure 11, interruption time is lengthened as the increasing of the number of channels. In Figure 12, the rate of packet loss becomes high as the increasing of the number of channels. When the number of channels is 2, since the packet loss is not occurred, the rate of packet loss is closely the same in both cases of where the available bandwidth is 25 Mbps or 20 Mbps. When the available bandwidth is 25 Mbps and the number of channels is less than 4, since the packet loss is not occurred, the packet loss is not changed. However, when the number of channels is more than 5, since the packet loss is occurred, the interruption time becomes lengthened.

7 Conclusion

In this paper, we proposed the division-based broadcasting system to evaluate the effectiveness using conventional scheduling method. In division-based broadcasting, many researches have proposed scheduling methods to reduce the waiting time. Although we have proposed the conventional division-based broadcasting system evaluate these methods, this system does not consider wireless LAN environment. In wireless LAN environment, we proposed the solution for the situation which the packet loss is occurred while delivering data. In addition, we implemented the system design and improved the conventional broadcasting system for wireless LAN environment. In our evaluation, we confirmed the effectiveness for reducing the waiting time using the scheduling method.

In the future, we will realize the system that reduces the communication traffic and the processing load of the server.

Acknowledgement. This work was supported by JSPS KAKENHI Grant Number 26730059, 15H02702, and 16K01065.

References

1. WHITE PAPER Information and Communications in Japan (2015). http://www.soumu.go.jp/johotsusintokei/whitepaper/eng/WP2015/2015-index.html.
2. Y. Gotoh and A. Kimura, "Implementation and Evaluation of Division-based Broadcasting System for Webcast," Journal of Digital Information Management (JDIM), Vol.13, Issue.4, pp.234-246, 2015.
3. L. Juhn and L. Tseng, Fast Data Broadcasting and Receiving Scheme for Popular Video Service, IEEE Trans. on Broadcasting, vol.44, no.1, pp.100-105, 1998.
4. L.-S. Juhn, and L.M. Tseng, "Harmonic broadcasting for video-on-demand service," IEEE Trans. Broadcasting, vol.43, no.3, pp.268-271, 1997.
5. Y. Zhao, D.L. Eager, and M.K. Vernon, "Scalable On-Demand Streaming of Non-Linear Media," Proc. IEEE INFOCOM, vol.3, pp.1522-1533, 2004.
6. M. Tantaoui, K. Hua, and T. Do, "BroadCatch: A Periodic Broadcast Technique for Heterogeneous Video-on-Demand," IEEE Trans. Broadcasting, vol.50, Issue 3, pp.289-301, 2004.
7. X. Zhu, R. Pan, N. Dukkipati, V. Subramanian, and F. Bonomi, "Layered internet video engineering (LIVE): Network-assisted bandwidth sharing and transient loss protection for scalable video streaming," Proc. IEEE INFOCOM, pp.226-230, 2010.
8. W. Xiao, S. Agarwal, D. Starobinski, and A. Trachtenberg, "Reliable Wireless Broadcasting with Near-Zero Feedback," Proc. IEEE INFOCOM, pp.2543-2551, 2010.
9. N. Fountoulakis, A. Huber, and K. Panagiotou, "Reliable Broadcasting in Random Networks and the Effect of Density," Proc. IEEE INFOCOM, pp.2552-2560, 2010.
10. K.A. Hua, O. Bagouet, and D. Oger, "Periodic Broadcast Protocol for Heterogeneous Receivers," Proc. Multimedia Computing and Networking (MMCN '03), vol.5019, No.1, pp.220-231, 2003.
11. T. Yoshihisa, M. Tsukamoto, and S. Nishio, A Broadcasting Scheme for Continuous Media Data with Restrictions in Data Division, Proc. IPSJ International Conference on Mobile Computing and Ubiquitous Networking (ICMU'05), pp.90-95, 2005.
12. "Totem," http://projects.gnome.org/totem/.

High Quality Multi-path Multi-view Video Transmission considering Path Priority Control

Tetta Ishida, Takahito Kiro, Iori Otomo, Takuya Fujihashi, Yusuke Hirota, Takashi Watanabe

Abstract Multi-view video is one of promising techniques for future video services. In multi-view video delivery, transmitted video frames are highly vulnerable for losses during transmission. When a video frame in a certain camera is lost, the effect of the loss may propagate to the subsequent video frames in the same camera and video frames in the neighbor cameras. This is referred to as multi-view error propagation. To improve the loss resilience, some multi-path-based video transmission schemes have been proposed for multi-view video delivery. The existing transmission schemes assume that each transmission path has the same packet loss characteristic and thus do not account for the path diversity. However, each transmission path often has different loss characteristic, e.g., loss-free and loss-prone. To exploit the path diversity for high video quality, we propose two multi-view video transmission schemes. The first scheme utilizes features of multi-view video encoding. Specifically, reference, i.e., important, video frames are transmitted over loss-free transmission path while the other video frames are transmitted over loss-prone transmission paths. The second scheme utilizes decoding operation of multi-view video. In this case, a sender sends odd video frames over loss-free transmission

Tetta Ishida
Osaka University, 1-5 Yamadaoka, Suita, Osaka, 565-0871 e-mail: ishida.tetta@ist.osaka-u.ac.jp

Takahito Kito
Osaka University, 1-5 Yamadaoka, Suita, Osaka, 565-0871 e-mail: kito.takahito@ist.osaka-u.ac.jp

Iori Otomo
Osaka University, 1-5 Yamadaoka, Suita, Osaka, 565-0871 e-mail: otomo.iori@ist.osaka-u.ac.jp

Takuya Fujihashi
Osaka University, 1-5 Yamadaoka, Suita, Osaka, 565-0871 e-mail: fujihashi.takuya@ist.osaka-u.ac.jp

Yusuke Hirota
Osaka University, 1-5 Yamadaoka, Suita, Osaka, 565-0871 e-mail: hirota.yusuke@ist.osaka-u.ac.jp

Takashi Watanabe
Osaka University, 1-5 Yamadaoka, Suita, Osaka, 565-0871 e-mail: watanabe@ist.osaka-u.ac.jp

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_57

path and exploits decoding operations for transmissions of even video frames. Evaluations show that our proposed scheme improves video quality (in terms of Peak Signal Noise Ratio) by 3 dB compared to existing transmission schemes when loss rate of transmission paths is 10%. In addition, our scheme achieves 1 MB lower video traffic in the same environment.

1 Introduction

Video streaming is becoming one of major applications over networks. Multi-view video delivery one of future send video frames of multiple cameras to provide 3-dimensional applications such as 3DTV [1] and free viewpoint TV [2] for users. In general, video traffic of multi-view video systems is large compared to that of single-view video systems because multi-view video systems send video frames of multiple cameras. To reduce the traffic, multi-view video systems utilize correlations between video frames in time and spatial domains for encoding. Specifically, a sender encodes a video frame of a camera as reference and then encodes the difference between the reference and same and other cameras' video frames to remove redundant information. However, video encoding causes high vulnerability for frame loss. During video transmissions, some video frames are lost due to network congestion. When a frame loss occurs, video quality sharply degrades due to multi-view error propagation [3]. To mitigate an effect of multi-view error propagation and quality degradation, several multi-view video transmission schemes have been proposed in [4, 5]. These schemes partially encode video frames and send them over multiple transmission paths.

The existing transmission schemes [4, 5] consider that each transmission path has the same quality and thus do not account for the path diversity. However, multiple transmission paths sometimes have different packet loss characteristic, e.g., loss-free and loss-prone. In this case, reference video frames are likely to be lost in loss-prone transmission paths because the video frames have a larger data size than the other frames. When a reference video frame is lost, the effect of the frame loss propagates to decoding operations of subsequent video frames and thus video quality will be significantly degraded.

In this paper, we propose two multi-view video transmission schemes to utilize the path diversity for high quality multi-view video delivery. Our first proposed scheme utilizes features of multi-view video encoding, namely, Frame Significance based Transmissions (FST). Specifically, a sender sends reference video frames and the other frames over loss-free and loss-prone transmission paths, respectively, to mitigate an effect of multi-view error propagation. To control multiple path's quality, we suppose Software Defined Networking (SDN) technology for path configuration. We also propose a transmission scheme, namely, Concealment-aware Transmissions (CT), which utilizes decoding and concealment operations for video transmissions. Specifically, a sender first encodes and sends odd frames over loss-free transmission paths. For even frames, the sender calculates residuals between the original pixel

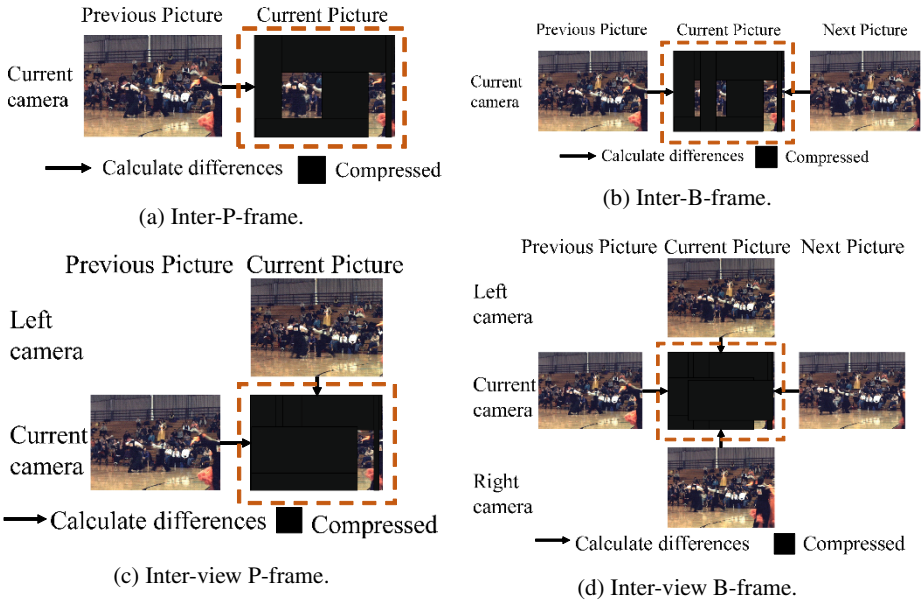


Fig. 1 P-frames and B-frames in multi-view video. The video frames surrounded by dotted lines indicate that the video frame is used for encoding and to compress redundant information in the other video frames.

values and average pixel values obtained from adjacent odd video frames. Finally, the residuals are encoded and transmitted over loss-prone transmission paths.

Evaluations show that our proposed schemes achieve better video quality compared to an existing transmission scheme. In addition, our second proposed scheme achieves lower video traffic compared to existing and our first proposed schemes.

2 Related Work

Our study is related to existing studies on multi-view video encoding, multi-path data transmissions, video delivery using path priority control, and multi-path multi-view video transmission schemes.

2.1 Multi-view Video Encoding

Figure 1 shows examples of general video frames, where the first two frames are common formats used for single-view video and the others are specialized for multi-view video. To encode a video sequence, three types of video frames are used in standard H.264/AVC: I-frames, inter-P-frames (Fig. 1 (a)), and inter-B-frames

(Fig. 1 (b)). I-frames have the lowest encoding/decoding computation requirements and their traffic is highest among all the video frame types. I-frames are encoded independently of the other video frames. Inter-P-frames have higher coding computation requirements and lower video traffic than I-frames. Inter-P-frames utilize motion compensation prediction between the current video frame and the previous video frame from the same camera to reduce the time domain redundancy, as shown in Fig. 1 (a). Inter-B-frames have higher coding computation requirements and lower video traffic than inter-P-frames. Inter-B-frames utilize the previous and next video frames from the same camera for motion compensation prediction to reduce the time domain redundancy greatly, as shown in Fig. 1 (b).

Multi-view video encoding in [6, 7] exploits two additional video frames to reduce the inter-camera domain redundancy: inter-view P-frames (Fig. 1 (c)) and inter-view B-frames (Fig. 1 (d)). Inter-view P-frames select video frames with previous time frames from the same camera and the same time video frame from another camera. Inter-view P-frames reduce the time and inter-camera domain redundancy by using selected video frames, as shown in Fig. 1 (c). Inter-view B-frames exploit a maximum of four video frames for motion compensation prediction: the previous and next time video frames from the same camera and the same time frame from other cameras. Inter-view B-frames reduce the time and inter-camera domain redundancy greatly by using four video frames, as shown in Fig. 1 (d).

Our scheme uses I-frame and inter-P-frame for video encoding. These video frames have larger video traffic than the other video frames while decrease the effect of frame losses.

2.2 Multi-path Data Transmission

Multi-path Transmission Control Protocol (TCP) has been proposed in [8, 9] to send data over multiple transmission paths. Specifically, this scheme divides each data flow into multiple subflows and transmits the subflows using multiple paths. The assumption in these multi-path-based studies is that partially received data is valuable for the receiver. However, when an encoded bit stream is transmitted using multiple transmission paths, data loss in a certain path induces loss of the whole frame, regardless of the status of the other transmission paths. This is because the transmitted data in each path is strongly correlated with the data in the other paths.

Our scheme encodes video frames into multiple descriptions that are mutually independent, to utilize the independence of multiple transmission paths. There are two advantages in our scheme. First, a frame loss in a certain transmission path does not affect quality of the other transmission paths. Second, a frame loss in a transmission path can be concealed by using received video frames from the other paths.

2.3 Video Delivery using Path Priority Control

Software Defined Networking (SDN) is one of the promising technologies for future networks and it can also construct multiple transmission paths between a sender and a user. In particular, SDN can set priority for transmission paths by using path priority control. For example, available transmission paths can be divided into loss-prone and loss-free paths by SDN controller.

Several studies on multi-path based video transmissions have been proposed for keeping high video quality by utilizing path priority control. References [10, 11] propose a multi-path video transmission for Quality of Experience (QoE)-aware path allocation by using SDN. These schemes divide encoded video frames into multiple descriptions based on the their significance, i.e., the frame has a great/little effect on video quality when the frame is lost. The descriptions are transmitted over multiple transmission paths with different loss rates to keep high video quality at a receiver side. For example, significant video frames are transmitted over loss-free transmission paths.

Our scheme follows the same motivation and extends their concepts to multi-view video transmissions. A sender transmits significant video frames of multiple cameras over loss-free transmission paths while the other video frames are transmitted over loss-prone paths.

2.4 Multi-path Multi-view Video Transmission

Reference [4] proposes a multi-path multi-view video transmission scheme. This scheme sends video frames of two cameras by using two transmission paths. Reference [5] extends the proposed scheme in [4] for a large number of cameras and transmission paths. These schemes encode video frames of multiple cameras into multiple versions and send each version over multiple paths to reduce quality degradation due to multi-view error propagation. However, these schemes do not utilize the quality diversity of transmission paths to maintain high video quality.

In this paper, we propose two multi-view transmission schemes. To utilize the diversity of transmission paths, the proposed schemes exploit features of encoding and concealment operations, respectively.

3 Proposed Framework

3.1 System Model

This paper proposes two multi-path multi-view transmission schemes to utilize transmission path diversity for high quality multi-view video delivery. Figure 2

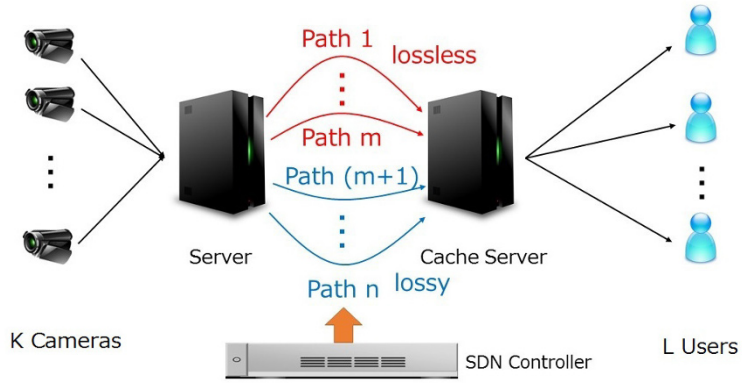


Fig. 2 Assumption model.

shows the system model of our study. In this scheme, K cameras capture the same object. The captured videos are transmitted to the encoding server over wired. The transmission paths between encoding server and cache server are constructed by SDN. The total number of transmission paths is n and m ($m < n$) paths are loss-free. We consider frame losses occur in $n - m$ transmission paths due to network congestion. In m transmission paths, SDN set a bandwidth limitation, i.e., B bps, for multi-view video transmissions to prevent congestion. In this paper, we assume three independent paths between encoding server and cache server. One path is loss-free and the others are loss-prone. The encoding server encodes and transmits video frames to the cache server by using multiple transmission paths. The cache server decodes the received video frames and conceals lost video frames by using the received video frames. Finally, the cache server sends video frames for L users.

3.2 Encoding and Transmission

In this paper, we propose two schemes for multi-view encoding and transmissions. In both schemes, the server encodes each camera's video frames in every Group of Pictures (GoP). A GoP is a set of video frames, typically consisting of eight frames. In each GoP, video frames are encoded into I-frame or inter-P-frame.

3.2.1 Frame Significance based Transmissions (FST)

This scheme utilizes features of multi-view video encoding for transmissions, i.e., transmitting reference video frames over loss-free transmission path to mitigate the effect of multi-view error propagation. Let $F_{i,j}$ denote a video frame of camera i at

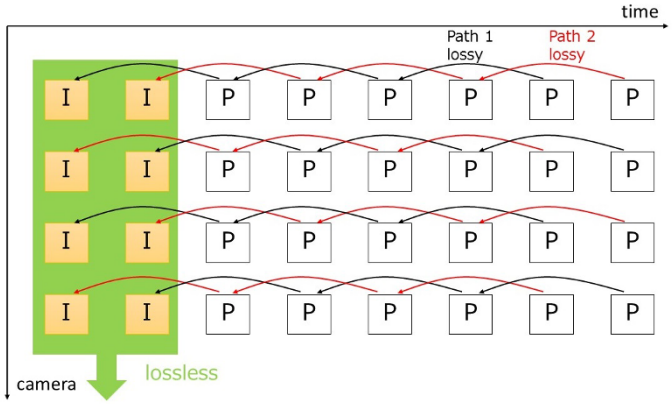


Fig. 3 Frame Significance based Transmissions.

time instant j. The number of video frames in one GoP is represented by N_{GoP}. The following list describes the steps of the video encoding operation.

1. Encode $F_{k,1}$ and $F_{k,2}$ into I-frames.
2. Encode $F_{k,2l}$ into P-frames using $F_{k,2l-2}$ as the reference frame.
3. Encode $F_{k,2l-1}$ into P-frames using $F_{k,2l-3}$ as the reference frame.

Here, $k = 1, 2, ...K$ and $l = 2, 3, ..., N_{\mathrm{GoP}}/2$.

Figure 3 shows encoding structure when the number of cameras is four. First, the server encodes first and second frames in GoP of each camera as I frame. The subsequent video frames of each camera are encoded into P frame. Each P frame uses every other video frame of the same camera as the reference frame. The server sends the first and second frames of each camera over a loss-free transmission path. By protecting the first and second frames, the impact of multi-view error propagation can be decreased. Other odd frames in cameras 1 and 3 and even frames in cameras 2 and 4 are transmitted to the cache server using path 1. The other frames are transmitted to the cache server using path 2.

3.2.2 Concealment-aware Transmissions (CT)

This scheme considers decoding and concealment operations for transmission in order to reduce video traffic. The following list describes the steps of the video encoding operation.

1. Encode $F_{k,1}$ into I-frames.
2. Encode $F_{k,2l-1}$ into P-frames using $F_{k,2l-3}$ as the reference frame.
3. Synthesize $F_{k,2l}$ by using the reconstructed $F_{k,2l-1}$ and $F_{k,2l-3}$.
4. Calculate residuals of $F_{k,2l}$ between original pixel values and synthesized pixel values.

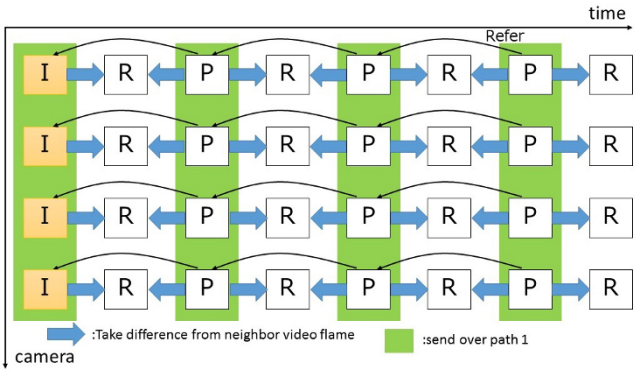


Fig. 4 Concealment-aware Transmissions.

5. Encode $F_{k,2l}$ into I-frame. We define these encoded video frames as Residual Frames (R-frames).

Figure 4 shows encoding structure of our second proposed scheme at four cameras. Odd frames of each camera are transmitted over a loss-free transmission path and R-frames are alternately transmitted over two loss-prone transmission paths. After the transmission, the cache server first decodes the video frames received from the loss-free path and then synthesizes the intermediate frames by using the decoded frames. Finally, correctly received R-frames are added to the synthesized frames for decoding operation.

3.3 Loss Concealment

During video transmissions from the server to the cache server, several frame losses occur in loss-prone transmission paths due to congestion. To conceal the lost video frames, a cache server uses neighbor decoded video frames [4]. The cache server first uses adjacent video frames in time direction. If these video frames are not be used for the concealment due to the frame loss, the cache server uses adjacent video frames in camera direction.

4 Evaluation

We use Joint Multi-view Video Coding (JMVC) encoder [12] and Ballroom that multi-view test video sequences provided by Mitsubishi Electric Research Laboratories (MERL) [13]. The number of cameras is 8. The frame rate is 25 fps, and the GoP size is 8 frames. We evaluated the performance between an encoding server

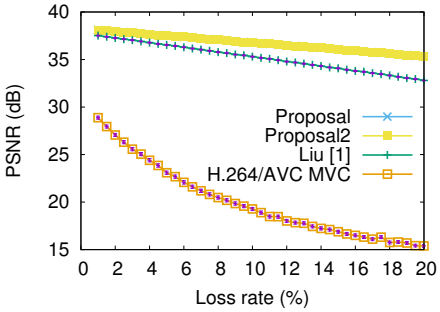


Fig. 5 Video quality vs. Loss rates.

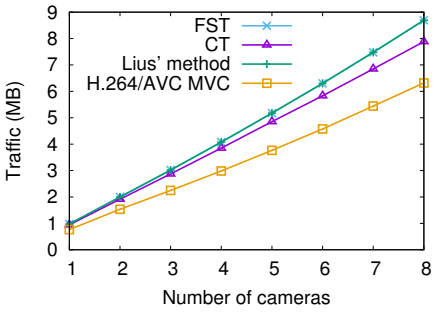


Fig. 6 Traffic vs. Number of cameras.

and a cache server. The two servers are connected by three independent paths. Path 1 is loss-free path. Path 2 and path 3 are congested paths. Transmission capacity in loss-free path is 5 Mbps. The loss rate in Path 2 is 5%. To evaluate an impact of loss rates, we use different loss rates in path 3. For comparison, we use three reference schemes: H.264/AVC Multi-view Video Coding (MVC), multi-path transmission scheme [4] (Liu's method), and our proposed schemes, namely, frame significance based transmissions (FST) and concealment-aware transmissions (CT).

Figure 5 shows the video quality as a function of the loss rate in path 3. Here, odd P-frames in each camera are transmitted in path 2 while even P-frames are transmitted in path 3. From this figure, we can observe three following points:

1. Proposed FST achieve higher video quality compared to existing schemes.
2. FST yields better performance than our second proposedCT. For example, FST improves video quality by 1.7 dB compared to CT at a loss rate of 10%.
3. H.264/AVC MVC degrades video quality as a loss rate increases. When one video frame is lost, the effect propagates to video frames in the same and adjacent cameras.

Figure 6 shows the video traffic as a function of the number of camera. The second and sixth R-frames in each camera are transmitted in path 2 while fourth and sixth R-frames are transmitted in path 3. From this figure, we can see the following results:

1. CT yields better performance than FST and Liu's method. For example, CT decreases the video traffic by 1 MB compared to FST and existing scheme in Liu's method at a number of cameras of 8.
2. H.264 AVC/MVC decreases video traffic compared to both proposed schemes. H.264/AVC MVC fully exploits correlations between video frames for compression.

5 Conclusion

In this paper, we proposed two multi-path based multi-view video transmission schemes to utilize path priority control of SDN for high quality multi-view video delivery. To this end, our first proposed scheme utilizes features of multi-view video encoding and our second proposed scheme utilizes decoding and concealment operations at a receiver for video transmissions. Evaluations show that our proposed scheme mitigates degradation of video quality compared to existing schemes.

Acknowledgements This work was supported by JSPS KAKENHI Grant Number 15K12018.

References

1. K. Muller, H. Schwarz, D. Marpe, C. Bartnik, S. Bosse, H. Brust, T. Hinz, H. Lakshman, P. Merkle, F.H. Rhee, G. Tech, M. Winken, and T. Wiegand, 3D high-efficiency video coding for multi-view video and depth data, IEEE Transactions on Image Processing, vol.22, no.9, pp.3366-3378, Sept. 2013.
2. M. Tanimoto and K. Suzuki, Global View and Depth (GVD) format for FTV/3DTV, Three-Dimensional Imaging, Visualization, and Display, 10 pages, May 2013.
3. T. Fujihashi, K. Kodera, S. Saruwatari, and T. Watanabe, "Wireless Multi-view Video Streaming with Subcarrier Allocation," IEICE Transactions on Communications, Vol.E99-B, No.02, pp.542-554, Feb. 2016.
4. Zhi Liu, Gene Cheung, Jacob Chakareski, and Yusheng Ji, "Multiple Description Coding & Recovery of Free Viewpoint Video for Wireless Multi- Path Streaming", IEEE Journal of Selected Topics in Signal Processing, vol. 9, no. 1, pp. 151-164, 2015.
5. Iori Otomo, Takuya Fujihashi, Yusuke Hirota, and Takashi Watanabe, "Loss Resilient Multi-view Video Streaming over Multiple Transmission Paths," IEEE International Conference on Communications, pp. 1-6, 2016.
6. H. Zeng, X. Wang, C. Cai, J. Chen, and Y. Zhang, "Fast multiview video coding using adaptive prediction structure and hierarchical mode decision," IEEE Transactions on Circuits and Systems for Video Technology, pp.1–13, 2014.
7. K. Muller, H. Schwarz, D. Marpe, C. Bartnik, S. Bosse, H. Brust, T. Hinz, H. Lakshman, P. Merkle, F.H. Rhee, G. Tech, M. Winken, and T. Wiegand, "3D high-efficiency video coding for multi-view video and depth data," IEEE Transactions on Image Processing, vol.22, no.9, pp.3366–3378, 2013.
8. L. Li and M. Chen, "AMTCP : An Adaptive Multi-path Transmission Control Protocol." Proceedings of the 12th ACM International Conference on Computing Frontiers. ACM, 2015.
9. X. Corbillon, et al. "Cross-layer scheduler for video streaming over MPTCP." Proceedings of the 7th ACM International Conference on Multimedia Systems, 2016.
10. H. Egilmez, S. Civanlar, and A. Murat, "An Optimization Framework for QoS-Enabled Adaptive Video Streaming Over Open-Flow Networks", IEEE Transactions on Multimedia, vol. 15, no. 3, pp. 710-715, 2013.
11. M. Karl, J. Gruen and T. Herfet, "Multimedia optimized routing in OpenFlow networks," 19th IEEE International Conference on Networks, pp. 1-6, 2013.
12. Joint Video Team of ITU-T VCEG and ISO/IEC MPEG: "JMVC (Joint Multiview Video Coding) Software" (2008).
13. Text of ISO/IEC 14496-10:2008/FDAM 1 ISO/IEC JTC1/SC29/WG11: "Multiview Video Coding" (2008).

A Recursive Continuous Query Language for Integration of Streams and Graphs

Yousuke WATANABE

Abstract Query processing data streams obtained from sensors becomes more and more important today. We have many kind of query processing requirements for not only a single stream, but also multiple streams from different sensors. To utilize multiple streams, we have to consider positional relationships between sensors. Conventionally, they are represented by distances in the 3-dimensional space, or a hierarchy of logical areas (e.g. rooms/floors/buildings). Some stream applications need more sophisticated relationships that are represented in a graph structure. However, query languages for streams are not integrated with ones for graphs, thus we cannot uniformly treat them in the same language. This paper proposes a DataLog-like query language which can utilize multiple streams from streams having graph relationships. The proposed query language supports continuous queries on streams by containing event specifications and sliding-windows for streams. And, it also supports recursive search for graph relationships.

1 Introduction

Today, large volume of sensor data are obtained from many embedded devices such as mobile phones, surveillance cameras, connected vehicles and so on. Each sensor produces a new data item continuously. An infinite sequence of data items from the sensor is regarded as a data stream. Due to the increase of data streams, query processing on data streams becomes more and more important. We have many kind of query processing requirements for not only a single stream, but also multiple streams from different sensors. Typical examples of such requirements are joining multiple streams and applying aggregate functions (e.g. max, average).

When we utilize multiple streams from different sensors, we often select a subset of sensors that are highly related each other. A placement of each sensor is an im-

Yousuke WATANABE
Nagoya University, e-mail: watanabe@coi.nagoya-u.ac.jp

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_58

portant property, and spatial or logical relationships are derived from placements of sensors. Here, we introduce two types of positional relationships between sensors. The first one is a relationship in the 3-dimensional space. A position of each sensor is represented as a point in the 3-dimensional space. Sensors, whose distances are "close" to each other, can possibly monitor the same event in the real world. Thus, we often utilize them together. The second one is a relationships in a hierarchy of logical areas. For example, suppose a hierarchy of rooms/floors/buildings. Each sensor belongs to a room, a floor and a building. Streams from these sensors are often coupled based on the hierarchy. Sensors in the same room are strongly related, but sensors in the same building are weakly related.

Some stream applications need more sophisticated relationships between sensors. For example, suppose a traffic application which treats in-vehicle-sensors. Vehicles are running on a road network which is a graph consisting of roads and intersections. Relationships between in-vehicle-sensors on the road network can be regarded as a graph structure.

However, since query languages for streams are not integrated with ones for graphs. To process stream data, we need continuous queries which are repeatedly evaluated on stream data according to time progress and data arrival. On the other hand, for graph data, we need recursive search to traverse multi-hop relationships. Two types of query languages have been studied separately. Existing query languages cannot uniformly treat them.

This paper proposes a DataLog-like query language which can utilize multiple streams from sensors having graph relationships. DataLog [4] is a logic query language for relational models, and it supports recursive search for relations. The proposed query language is an extension of DataLog, thus it can process recursive queries for graph relationships stored in relations. And, it also supports continuous queries by introducing event specifications and sliding-windows for streams. This paper provides an example scenario in which in-vehicle-sensors on the road network are utilized.

The remaining part of this paper is organized as follows: Section 2 introduces related work. Section 3 presents an example scenario. Section 4 proposes the proposed query language. Section 5 concludes this paper and describes future research issues.

2 Related work

This section describes related work.

Recursive queries for relational model

DataLog (Database Logic) [4] is a logic query language for relational models. It has higher expressive power than the relational algebra, because DataLog supports

recursive queries. When graph data are stored in relations, DataLog is applicable to the graph data. This research extends DataLog to handle stream data. SQL:1999 also supports recursive queries. "WITH RECURSIVE" clause is used to achieve recursion. The syntax of DataLog is more simpler than one of SQL, so it is easy to extend and implement.

Continuous queries for stream data

CQL [1] is a continuous query language for relational streams. A relational stream is a sequence of tuples belonging to the relation. CQL defines three types of operators: stream-to-relation, relation-to-stream and relation-to-relation. A stream-to-relation operator is a sliding window which is a finite set of tuples within the specified time range. Relation-to-stream operators (I-Stream, D-Stream, R-Stream) are used to extract differential results from relations. Relation-to-relation operators are the same as ones of the traditional relational algebra. Since CQL does not focus on searching graph data, recursive search is not considered in CQL. C-SPARQL [3] is a continuous query language for RDF streams. RDF-stream is one of graph streams, and it is an infinite sequence of triples (subject, predicate, object) with timestamps. Though C-SPARQL handles graph data, it does not support recursive queries.

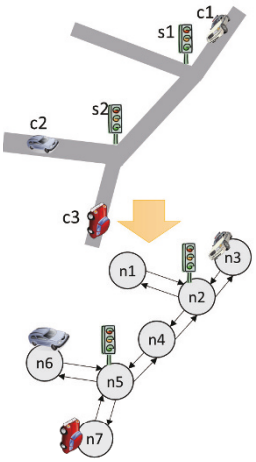
The purpose of the proposed query language in this paper is utilizing multiple streams with graph relationships. The main purpose is to uniformly handle relational streams and graphs stored in relations.

3 Example Scenario

This section presents an example scenario for introduction of the proposed query language. Here, we assume a traffic application which treats streams from in-vehicle-sensors and a graph of a road network. Recent connected vehicles have wireless network devices and many types of embedded sensors. Stream data from in-vehicle-sensors are continuously sent to the data center.

A road network is represented by a directed graph stored in two relations (Figure 1). In this research, roads and intersections are regarded as nodes in the graph. A relation named *Road* has two attributes: *Road_id* and *Length*. Connections between roads and intersections are directed edges in the graph. A relation named *Connection* has two attributes: *Source* and *Destination*.

Streams in this scenario are obtained from in-vehicle-sensors and traffic lights (Figure 2). We employ the relational stream model. A relational stream is an infinite sequence of tuples belonging to the relation. A stream named *Vehicle* has five attributes: *Timestamp*, *Vehicle_id*, *Road_id*, *Position* and *Velocity*. Each vehicle has a unique identifier, and it continuously reports the current road id on which the vehicle is running. A stream named *Signal* has five attributes: *Timestamp*, *Signal_id*,



Road

Road_id	Length
n1	180
n2	40
n3	150
n4	200
n5	40
n6	160
n7	200

Connection

Source	Destination
n1	n2
n2	n1
n2	n3
n2	n4
n3	n2
n4	n2
n4	n5
n5	n4
n5	n6
n5	n7
n6	n5
n7	n5

Fig. 1 A graph data of road network

Vehicle

Timestamp	Vehicle_id	Road_id	Position	velocity
t1	c1	n3	(x1,y1)	60
t1	c2	n6	(x2,y2)	55
t1	c3	n7	(x3,y3)	58
t2	c1	n3	(x4,y4)	60
t2	c2	n6	(x5,y5)	55
t2	c3	n7	(x6,y6)	58
...	...	...	...	...

Signal

Timestamp	Signal_id	Intersection	Control_Road	Color
t1	s1	n2	n1	Red
t1	s2	n5	n6	Red
t2	s1	n2	n1	Red
t2	s2	n5	n6	Red
...	...	...	...	...

Fig. 2 Streams from in-vehicle-sensors and signals

Intersection, *Control_Road* and *Color*. Each traffic light (signal) is located on an intersection, and it controls admission of vehicles from *Control_Road*.

Since vehicles and signals are on the road network, relationships between sensors compose a graph. To traverse multiple hops on the road network, we need recursive search. To process streams from sensors, we need continuous queries.

$$
\begin{aligned}
\mathit{Query} &:= \mathit{Event_Spec}\ \mathit{Rules} \\
\mathit{Event_Spec} &:= \text{"On"}\ \mathit{Events}\ \text{":"} \\
\mathit{Events} &:= <\mathit{Identifier}> \mid (\ <\mathit{Identifier}>\ \text{","}\ \mathit{Events}\ ) \\
\mathit{Rules} &:= \mathit{Rule} \mid (\ \mathit{Rule}\ \mathit{Rules}\ ) \\
\mathit{Rule} &:= \mathit{Head} \leftarrow \mathit{Body}\ \text{";"} \\
\mathit{Head} &:= \mathit{Relational_Atom} \\
\mathit{Body} &:= \mathit{Atom} \mid (\ \mathit{Atom}\ \text{"AND"}\ \mathit{Body}\ ) \\
\mathit{Atom} &:= \mathit{Relational_Atom} \mid \mathit{Stream_Atom} \mid \mathit{Arithmetic_Atom} \\
\mathit{Relational_Atom} &:= <\mathit{Identifier}>\ \text{"("}\ \mathit{Variables}\ \text{")"} \\
\mathit{Stream_Atom} &:= <\mathit{Identifier}>\ \text{"["} <\mathit{Window}>\ \text{"]"}\ \text{"("}\ \mathit{Variables}\ \text{")"} \\
\mathit{Arithmetic_Atom} &:= <\mathit{Identifier}>\ \mathit{Comparator} <\mathit{Constant}> \\
\mathit{Variables} &:= <\mathit{Identifier}> \mid (\ <\mathit{Identifier}>\ \text{","}\ \mathit{Variables}\ ) \\
\mathit{Comparator} &:= \text{"<"} \mid \text{"<="} \mid \text{">"} \mid \text{">="} \mid \text{"="} \mid \text{"!="}
\end{aligned}
$$

Fig. 3 A part of syntax tree of the proposed query language

4 Proposed Query Language

This section explains the proposed query language. As described in Section 1, it is an extension of DataLog [4]. Event specification and sliding-windows are introduced to handle stream data. Since DataLog originally supports recursive queries, the proposed query language also supports them.

4.1 Syntax

The syntax the proposed query language is also an extension of DataLog. Figure 3 presents a part of syntax tree of the query language. $<\mathit{Identifier}>$ means a name of relation/stream/variable.

A query $Q(E,R)$ consists of an event specification E and a set of rules R. An event specification E is a set of stream names. An evaluation of Q is triggered on each data arrival from the streams included in E.

A form of each rule r ($\in R$) is $\mathit{Head} \leftarrow \mathit{Body}$. Head is an output relation or an intermediate relation. *Body* is a conjunction (logical AND) of atoms. The evaluation results of *Body* are stored into the relation written in Head.

There are three types of atoms: relational atom, stream atom and arithmetic atom. A relational atom consists of a name of relation and attribute variables. A stream atom consists of a name of relation, a sliding-window, and attribute variables. A sliding-window is specified by time-interval. An arithmetic atom consists of a condition between an attribute variable and a constant. Currently, we do not consider logical NOT in arithmetic atom.

4.2 Examples

Now, we provide some example queries according to the scenario in Section 3.

Example1: Filtering Vehicle stream

The following example is a simple query which performs filtering on *Vehicle* stream.

On Vehicle :

$$Vehicle_c1(t,i,r,p,v) \leftarrow Vehicle[1second](t,i,r,p,v)\ AND\ i =' c1';$$

This query consists of an event specification and one rule. Sicne *Vehicle* is specified at the event specification (on-clause), this query is triggered by each arrival from *Vehicle* stream. In the rule, $Vehicle_c1$ is a name of output stream. t, i, r, p and v are attribute variables. $[1second]$ is a window specification on *Vehicle* stream. $i =' c1'$ is the filtering condition.

Example2: Union of two filtering results on Vehicle stream

The following example is a query which merges two filtering results on *Vehicle* stream.

On Vehicle :

$$Vehicle_c1c2(t,i,r,p,v) \leftarrow Vehicle[1second](t,i,r,p,v)\ AND\ i =' c1';$$
$$Vehicle_c1c2(t,i,r,p,v) \leftarrow Vehicle[1second](t,i,r,p,v)\ AND\ i =' c2';$$

This query consists of an event specification and two rules. The output relation $Vehicle_c1c2$ stores union of two filtering results from two rules.

Example3: Joining two streams

The following example is a query which provides the current signal color on the intersection to a vehicle $c1$.

On Signal :

$$\begin{aligned} Signal_c1(t2,s2,col2) \leftarrow\ & Vehicle[1second](t1,i1,r,p1,v1) \\ & AND\ Signal[1second](t2,s2,i2,r,col2) \\ & AND\ i1 =' c1'; \end{aligned}$$

The query is triggered by each arrival from *Signal* stream, then it joins *Vehicle* and *Signal*. Since variable r appears two times ($Road_id$ in *Vehicle* and $Control_Road$ in

Signal), this implies joining *Vehicle* and *Signal* based on the condition $Road_id = Control_Road$. The output relation contains three attributes: *Timestamp*, *Signal_id* and *Color*.

Example4: Finding the next road of Vehicle 'c1'

The following example is a query which integrates *Vehicle* stream and a graph. The query finds the next road (or intersection) of vehicle $c1$.

On Vehicle :

$$\begin{aligned} NextRoad_c1(t1,i1,d2) \leftarrow\ & Vehicle[1second](t1,i1,r,p1,v1) \\ & AND\ Connection(r,d2) \\ & AND\ i1 =' c1'; \end{aligned}$$

Since variable r appears two times (*Road_id* in *Vehicle* and *Source* in *Connection*), this implies joining *Vehicle* and *Connection* based on the condition $Road_id = Source$. The output relation contains three attributes: *Timestamp*, *Vehicle_id* and *Destination*.

Example5: Finding reachable roads recursively

The following example is a query which recursively finds reachable roads from the current road of vehicle $c1$.

On Vehicle :

$$\begin{aligned} MultiHop_c1(t1,i1,d2,1) \leftarrow\ & Vehicle[1second](t1,i1,r,p1,v1) \\ & AND\ Connection(r,d2)\ AND\ i1 =' c1'; \\ MultiHop_c1(t1,i1,d2,h1+1) \leftarrow\ & MultiHop_c1(t1,i1,r,h1) \\ & AND\ Connection(r,d2)\ AND\ h1 \leq 2; \end{aligned}$$

This query contains two rules. Figure 4 illustrates an execution of this query. The first rule is applied once (the 1st round), and the second rule is applied twice (the 2nd round and the 3rd round). The first rule generates the initial result of recursion. The output relation contains four attributes: *Timestamp*, *Vehicle_id*, *Destination* and the number of hops. The second rule stops when the number of hops becomes 3 (because the terminal condition is $h1 \leq 2$).

Example6: Finding vehicles on reachable roads

The following example is a query which finds vehicles on reachable roads from the current road of vehicle $c1$.

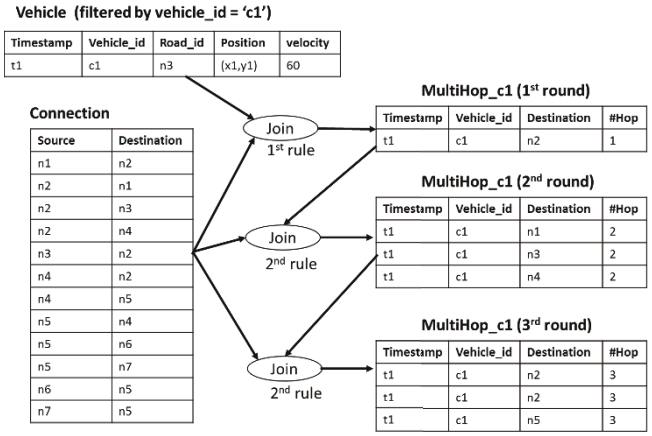


Fig. 4 The query result of Example 5

$$
\begin{aligned}
&\textit{On Vehicle}: \\
&\textit{Other_Vehicle}(t1,i1,r,p1,v1) \leftarrow \textit{Vehicle}[1second](t1,i1,r,p1,v1) \\
&\qquad\qquad \textit{AND MultiHop_c1}(t2,i2,r,h2) \\
&\qquad\qquad \textit{AND } i1 \; !=' \; c1'; \\
&\textit{MultiHop_c1}(t1,i1,d2,1) \leftarrow \textit{Vehicle}[1second](t1,i1,r,p1,v1) \\
&\qquad\qquad \textit{AND Connection}(r,d2) \textit{ AND } i1 =' c1'; \\
&\textit{MultiHop_c1}(t1,i1,d2,h1+1) \leftarrow \textit{MultiHop_c1}(t1,i1,r,h1) \\
&\qquad\qquad \textit{AND Connection}(r,d2) \textit{ AND } h1 \leq 2;
\end{aligned}
$$

This query contains three rules. Similar to Example 5, the second rule and the third rule are recursively applied to find reachable roads for the current road of vehicle $c1$.

4.3 Processing queries

Queries written in the proposed query language are executed by our Data Stream Management System (DSMS). Here, we show a brief summary of query processing.

1. A query is registered to the DSMS. Each rule in the query is converted into a operator tree consisting of selection, projection and join operators. A rule in DataLog is equivalent to a combination of operators in relational algebra.
2. Tuples belonging to relational streams from sensors arrive at the DSMS. If arriving tuples do not belong to the streams in the event specification, the query is not triggered.

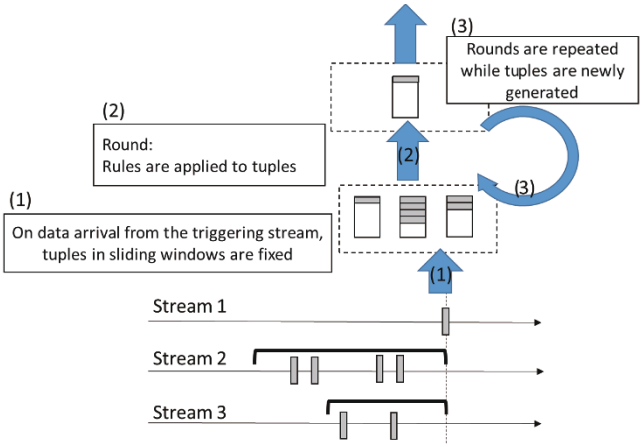


Fig. 5 Query processing on each event

3. When a tuple arrives from the stream whose name is included in the event specification, the DSMS evaluates the query. Figure 5 illustrates query evaluation on each data arrival.

   a. For stream data, tuples within sliding-windows are detected and fixed.
   b. The DSMS starts *round*, in which the operator trees are applied to tuples included in sliding-windows and relations. It generates the results for intermediate relations and the output relation.
   c. When tuples are newly generated in the last round, the DSMS tries the next round for recursive queries. If no tuples are generated in the last round, then the DSMS finishes the evaluation triggered by the data arrival.

5 Conclusion

This paper proposed a DataLog-like query language which can utilize multiple streams from sensors with graph relationships. As an example scenario, a traffic application for in-vehicle-sensors on a road network was presented.

There are some future research issues. The query parser for the query language is under development. It will be implemented in our DSMS. An efficient query processing method is also to be studied. We need an incremental evaluation algorithm for a combination of sliding-window and recursive queries.

Acknowledgements This work was supported in part by JSPS KAKENHI Grant Number 25240014.

References

1. A. Arasu, S. Babu, and J. Widom. The cql continuous query language: Semantic foundations and query execution. The VLDB Journal, 15(2):121-142, 2006.
2. SPARQL Query Language for RDF. http://www.w3.org/TR/rdf-sparql-query/
3. D. F. Barbieri, D. Braga, S. Ceri, E. D. Valle, M. Grossniklaus. Querying RDF streams with C-SPARQL. SIGMOD Record, Vol.39 No.1, pp.20–26, 2010.
4. Hector Garcia-Molina, J. D. Ullman, J. Widom. Database Systems: The Complete Book (2nd Edition). Pearson, 2008.

A Design of a Video Effect Process Allocation Scheme for Internet Live Broadcasting

Tomoya Kawakami, Yoshimasa Ishi, Satoru Matsumoto, Tomoki Yoshihisa, and Yuuichi Teranishi

Abstract In this paper, we propose and implement a method to distribute loads on a video processing system for Internet live broadcasting. Recently, Internet live broadcasting services are widely spread. In the Internet live broadcasting services, users add special effects to videos in real time. However, a delay occurs to process and broadcast videos because of the poor performances of users' devices such as PCs and mobile devices. Therefore, our proposed method assumes that high-performance computers on the Internet and users request the computers to add video effects. Our proposed method selects a specific computer from several computers to process each video based on a P2P mobile agent system. Our experimental results show that our proposed method can reduce the time to process and broadcast videos.

1 Introduction

Internet live broadcasting services are widely spread such as Ustream and YouTube Live because of recent advances in video delivery technologies [1]. In those services, users can easily broadcast real-time videos from their cameras.

Tomoya Kawakami
Nara Institute of Science and Technology, Ikoma, Nara, Japan, e-mail: kawakami@is.naist.jp
Osaka University, Ibaraki, Osaka, Japan

Yoshimasa Ishi
PIAX Inc., Kita-ku, Osaka, Japan
Osaka University, Ibaraki, Osaka, Japan

Satoru Matsumoto and Tomoki Yoshihisa
Osaka University, Ibaraki, Osaka, Japan

Yuuichi Teranishi
National Institute of Information and Communications Technology, Koganei, Tokyo, Japan
Osaka University, Ibaraki, Osaka, Japan

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_59

In Internet live broadcasting services, users add various effects to their videos such as image recognition, superimposing, image processing. By adding video processing quickly, users can add more special effects to make their videos more exciting. The users' requirements to add video effects on Internet live broadcasting are the following;

- Users can specify the time to add video effects
  Users desire to add video effects anytime to show temporal information, warm up the audiences, and so on, not only the effects that show always such as a clock and a logo. As an example of these temporal effects, users add the names of participants in real time video.
- Users can add requesting effects
  Users desire to add high level effects to make easy to watch the video and warm up audiences, and so on, not only simple effects such as showing a clock and a logo. As an example of these high level effects, the system recognizes human faces and emphasizes them in the real time video by adjusting those colors. To satisfy these requests from users, the system needs to add and remove functions for video effects flexibly.

However, a delay occurs to process and broadcast videos with anytime and high level video effects because of the poor performances of users' devices such as PCs and mobile devices. For example, users need to put additional information previously because it takes long time to recognize objects. Although high performance computers can reduce time to add video effects and solove the above problems, those computers are too expensive to be purchased by end-users.

In this paper, we assume that video recording nodes (clients) request video processing servers to add video effects through the Internet. The recording nodes are computers to record and broadcast videos through the Internet. The recording nodes also have a library to process videos and send the library in the case of requesting video processing. The video processing servers are high performance computers. The libraries are computer programs to add various video effects such as image recognition, superimposing, and image processing. Video broadcasting is executed by computer software on the recording nodes.

If a huge number of recording nodes requests the same server to process videos, users cannot smoothly broadcast videos because the video processing server has a heavy load and takes long time to process the requested tasks. Currently, several distributed video processing systems have been researched to process videos on different devices from those recording devices. However, the existing works do not satisfy the broadcasters' requirements to add high level video effects anytyme. Therefore, we propose a load distribution method that selects a specific processing server from several servers through the Internet. Video processing servers construct a P2P network to distribute the loads for video processing. Broadcasters send video frame data with a video effect library, requesting video effects are satisfied by the video processing server. We implement the proposed system by PIAX[1], which is platform middleware for P2P agent systems [2].

[1] PIAX (in Japanese): http://www.piax.jp/

In the following, our assumptions are defined in Section 2. The proposed method and its implementation are described in Section 3, and its experiment is presented in Section 4. We describe the related work in Section 5. Finally, the conclusion of the paper is presented in Section 6.

2 Assumptions

In this section, we explain our assumptions of a video processing system for Internet live broadcasting.

2.1 Video Processing Model

For personal users, it is difficult to prepare high performance computers only to process videos smoothly. In this paper, we assume that recording nodes request video processing servers to add video effects through the Internet. Figure 1 shows the flow to process and broadcast a video in our assumptions.

The recording nodes are computers to record and broadcast videos through the Internet. The recording nodes also have a library to process videos and send the library in the case of requesting video processing. The video processing servers are high performance computers that are managed by providers of Internet live broadcasting services. The libraries are computer programs to add various video effects such as image recognition, superimposing, and image processing. Video broadcasting is executed by computer software on the recording nodes. The broadcasting software includes a type to execute on a web browser and upload the processed

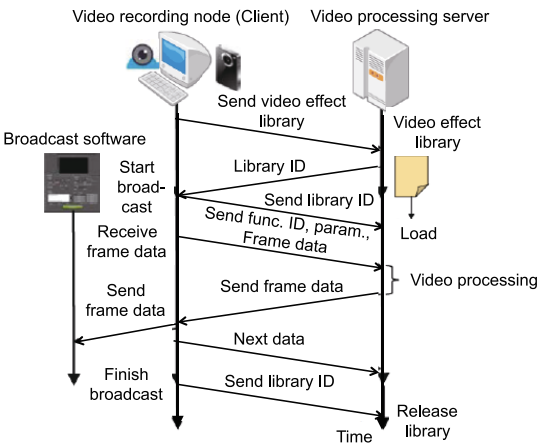


Fig. 1 Flow to process and broadcast a video

videos to broadcasting services. Figure 2 shows a screentshot of the recording node software.

Recently, a cloud computing environment becomes popular. Therefore, our assumptions are realistic that providers of Internet live broadcasting services manage the high performance computers.

2.2 Video Recording Node

Each video recording node has computer programs as a library to add various video effects. The recording node sends the library to the video processing server before sending video frame data because communications traffic is increased by sending the library. Each library and its including functions have IDs and specified by the IDs. Each function also has parameters to be executed.

In the Internet live broadcasting, frame merging methods have been proposed to reduce communications traffic between the sender and the receiver. However, the communication delay becomes longer because the sender waits to collect several frames. If the delay becomes longer, it becomes more difficult to add video effects. Therefore, we assume that a recording node sends each video frame to a video processing server one by one. When the recording node requests the video processing server to add video effects, the recording node sends library IDs and parameters to the video processing server.

After a recording node sends data of a video frame and receives the processed data, the recording node broadcasts the video frame by broadcasting software and sends the next video frame data to the video processing server. After the recording node finishes broadcasting, the recording node announces it to the video processing server.

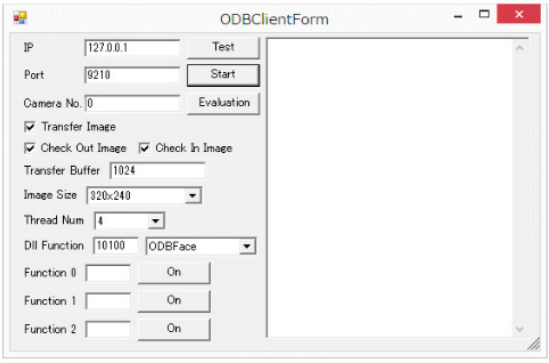


Fig. 2 Screentshot of the recording node software

2.3 Video Processing Server

In most Internet live broadcasting services, users login to the services for user identification before their broadcasting. If a video processing server directly broadcasts videos on the Internet, the server needs to keep users' login information for the broadcasting services. Keeping users' login information has problems in term of security. Therefore, we assumes that the video processing server returns the processed video frame data to the requesting recording node, and the recording node broadcasts the video. In the case of no user authentications, the video processing server can directly broadcast the processed video.

When a video processing server receives frame data from a recording node, the video processing server adds video effects and returns the frame data to the requesting recording node. After the recording node finishes broadcasting, the video processing server releases the library to process the video.

3 Proposal and Implementation of a Distributed Video Processing System

In this section, we explain our proposed system and its implementation.

3.1 Overview

If a huge number of recording nodes requests the same server to process videos, users cannot smoothly broadcast videos because the video processing server has a heavy load and takes long time to process the requested tasks. Therefore, we propose and implement a load distribution method that selects a specific processing server from several servers through the Internet. Figure 3 shows our proposed model of the distributed video process system. Video processing servers construct a P2P network to distribute the loads for video processing.

Figure 4 shows internal structures of recording nodes and video processing servers. Recording nodes have client software and a PIAX process. In addition, video processing servers have server software and a PIAX process. The PIAX process on each recording node searchs a specific server from several video processing servers on the PIAX overlay network. PIAX is platform middleware for P2P agent systems. PIAX realizes "Max Less Than" search by Skip Graphs [3]. "Max Less Than" search is to search the maximum element less than the specific key. For example, if elements are 1, 6, 15, 24 and the specified key is 10, the element 6 is searched as the result.

Our proposed method autonomously selects one server from several video processing servers by "Max Less Than" search. Figure 5 shows a sequence to request

video processing, and Figure 6 shows a schematic diagram to select a video processing server. First, PIAX processes on each video processing server make a random value at the time of starting up. Those values are shared on P2P networks as keys of video processing servers. The video processing servers on P2P networks are sorted by the keys, and recording nodes execute "Max Less Than" search for a video processing server with a random key. In the case of Figure 6, the PIAX process on the recording node generates 7243 as a key and executes "Max Less Than" search. By searching on the PIAX network, the video processing server that has the key 5021 is selected.

Our proposed method selects the video processing server at random to make the loads of servers probabilistically equal. If the system is required to make the loads exactly equal, we can adopt the method that each server measures their own loads and shares the information to select the low-load servers preferentially.

3.2 Avoid Overloading of Video Processing Servers

In our proposed method, heavy load servers temporarily leave from the system to avoid increasing the load by new requests. Figure 7 shows a sequence to request video processing with a heavy load server. First, the PIAX process on heavy load servers leaves from the PIAX network by removing its own key. By removing the key from a key space, its video processing server is not searched for new requests, and the load of the video processing server is not increased from the current value. If the load of the video processing server reduces to the permitted value, the PIAX process on the video processing server joins again to the system sharing its own key and accepts new requests from recording nodes.

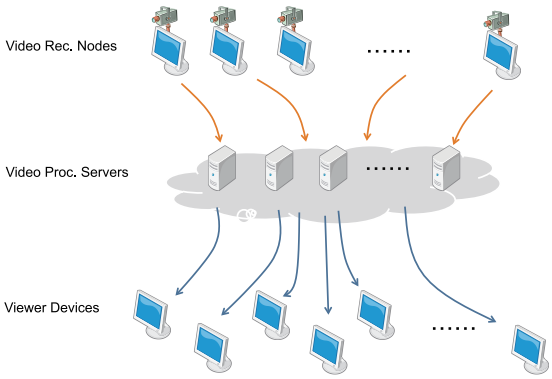


Fig. 3 System model

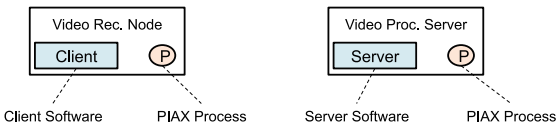


Fig. 4 Internal structures of recording nodes and video processing servers

By the above procedures, our proposed method autonomously controls video processing servers against new requests from recording nodes and avoids overloading of video processing servers. In addition, if there are no available servers on the PIAX network, recording nodes fail to connect to video processing servers. By failing to connect servers, users or service providers understand the lack of computing resources against the current demands and need to expand the system.

As similar techniques to our proposed method, load distribution by a reverse proxy is widely used for Web servers. In our proposed system, each server collaborate autonomously based on P2P technologies and avoid generating a bottleneck server such as a controller or surveillance server.

4 Experiment

In this section, we explain the experiment of our proposed system using real devices.

4.1 Environment

We use two computers as servers A and B. The server A has a dual-core CPU, and the server B has a quad-core CPU. The number of video recording nodes is three. Each recording node has a USB camera and sends pictures to servers. The client 1

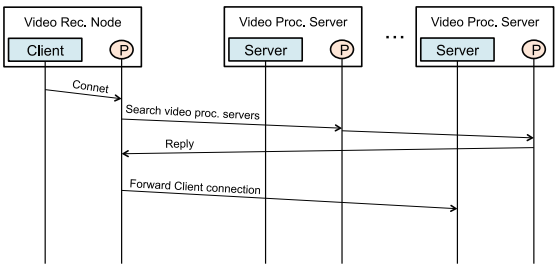


Fig. 5 Sequence to request video processing

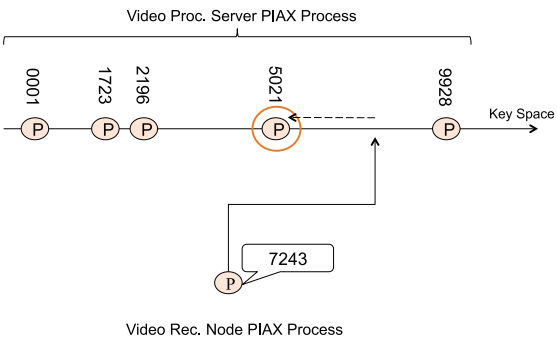


Fig. 6 Selection of a video processing server

uses the server A, and the client 2 uses the server B. In order to measure the loads to use each server, the client 3 uses the server B after using the server A. Those devices connect to the same wireless network.

Video processing systems can realize a high frame rate by reducing the time to process each frame. Therefore, we define the turnaround time as a load. The turnaround time means the time between sending one frame and finishing processing it.

4.2 Results

Figure 8 shows the experimental result. The horizontal axis shows the number of frames sent from the client A or B. The vertical axis shows the turnaround time. The client 1 uses server A, and the client 2 uses the server B. The client 3 uses the server A between the frame 50 and 150. After standing by between the frame 150 and 200, the client 3 uses the server B between the frame 200 and 300.

The turnaround time of the client 1 is about 50 ms before sending the frame 50. After client 3 uses the server A on the frame 50, the loads of server A increases, and the turnaround time of the clietn 1 becomes about 180 ms. Between the same period, also the turnaround time of the client 3 is about 180 ms. In addition, the turnaround time of the client 2 is about 32 ms before sending the frame 200. After the client 3

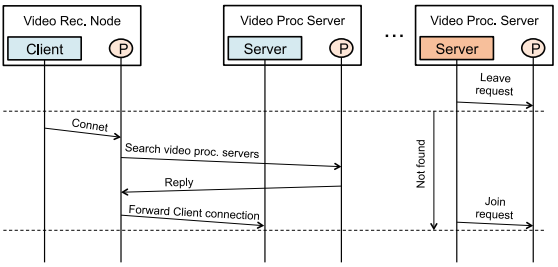


Fig. 7 Sequence to request video processing with a heavy load server

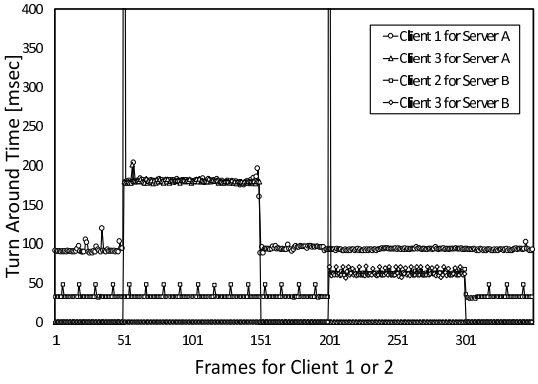


Fig. 8 Turnaround time

uses the server B on the frame 200, the loads of the server B increases, and the turnaround time of the clietn 2 becomes about 65 ms. Considering these results, the client 3 can reduce the turnaround time by using the server B if the server A recognizes own loads and stops to accept new requests based on our proposed method. Although the turnaround time of the client 2 increases to nearly 32 ms, the increased amount is less than the case of the client 1 and the server A, about 90 ms.

The turnaround time of the client 1 or 2 increases when the client 3 connects to the server A or B. These are because of the time of connection processes for the client 3. In addition, the turnaround time of the client 2 periodically increases to nearly 50 ms. This is because of the time to retrieve pictures from the camera and affected by the performances of devices.

5 Related Work

Distributed video processing systems have been researched to process videos on different devices from those recording devices. MediaPaaS purposes to realize a cloud computing service for live broadcasting based on PaaS (Platform as a Service) [4]. In MediaPaaS, computers on the service provider encode, decode and process videos. However, users in MediaPaaS can add only the video effects that show always such as a clock and a logo. Therefore, the users cannot add video effects in a specific time. The P2P-based method has been proposed to encode videos based on users' communication states [5]. However, the users in this research cannot specific the time to encode videos. PictuAR[2] provides a service to search related videos based on uploaded images from users. Our proposed system is different from PictuAR to treat the requested processes with various libraries from users.

Also video delivery systems have been researched for Internet live broadcasting. The processed videos from panoramic videos are delivered such as being zoomed and panned [6]. In addition, a function of chasing playback are provided on P2P-based live broadcasting [7]. Our proposed system is different also from these researches to treat the requested processes with various libraries from users.

In Internet live broadcasting, several methods have been proposed to reduce the delay to deliver videos. SmoothCache 2.0 reduces the communication loads of devices and the delay by caching videos on a P2P network [8]. The method on a P2P network has been proposed to select the delivery path that has the theoretical shortest delay [9]. HD method reduces communications traffic by network broadcasting not only unicast [10]. Although our proposed system can apply these delay reducing techniques, our prosed system has different target.

Several systems have been proposed to process the recorded videos. The system has been proposed to forward videos to high performance computers [11]. In addition, the system has been proposed to record videos on external storages directly

[2] PictuAR (in Japanese) - http://www.nttcom.co.jp/ar-saas/

such as cloud [12]. These systems target the recorded videos and cannot be applied to Internet live broadcasting.

6 Conclusion

In this paper, we propose and implement a method to distribute loads on a video processing system for Internet live broadcasting. Our proposed method assumes that high-performance computers on the Internet and users request the computers to add video effects. Our proposed method selects a specific computer from several computers to process each video based on a P2P mobile agent system. We confirmed in our experimental results that our proposed method can reduce the time to process and broadcast videos.

In the future, we will study a technique to measure and share the loads of video processing servers to select the low-load servers preferentially.

Acknowledgements A process allocation scheme in this research was supported by JSPS KAKENHI Grant-in-Aid for Scientific Research (B) numbered 15H02702. A P2P network construction technique in this research was supported by the collaborative research of National Institute of Information and Communications Technology (NICT) and Osaka University (Research on high functional network platform technology for largescale distributed computing).

References

1. I. Sodagar, IEEE MultiMedia **18**(4), 62 (2011)
2. M. Yoshida, T. Okuda, Y. Teranishi, K. Harumoto, S. Shimojo, IPSJ Journal (in Japanese) **49**(1), 402 (2008)
3. J. Aspnes, G. Shah, ACM Transactions on Algorithms (TALG) **3**(4 (37)) (2007)
4. B. Cheng, in *Proceedings of the 7th IEEE International Conference on Cloud Computing (CLOUD 2014)* (2014), pp. 713–720
5. J.C. Wu, P. Huang, J.J. Yao, H.H. Chen, IEEE Transactions on Circuits and Systems for Video Technology pp. 220–224 (2011)
6. V.R. Gaddam, R. Langseth, H.K. Stensland, P. Gurdjos, V. Charvillat, C. Griwodz, D. Johansen, P. Halvorsen, in *Proceedings of the 5th ACM Multimedia Systems Conference (MMSys 2014)* (2014), pp. 168–171
7. Y. Gotoh, T. Yoshihisa, H. Taniguchi, M. Kanazawa, Journal of Networking Technology **2**(4), 169 (2011)
8. R. Roverso, R. Reale, S. El-Ansary, S. Haridi, in *Proceedings of the 6th ACM Multimedia Systems Conference (MMSys 2015)* (2015), pp. 61–72
9. J. Dai, Z. Chang, G.S.H. Chan, in *Proceedings of the IEEE International Conference on Communications (ICC 2015)* (2015), pp. 6959–6964
10. T. Yoshihisa, S. Nishio, IEEE Transactions on Broadcasting **59**(1), 62 (2013)
11. D. Gibbon, L. Begaja, IEEE ComSoc MMTC E-Letter **9**(3), 29 (2014)
12. W.C. Ting, K.H. Lu, C.W. Lo, S.H. Chang, P.C. Liu, in *Proceedings of the 2014 IEEE International Conference on Internet of Things (iThings 2014)* (2014), pp. 169–176

Part IV
Workshop MWVRTA-2016: The 6th International Workshop on Multimedia, Web and Virtual Reality Technologies

Numerical Simulation of Impact Acceleration on the Key Parts of Packing Ammunition on Condition of Dropping

Xue Qing
Department of Information Engineering,
Engineering University of PAP,
WuJing Road #1, WeiYang Section,710086, Xi'an, Shaanxi, China;
455748966 @qq.com

Abstract. In order to evaluate ammunition safety on condition of dropping exactly, choose reasonable parameters for ammunition and package design, and guide ammunition service management scientifically, the thesis defined fuze as a pivotal element. Then the dropping process of package ammunition was emulated on computer by use of the finite element analysis software, and the impact acceleration on the fuse was simulated, the reliability of the model was also validated by the test.

Keywords. Dropping impact; Acceleration; Fuse; Finite element analysis; Numerical simulation

1 Introduction

Ammunition has to suffer impact effect inevitably when they're conveyed or transported. A larger impact effect makes some ammunition arm and evokes security accident. Elements inside the ammunition which are sensitive to impact circumstance are research emphases. Defining the elements as pivotal elements can both predigest questions and make the research and analysis abound in pertinence. The thesis took fuse as pivotal elements, and studied on impact acceleration of fuse when dropping accident of packing ammunition happened.

At present, there are two methods of calculating impact acceleration of packing ammunition on condition of dropping. Firstly, packing ammunition is simplified to spring-mass model of double degrees of freedom. Then we can solve impact acceleration by considering the binary quadratic differential equations [1]. But the limitation of this method is neglecting the effect of elasticity and material, and not considering the structure of the package, and also the parameter of spring-mass model isn't determined easily. So the application is limited. Another method is measuring the impact acceleration by sensor. Although its value has a high reliability, the cost of test is high too. And drop gesture of parking is also controlled difficultly. So does the accidental drop. Therefore it's very necessary to find an ideal method of calculation. According to the theory of finite element, the thesis applied computer simulation

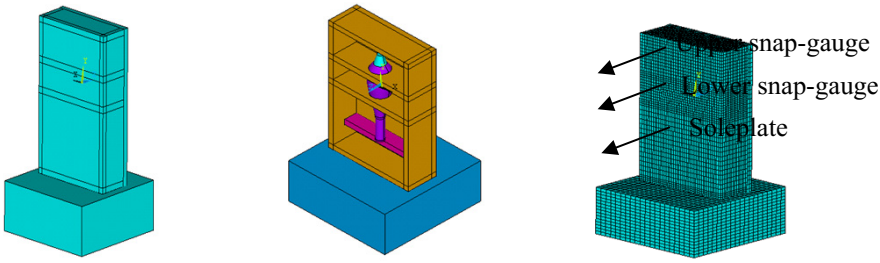
© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_60

method to establish the mathematical model of packing ammunition on condition of dropping, and research the changeable situation of fuze of impact acceleration.

2 Description of finite element model

There are two kinds of package form of fuse primarily. First is that fuses are packed in the plastic box individually. Second is that fuses are installed in the bullet, packed with ammunition. The second package form is chiefly used for early ammunition, and the safety of that ammunition is also poorer. Once an accident happens, the consequence could be very serious.



(a) External structure of pack (b) Internal structure of pack (c) Gridded model

Fig.1. Finite element model of packing ammunition

The thesis researched the second package form. Software of PRO/E was applied to establish the entity model of packing ammunition (Fig. 1-a), and its internal structure was shown on fig. 1-b. The size of package of ammunition is 437mm×295mm×98mm, and the cabinet is 18mm thick, and snap-gauge is 20mm thick. There were three rounds of a certain type of mortar shells installed in the package. In order to save up computation time, the tail part and obturating ring of the mortar shells were simplified, and equivalent mass of another two shells was exerted to the package, and shells had a common node with lower snap-gauge. Then the model was imported into ANSYS and meshed, and 63607 units and 83023 nodes formed (Fig. 1-c). When they dropped vertically, the shells were fixed by lower snap-gauge and soleplate, and didn't rock. Package was wooden structure, and shell was made of ductile cast iron, and fuze body was phenolic glass fiber plastic, and ground was made of cast iron plate which corresponded with verification test. There were mechanical properties shown on table 1.

Timber has some especial mechanical properties, mainly as elasticity at low load, and plastic at high load, and the behavior of creep and relaxation similar with metal under certain conditions [3-4]. Based on above, the thesis regarded wooden package and snap-gauge as ideal elastic-plastic material. Same with ductile cast iron and gray cast iron, it also chose model of bilinear kinematic. Phenolic glass fiber plastic expressed as elastic deformation under impact. Based on the parameters, model chose linear elastic material. The contact between package and ground, between shell and upper snap-gauge, between collembolan and soleplate used automatic general contact algorithm (AG), and numerical model chose system of units of mm-kg-s.

Table 1. Mechanical properties [2]

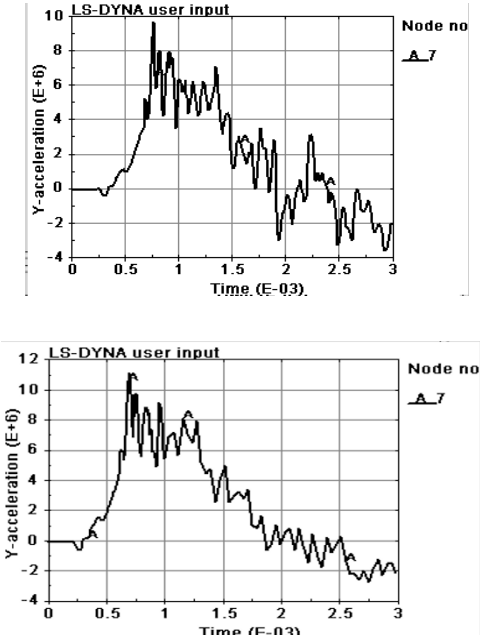
Material type	Density /kg·mm-3	Modulus of elasticity /KPa	Poisson's ratio	Yield modulus /KPa	Tangent modulus /KPa
Red pine	4.8E-7	1.02E7	0.49	3.1E4	5.5E5
FRP	1.6E-6	4E7	0.15	—	—
Ductile cast iron	7E-6	1.68E8	0.24	3.65E5	6.74E7
Gray cast iron	7.25E-6	1.1E8	0.24	1.8E5	4.8E7

3 Simulations

When package ammunition dropped, impact acceleration and impact duration of location on fuse component were relevant to package structure, as well as rigidity of ground, dropping initial velocity, dropping height, dropping gesture. The thesis simulated that the package dropped vertically on the cast iron plate with fuse headed up. According to the figuration of fuse of straight-line inertia insurance type, this gesture relieved insurance most easily [5]. Then the thesis chose two dropping heights of 500m and 1000m respectively to simulate. In order to save up computing time, package ammunition was given equivalent dropping speed when it impacted the cast iron plate according to formula of falling freely [6]. By the curve, when drop height was 500mm, the peak of acceleration was 9.5E6mm/s2 and duration of impact was 1.4ms, and when drop height was 1000mm, the peak of acceleration was 11E6mm/s2 and duration of impact was 1.75ms.

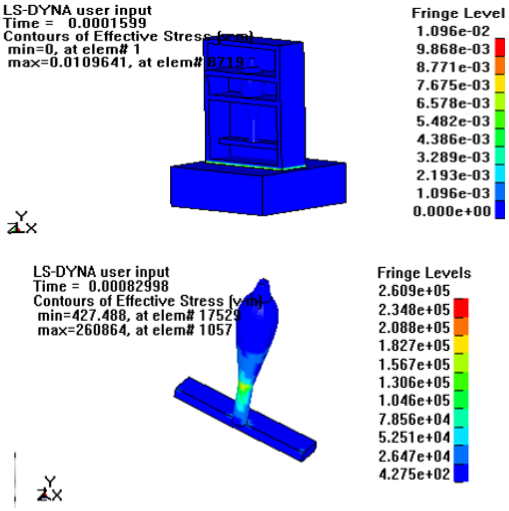
Through the post-processor of LS-PREPOST, we got the stress cloud picture of different time in the process of package ammunition dropping. Fig. 3 showed us the stress cloud picture when package contacted with ground and ammunition with lower snap-gauge on the height of 500mm.

As is known in the fig. 3-a, there wasn't impact reach to the fuse component when package contacted with ground in 0.15ms. And in the fig. 2-a, when impact effected reach to the fuse component in 0.3ms, acceleration augmented gradually to the peak until 0.75ms. Through the fig. 3-b, lower snap-gauge contacted with ammunition tail section in 0.8ms, and impact of mortar shell and lower snap-gauge effected on the fuse component. By the fig. 2-a, the impact of mortar shell and lower snap-gauge made the curse of impact acceleration have some disturbance in 0.8ms.



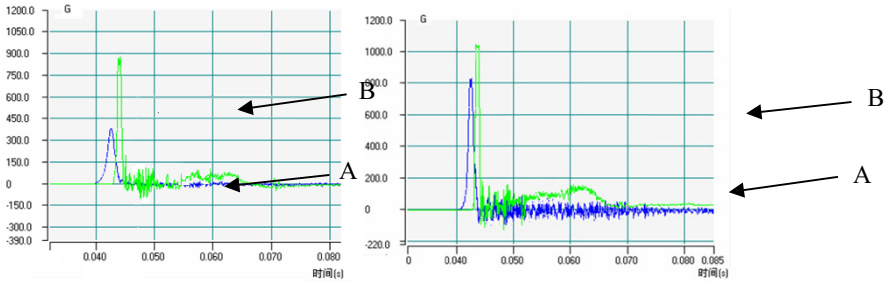
(a) Drop height of 500mm (b) Drop height of 1000mm

Fig.2. Impact acceleration curse of fuse component



(a) Package contacted with Cast iron plate (b) Ammunition tail contacted with soleplate

Fig.3. Stress cloud picture of different time on condition of ammunition dropping



(a) Drop height of 500mm (b) Drop height of 1000mm

Fig.4. Impact acceleration curse of dropping

4 Dropping test

To verify the reliability of simulation model, the thesis administrated the dropping test of package ammunition according to the 29th section of < GJB5389.29-2005 Test method of artillery missiles >: 1.5m dropping test of package and <GJB573.14-88 Environment and performance test method for fuse>: 1.5m dropping test. Then the thesis made package installing with ammunition drop freely on the cast iron plate by the most rigorous gesture (fuse headed up), and tested the curse of impact acceleration on the fuse component.

Fig. 4 showed the impact acceleration of dropping height for 500mm and 1000mm. Curse A represented the impact acceleration of wooden package, and B was impact acceleration curse of fuse component. As is known from curse A and B, duration time of impact on fuse component had a minor difference between 1ms and 2ms, and higher the dropping height, larger the acceleration peak of impact on package and fuse component. Because of cushion effect, acceleration peak on fuse component appeared latter than peak of package. Comparison between test peak and simulation value of impact acceleration on fuse component was shown in table 2. Through the comparison, we found that difference between test peak and simulation value was not large, and the simulation model established by the thesis was reliable.

Table 2. Comparison between test value and simulation value

Height /mm	Test value /106mm • s-2	Simulation value /106mm • s-2	Simulation error
500	8.7	9.5	9.2%
1000	10.3	11	6.8%

5 Conclusions

The thesis took fuse component as pivotal part. According to finite element theory, the thesis chose two kinds of 500mm and 1000mm height through the establishment of the simulation model, and analyzed simulative on package ammunition of dropping, and gained the impact acceleration value on fuse component. Through the comparison of test value, it verified the credibility of simulation model that the difference between simulation value and test value was not large. The thesis also adopted computer simulation technology to simulate the impact acceleration on the pivotal part on the condition that package ammunition dropped, and evaluated the impact situation on the pivotal part of ammunition with quantitative indicators. There's certain guidance significance for safety assessment of ammunition dropping.

References

1. Li Jin-ming, An Zhen-tao, Ding Yu-kui. Study of Ammo Packaging Drop Impact Response[J]. Packaging Engineering, 2006, vol. 27(4), pp. 137-138.
2. Grid. Pizza ke. Mechanics of materials manual[M]. Hebei: Hebei People's Publishing Company, 1981.
3. Yin Si-ci. Wood Science[M]. Beijing: China Forestry Publishing Company, 1996.
4. Tao Jun-lin, Jiang Ping, Yu Zuo-sheng. Research in Constitutive Relations of Wood Static Pressure Large Deformation[J]. Mechanics and Practice, 2000, vol. 22(5), pp. 25-27.
5. Writing Group of Fuse Design Manual. Fuse Design Manual[M]. Beijing: National Defence Industry Press, 1978, pp. 174.
6. Hua Li. Dropping Simulation Analysis of the Transport Package for Washing Machine Based on Ansys/Is-dyna[J]. Mechatronics, 2008, vol. 02, pp. 71-73.
7. Zhai Yuwen, Xu Hongliang. Design of Practical Multi-functional Electronic Timekeeping Watch. Journals of Jilin Chemi-industry College, 2001(1): 31-33.
8. Zhong Hanmei. Brief Introduction of Lingyang Single-chip Machine. Overseas Electronic Elements, 2002(9): 69-72
9. Luo Yafei. Application Basis of Lingyang 16-digit Single-chip Machine[M]. Beijing: Beijing University of Aeronautics and Astronautics Press, 2003
10. Zhai Yuwen, Xu Hongliang. Design of Practical Multi-functional Electronic Timekeeping Watch. Journals of Jilin Chemi-industry College, 2001(1): 31-33.
11. Zhong Hanmei. Brief Introduction of Lingyang Single-chip Machine. Overseas Electronic Elements, 2002(9): 69-72
12. Luo Yafei. Application Basis of Lingyang 16-digit Single-chip Machine[M]. Beijing: Beijing University of Aeronautics and Astronautics Press, 2003
13. Ma Zhongmei. C-language Application Programming Design -3 Version of Single-chip Machine. Press of Beijing University of Astronautics and Aeronautics , Nov 2003.
14. Hu Hancai. Principle of Singlke-chip Machine and its Interface Technology. Press of Tsinghua University, Feb 2004.
15. Wu Han. The graduate student information management system based on VC + + [J]. The design and implementation of computer technology and development, 2006, 16 (12) : 184-186.

Instruction Sequences Clustering and Analysis of Network Protocol's Dormant Behaviors

Yan-Jing Hu[1, 2]

1Key Laboratory of Cryptology & Information Security Under the Chinese PLA
Engineering University of the Armed Police Force
Xi'an 710086, China
2State Key Laboratory of Integrated Services Networks
Xidian University
Xi'an 710071, China
Huyanjing2007@126.com

Abstract. Protocol's dormant behavior are the behaviors that cannot be observed during dynamic analysis, they can only be triggered under particular conditions. In the last few years, several approaches have been proposed to detect anomalous behaviors in malwares, but protocol's dormant behaviors are rare studied due to the large invisibility and variability. This paper presents a novel approach to the analysis of protocol's instruction sequences that uses a novel instruction sequences clustering technique to mine dormant behaviors in unknown protocols. Our technique provides a more precise solutions to mine unknown protocol's dormant behaviors, and in addition, it is able to analyze encryption protocols and mine the behavior instruction sequences protected by obfuscation technologies. Experimental results show that the present method can mining the protocol's dormant behaviors accurately and effectively.

Keywords—***protocol reverse; protocol's behavior; protocol's dormant behaviors; instruction sequences***

1 Introduction

In this paper we develop techniques for mining and analyzing security-relevant dormant behaviors hidden in unknown protocols. Protocol reverse technique we invited is a very valuable and recurring fundamental problem in many fields, such as deep packet inspection[1, 2], botnet analysis[3], vulnerability discovery[4]and signature generation[5]. Most protocol reverse works focus on inferring the grammar of undocumented or unknown protocols, such as message format and fields[6, 7], this information is important for applications like network monitoring, signature generation, or botnet infiltration[8-10]. We develop a new approach to protocol behavior reverse based on instruction sequences clustering analysis. A lot of case studies show that the protocol behavior analysis is more important, because a protocol's behavior, especially its dormant behavior affects the protocol's execution security, which is directly related to the foundation of network security.

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_61

Due to the lack of detailed understanding of an unknown protocol's specification, protocol reverse analysis may be the best way to study its dormant behavior. Dormant behavior is usually hidden or disguised in protocol binary, therefore such important information is suitable to be reversed from program binary that implements the protocol. While this work is not trivial because the dormant behaviors we are confronted with are varied and usually difficult to mine. For some protocols, the malicious functions, special modules or key code segments are encrypted, disguised or control flow obfuscated. For other protocols, the malicious behaviors are embedded into normal behaviors, and they could be triggered only under specific conditions. Traditional network security technologies have less effective for dormant behavior mining, even worse, they may affect the normal communication, because a concealed malicious behavior does not replicate or spread, and even has no significant malicious characteristic. The growing invisibility, robustness and survivability of a dormant behavior make the traditional analysis, tracking and identification more difficult. How to grasp a general analysis method which can mine and explore the unknown protocol's dormant behavior quickly and accurately is becoming a new challenge for network security. Mining a protocol's dormant behavior is a key problem that could not be avoided for protocol reverse analysis.

Recently binary analysis plays a vital role in combating the rapidly growing unknown protocols[4, 5, 11]. Most existing analysis algorithms operate on either static message features or dynamic behavior features[5, 12, 13]. However, these two distinct approaches have their own strengths and weaknesses in handling various types of protocol dormant behaviors. Although the ideas of combining static and dynamic have been mentioned in both industry and academia[4, 14], very few ideas have addressed their systematic integration. In this paper, we propose a novel method of integrating dynamic and instruction clustering analysis to mine an unknown protocol's dormant behavior regardless the types and hidden techniques the protocol uses. It just relies on the protocol's binary raw data (including both binary code and protocol message) to quickly cluster protocol behaviors and to finally explore dormant behaviors.

In this paper, resting on the analysis of unknown protocol binary process message, we develop a virtual analysis platform HiddenDisc (Hidden Behavior Discovery) for automatically grouping protocol behaviors into families that share some common instruction level characteristics by using a instruction clustering solution, the proposed solution is an integration of dynamic binary and static clustering analysis.

To summarize, the major contributions of this paper are as follows:

-Mining dormant behavior quickly and accurately. We get public behavior instruction sequences by dynamic analyzing the protocol program parsing network message. With the help of public behavior instruction sequences, we use instruction clustering algorithm to categorize the protocol behaviors, then mine the dormant behaviors and trigger conditions.

-Unifying dynamic binary analysis and static clustering analysis, which could resist various encryption and obfuscation transform. Our proposed framework HiddenDisc first systematically uses a composition of dynamic binary analysis and static instruction clustering analysis regardless of the encryption and obfuscation ways the protocol used, achieving much better results than any single approach alone.

-Propose a novel metric of Protocol Execution Security and evaluate the unknown protocol's execution security in a concise manner.

The remainder of the paper is organized as follows. In Section 2 we describe the system overview as well as the basic idea and approach. We present the system design and key techniques for mining and analyzing unknown protocol's hidden behavior in Section 3. The experiment and evaluation are demonstrated in Section 4. Finally, we conclude this paper and present future work in Section5.

2 System Overview

Protocol's dormant behavior can be hidden in anywhere or encrypted in any type. The single static analysis is difficult to identify these special deformed instruction sequences and control flows, while dynamic analysis is also difficult to capture dormant behaviors in a single run, and completely cover all the instruction sequences is almost impossible. The proposed hybrid method is regardless of how the dormant behavior instruction sequences hide and deformed, they have a self-identification and self-conversion process, and the dormant behavior instruction sequences are completely different from the public behavior instruction sequences. So starting from the public behavior instruction sequences, using instruction clustering algorithm, we can identify and locate the dormant behavior instruction sequences and the trigger conditions.

To compensate the deficiencies of protocol reverse analysis in dormant behavior mining, and integrate the advantage of dynamic binary analysis and static clustering analysis, we design and implement a virtual analysis platform, protocol dormant behavior mining prototype system - *HiddenDisc*. This prototype system can be used to reverse analyze unknown protocol's dormant behavior regardless of the deformation mode, capturing the protocols' dormant behavior instruction sequences, triggering the execution and obtain more complete information about the protocol's behavior. Based on the analysis of most protocols' execution process, the proposed analysis architecture *HiddenDisc* is shown in Figure 1.

HiddenDisc is designed to automatically reverse analyze unknown protocol's dormant behavior. It marks all the message data as tainted bytes, and then tracks the propagations of all the tainted bytes when the program binary executes. After the raw data processing, all the instruction sequences relevant to the tainted data processing are recorded. Based on such dynamic binary analysis technology, basic

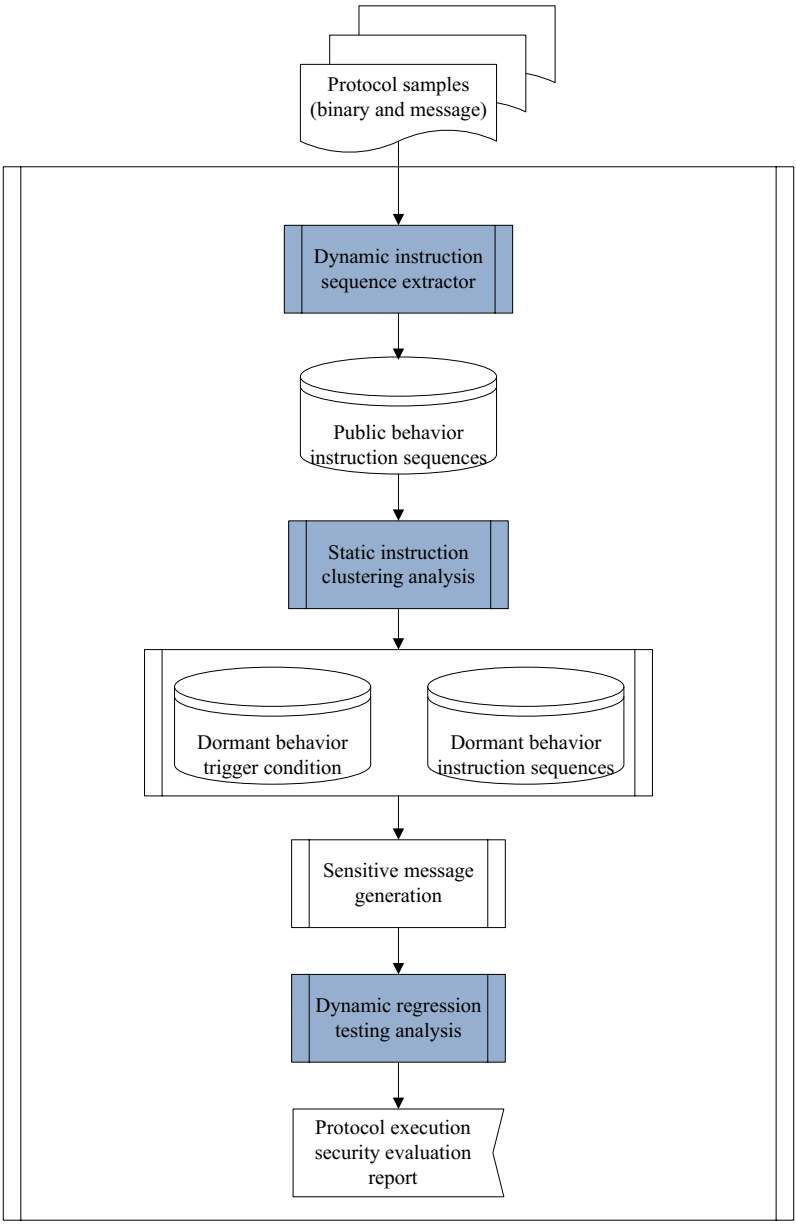


Fig. 1. System architecture of HiddenDisc

block based public behavior instruction sequences are obtained. Based on the recorded public behavior instruction sequences, protocol's dormant behavior and trigger conditions are perceived and mined by the instruction clustering analysis algorithm. Then the dormant behaviors are triggered, monitored and analyzed, and finally the protocol's execution security is evaluated.

Our main idea is capturing public behavior by dynamic binary analysis, mining dormant behavior by instruction clustering, generating trigger condition as test case and confirm dormant behavior as validation execution. Protocol's public behavior can be captured in the dynamic message parsing process[15]. Protocol's dormant behavior differs from its public behavior due to the significant characteristics on Machine-Instruction Level, so the dormant instructions can be explored by clustering analysis. The dormant behavior often uses encryption, obfuscation and other protection measures and has a self-discharge process. According to the dormant behavior trigger conditions and other obtained information, the test case to cover the dormant behavior instruction sequences can be generated[16-18], and then the dormant behavior is triggered and analyzed as validation execution. The intent of dormant behavior may be varied. The attacker can get confidential information through protocol's dormant behavior, even subtle control target host. Our main motivation is to find out the principles of dormant behavior, research the methods and techniques to prevent the protocol's dormant behavior.

3 Instruction sequences clustering for network protocol's dormant behavior

3.1 Protocol dormant behavior analysis steps

The protocol TestBot is composed of C(protocol program) and M(protocol message). The protocol program can be seen as a set of all the instruction sequences, including public and dormant behavior instruction sequences. All the disguised and obfuscated technologies are used to protect dormant behavior instruction sequences from being discovered. Case studies show that dormant behavior may seriously disrupt protocol's execution security. The proposed work is to mine and analyze all the dormant behaviors. The research can be divided into three main steps:

(1) Monitor the protocol program processing messages, and dynamically build the basic block-based instruction sequences. The protocol messages are captured by "RADCOM" hardware and "Wireshark" software Protocol Analyzer respectively. The active raw traffic data are reverse analyzed on the binary analysis platform. Each byte of data is marked as tainted data, and the tainted data can be tracked during the protocol's whole execution. All compare instructions related to the tainted data are recorded, including the jump(conditional jump) and the function call related instructions (call addresses) that can characterize the protocol's behavior. The analysis platform can not only mark the executed instruction sequences, but

also can infer the instruction sequences that have not yet been executed, that means the protocol contains potential dormant behavior.

(2) Statically analyze the instruction sequences by instruction clustering algorithm., The instructions are labeled with three types of genetic instructions described in detail in 4.3, and these labeled instruction sequences are converted into n-grams, using instruction clustering method, we can collect n-gram features on the instruction sequences and construct several behavior clusters.

(3) Generate new protocol message with the "sensitive information" as the trigger of validation execution. by monitoring the protocol program parses the new message, analyzing all instruction sequences that have never been executed before, comparing with the clustering result to verify the accuracy and completeness, ultimately, we can recognize the nature of dormant behavior instruction sequences. A collection of dormant behavior instruction sequences is the static description of the protocol's dormant behavior.

The potential dormant behavior can be mined by calculates the behavior distance between captured public behavior instruction sequences and each other behavior instruction sequences. Suppose behavior A and B are two public behaviors we captured, and we know their functions. We have a protocol sample P with unknown behaviors. In general, we compute two behavior patterns P_A and P_B to distinguish between behavior A and B, corresponding to the behavior instruction sequences A and B, respectively. To mine the dormant behavior of protocol P, we check the characteristics of its instruction sequences against both known patterns, and see which one it most closely matches. We assert the mined behavior P_i closely match behavior pattern P_A, i.e., d (P_i, P_A) =0, where d is the behavior distance function. This means that behavior P_i is very close to the known behavior A; they can be attributed to a same cluster. If the known behavior A is normal, the behavior P_i can also be a normal behavior. By the same token, if P_i is very close to the known behavior B, it can be clustered to behavior pattern P_B. Otherwise if the n-gram instruction characteristics of behavior P_i is too distant from all the behavior patterns, we may assert that P_i is a new dormant behavior pattern, and generate a new cluster.

Therefore, we can start from the public behavior instruction sequences, and then we search the associated dormant behavior instruction sequences using clustering algorithm.

3.2 Hybrid analysis for network protocol's dormant behavior

Protocol’s dormant behavior can be hidden in anywhere or encrypted in any type. The single static analysis is difficult to identify these special deformed instruction sequences and control flows, while dynamic analysis is also difficult to capture dormant behaviors in a single run, and completely cover all the instruction sequences is almost impossible. The proposed hybrid method is regardless of how the dormant behavior instruction sequences hide and deformed, they have a self-

identification and self-conversion process, and the dormant behavior instruction sequences are completely different from the public behavior instruction sequences. So starting from the public behavior instruction sequences, using instruction clustering algorithm, we can identify and locate the dormant behavior instruction sequences and the trigger conditions.

To compensate the deficiencies of protocol reverse analysis in dormant behavior mining, and integrate the advantage of dynamic binary analysis and static clustering analysis, we design and implement a virtual analysis platform, protocol dormant behavior mining prototype system - *HiddenDisc*. This prototype system can be used to reverse analyze unknown protocol's dormant behavior regardless of the deformation mode, capturing the protocols' dormant behavior instruction sequences, triggering the execution and obtain more complete information about the protocol's behavior.

We integrate Dynamic Taint Propagation Analysis technology into HiddenDisc analysis platform, each byte of the protocol message is marked as the taint source. Each instruction of the protocol program parses protocol message is recorded. With these instruction sequences and observe the propagation of taint source in register and memory, the public behavior of the protocol can be analyzed. In view of the research is mature and rich, we only focus on mining dormant behavior in this section.

The dormant paths are represented by basic blocks painted in dashed boxes, example is shown in Figure 2.

According to algorithm 1, we get the basic block-based representation of the behavior instruction sequences, generate conditional jump diagram of the protocol behavior. *B1, B2* . . . is the basic block, it may be an instruction, and may also be composed by a number of instructions. Its size depends on the situation of protocol parsing message fields. The rhombic basic block is Conditional Jump Block, its value tends to influence the protocol behavior, and it is the focus of our attention. Conditional jump diagram can intuitively represent the public behavior and dormant behavior as public paths and dormant paths. According to the dormant paths in conditional jump diagram, we can obviously find two dormant behavior instruction sequences, namely $H1 = \{B5, B13, B14\}$, $H2 = \{B5, B13, B15, B16, B17, B18\}$.

Observations show that in the protocol message parsing process there are a lot of compare and conditional branch instructions. The corresponding high-level language representation like structures If (a field) then {parsing or processing} or If (a constant) then {parsing and processing} else {other processing}

The protocol's dormant behavior may be embedded in them. Public behavior instruction sequences are just part of the protocol program's paths. We use doubly linked list dynamically record each node's information, including all conditional jump instructions, compare constants, unexecuted call instructions, and the corresponding address values. Based on this information, we can trace back the execu-

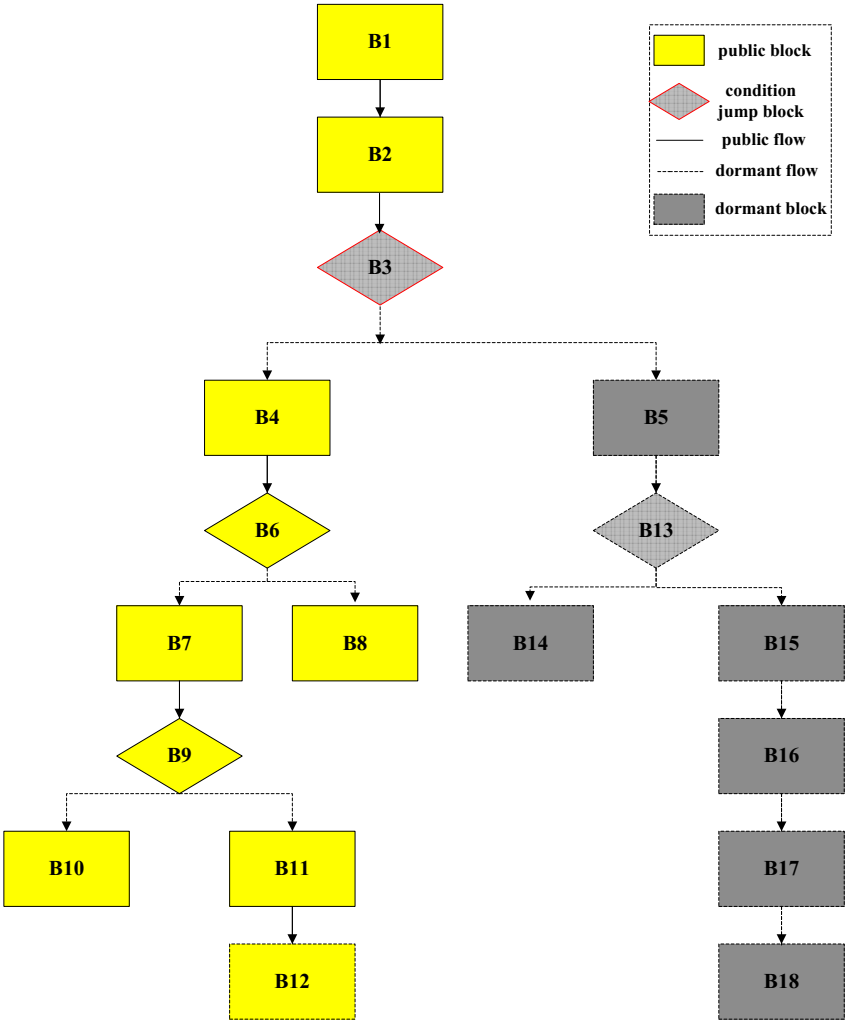


Fig. 2. Basic block based conditional jump diagram for TestBot (C&C) protocol

tion flow, analyze all tainted data and associated code blocks; analyze self-transformation, self-decrypting and other reverse dormant transformation processes. With algorithm 2 we can perform validation execution. The unexecuted instruction sequences may be a normal behavior, and may also be a potential threat, which need to trigger and further analysis. Under the guidance of "sensitive information", new message with trigger condition is generated, which trigger dormant behavior instruction sequences executed as validation execution, so we can observe and analyze protocol's dormant behavior. According to the dormant behavior analysis results, we can finally estimate protocol's execution security.

4 EXPERIMENTATION AND EVALUATION

In order to measure the effectiveness and accuracy, we have analyzed 1297 protocol samples, including 27 known benign samples to mine and classify protocols' behaviors. According to the characteristics of instruction sequences (we focus on examining the frequency and distribution of the three kinds of genetic instructions, it will be detailed described in the discussion section), the 1297 samples are classified into five groups evidently (the five behavior clusters are: web browsing, network communication, downloads, and two unknown dormant behavior clusters), and the dormant behavior can be distinguished significantly. To accurately capture the protocol data source, we use *RADCOM* hardware protocol analyzer to capture and analyze protocol packets on the analyzing server; to heuristic disassembly the protocol's raw byte binary code with disassemble engine Udis86; then clustering the protocol's public and dormant behavior instruction sequences. From the instructions and registers (memory) value, we can dynamically analyze the captured protocol message data in instruction-level. All experiments are implemented on the whole-system simulation platform *HiddenDisc* developed by us independently. *HiddenDisc* is based on *TEMU* virtual platform, which can realistically simulate various hardware and software resources. *HiddenDisc* enhances dynamic taint analysis technique, provides a user-defined-behavior interface (UDB-API). UDB-API facilitates the development of standard plug-ins, customizes and extends the platform's functions. Our platform was extended in two main directions. First, we increase the functions of *behavior instruction sequences clustering*, and we developed a mechanism to trace byte-level instructions. The second extension is related to *dormant behavior mining* and new protocol message generation. Using *RADCOM* hardware protocol analyzer, *Wireshark* software analyzer and 6 connected x86 architecture computers, on which *HiddenDisc* system are deployed. We analyzed 1297 unknown protocol samples including ourselves developed http-hidden protocol, and then discussed the problems we found during the analysis. The protocol binaries have been collected over a period of 2 years via various sources, including honey nets, spam traps, and security researchers. The 6 host machines' hardware configurations are all Intel Core i7 960@3.20GHz CPU, 4GB of memory. The host operating systems and 2 virtual lab environments are all Ubuntu V12.04 Linux; the other 2 virtual operating systems are Windows 7 SP1. One of the 2 servers is a control server and the other is an analysis server, the operating systems are both Windows server 2008.

5 CONCLUSIONS

In this research, a system solution for mining and analyzing protocol dormant behavior is proposed. Different from previous approaches, our solution combines dynamic taint analysis with Instruction Clustering of unknown protocols' instruction sequences to effectively integrate static and dynamic analyses. The dynamic

taint analysis is effective to monitor protocol's run time public behavior and instruction clustering can mine the dormant behavior which has not executed. We extend the function of instruction clustering algorithm to our developed analysis platform HiddenDisc, and then which can find a partitioning of a given protocol samples so that same cluster exhibit similar behavior and dormant behavior in different cluster. Our analysis begins by monitoring each protocol sample in our developed dynamic taint analysis environment HiddenDisc. Then, we extract all the behavior instruction sequences by instruction clustering. After instruction clustering analysis, dormant behavior instruction sequences are mined automatically. Finally, using new generated sensitive message to trigger dormant behavior execute as validation execution for extensive study.

The two proposed algorithms are tested using protocol samples acquired from a large real-life internet. Experimental results show that the approach can effectively analyze large number of unknown protocols without any prior knowledge, and can accurately mine 1297 protocol's dormant behaviors in less than three minutes. We demonstrate that our hybrid solution is more computationally efficient than the most recently proposed solutions and is more resilient to a wide range of attacks such as malicious code attack, encryption, Code Packing and Obfuscation, and various other re-reverse attacks. Our solution also provides the basis for evaluating protocol's execution security. Now we only have a preliminary understanding of protocol's execution security, the relationship between dormant behavior and execution security also needs further research. Future work will focus on improving the accuracy and versatility of the algorithms, and improve the ability to perceive various abnormal behaviors.

Acknowledgments This work was supported by the Natinal Natural Science Foundation of China under Grant no. 61103178 , 61373170, 61402530, 61309022 and 61309008.

References

[1] L. C. Li Xiang-Dong, "A survey on methods of automatic protocol reverse engineering[C]," *Proceedings of the 2011 Seventh International Conference on Computational Intelligence and Security.,* pp. 685-689, 2011.

[2] D. V. B. D. B. Hoang, "Multi-layer security analysis and experimentation of high speed protocol data transfer for GRID," *International Journal of Grid and Utility Computing,* vol. 3, pp. 81-88, 2012.

[3] W. Chang, A. Mohaisen, A. Wang, and S. Chen, "Measuring Botnets in the Wild: Some New Trends," presented at the Proceedings of the 10th ACM Symposium on Information, Computer and Communications Security, Singapore, Republic of Singapore, 2015.

[4] L.-z. G. Ying WANGa, Zhong-xian LIb, Yi-xian YANGa, "Protocol reverse engineering through dynamic and static binary analysis," *The Journal of China Universities of Posts and Telecommunications,* vol. 20, pp. 75–79, 2013.

[5] J. Narayan, S. K. Shukla, and T. C. Clancy, "A Survey of Automatic Protocol Reverse Engineering Tools," *ACM Comput. Surv.,* vol. 48, pp. 1-26, 2015.

[6] J. L. S. Yu, "Position-based automatic reverse engineering of network protocols," *Journal of Network & Computer Applications,* vol. 3, pp. 1070-1077, 2013.

[7] Y. W. Meijian Li, Shangjie Jin, Peidai Xie, "Reverse extraction of protocol model from network applications," *International Journal of Internet Protocol Technology,* vol. 7, pp. 228 - 245, 2013.

[8] D. C. C. Mansour, "Security challenges in the internet of things," *International Journal of Space-Based and Situated Computing,* vol. 5, pp. 141-149, 2015.

[9] S. K. T. G. X. F. B. L. C. Chen, "Optimizing Communication for Multi-Join Query Processing in Cloud Data Warehouses," *International Journal of High Performance Computing and Networking,* vol. 4, pp. 113-130, 2015.

[10] M. Ficco, "Security Event Correlation Approach for Cloud Computing," *International Journal of High Performance Computing and Networking,* vol. 3, pp. 173-185, 2013.

[11] W. Q.-Y. Zhang Zhao, Tang Wen, "Survey of mining protocol specifications[J]," *Computer Engineering and Applications.,* 2013.

[12] M. Fanzhi Meng , Yuan Liu , Chunrui Zhang, Tong Li "Inferring protocol state machine for binary communication protocol[C]," *Advanced Research and Technology in Industry Applications (WARTIA), 2014.,* pp. 870 - 874 2014

[13] J. Newsome, D. Brumley, J. Franklin, and D. Song, "Replayer: automatic protocol replay by binary analysis," presented at the Proceedings of the 13th ACM conference on Computer and communications security, Alexandria, Virginia, USA, 2006.

[14] B. Anderson, C. Storlie, and T. Lane, "Improving malware classification: bridging the static/dynamic gap," presented at the Proceedings of the 5th ACM workshop on Security and artificial intelligence, Raleigh, North Carolina, USA, 2012.

[15] Z. L. X. J. D. X. X. Zhang, "Automatic protocol format reverse engineering through context-aware monitored execution," *Network & Distributed System Security Symposium, 2008,* 2008.

[16] C. Leita, M. Dacier, and F. Massicotte, "Automatic Handling of Protocol Dependencies and Reaction to 0-Day Attacks with ScriptGen Based Honeypots," in *Recent Advances in Intrusion Detection: 9th International Symposium, RAID 2006 Hamburg, Germany, September 20-22, 2006 Proceedings*, D. Zamboni and C. Kruegel, Eds., ed Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 185-205.

[17] C. L. K. M. M. Dacier, "ScriptGen: an automated script generation tool for Honeyd," *Computer Security Applications Conference,* vol. 12, pp. 203-214, 2006.

[18] P. McMinn, M. Harman, D. Binkley, and P. Tonella, "The species per path approach to SearchBased test data generation," presented at the Proceedings of the 2006 international symposium on Software testing and analysis, Portland, Maine, USA, 2006.

Consideration of Educational Multimedia Contents Introducing Multimedia Switching Function

Kaoru Sugita
Dept. of Information and Communication Engineering
Fukuoka Institute of Technology
sugita@fit.ac.jp

Abstract. Recent years, information technology (IT) has introduced to various educational institutions such as digital teaching material, E-learning system, distance learning system and online course. On the other hand, we are faced with differences of scholastic abilities. For this reason, we focused on differences of progress of learning in a class. In this paper, we discuss an educational multimedia contents to narrow differences of progress of learning.

1 Introduction

Recent years, information technology (IT) has been introduced to many educational institutions such as digital teaching material, e-learning system, distance learning system and online course. These systems were caused to new learning models which student can study interest things anytime and anywhere [1][2]. A digital teaching material is the electronic media constituting the content of learning items. An e-learning system is online workbook for self-learning to manage progress of study [3]. A distance learning system is consists of remote communication tool and learning support tools to shorten the moving time for the distance. An online course is an online lecture using a video on demand service and video communication services. These systems can support to study efficiently for preparation of lesson, self-learning and homework. Of course, many studies about education support system have been conducted to improve an effectiveness of lecture, practical training and experiment. In these studies, many educational methods and system are proposed to improve a self learning [4], a management of learning history [5][6], an active learning method [7], problem/project based learning method [8]-[10], an inverted learning method hybrid of an active learning and a problem based learning method [11]. On the other hands, the lecture video have been published on the Internet and used for various ways in classes [12]-[16]. However, these studies are not focused on differences in their basic academic abilities. In this paper, we discuss multimedia contents to educate students

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_62

with different from scholastic abilities by providing appropriate multimedia expressions.

This paper is organized as follows. We propose a new concept of educational multimedia contents to introduce a multimedia switching function in Section 2. Section 3 explains switching functions applying to education. Finally, Section 4 concludes the paper.

2 Educational Multimedia Contents

Recently, Japanese students are different from their scholastic abilities. In entrance examination, we evaluate a part of learning content because of their differences with learning courses and curriculum in high school. For this reason, a learning support mechanism is important issue to narrow the differences of scholastic abilities.

The multimedia is become to poplar learning resource. As it is well known, the multimedia is spatiotemporally and semantically related with various types media such as video, image, audio and text. In our multimedia educational contents, a multimedia switching function is introduced to switch types of media according to the scholastic abilities as shown in Figure 1. We assume that these contents are prepared with lecturer semi-automatically. Students use their contents evaluated as follows.

(L0) Lack of ability: Unable to understand basic terms and commonsensible instructions in class.

(L1) Poor ability: Able to understand basic terms and commonsensible instructions, but unclear way and use in class.

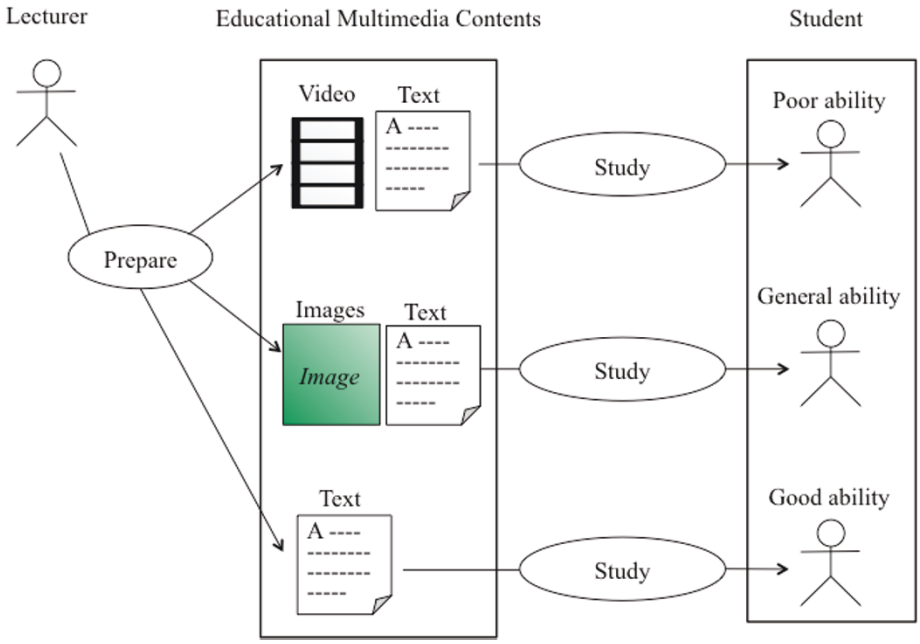


Fig. 1. Educational multimedia contents to introduce multimedia switching function.

(L2) General ability: Able to carry out instructions with some explanatory diagrams.
(L3) Good ability: Able to carry out instructions with simple explanation.

L0 student is required for improvement basic academic ability before registration. L1 student is supposed to select both with video and text contents. The student studies according to the program playing video and text before a lecture and exercise, experience. L2 and L3 student is intended to select learning content as needed.

In the contents, each media includes the information as follows:
(M1) Video: Lecture video.
(M2) Image: Snapshot of important essentials in exercise or experiment.
(M3) Text: Textbook, operation guide, additional explanation and so on.

These contents can be changed to another types of media during playing.

3 Construction Flow for Educational Multimedia Contents

The educational multimedia contents are constructed by 4 phases such as lecture phase, prepare phase, trial phase and improve phase as shown in Figure 2. In the lecture phase, lecture is given from a lecturer at normally. In prepare phase, each media is made from lecture video recorded at the lecture. In trial phase, educational multimedia contents consisted of each media and evaluated for their problem in trial use. In improve phase, text is revised in both misrecognition and supplement for insufficient explanation. After these phases, educational multimedia contents are released in practical use. But the phase goes back to previous phase when any problem is occurred in each phase and practical use.
We have trial use of some software to construct proposed contents. A lecture video is captured by a general video camera and desktop capture software as shown in Figure 3 and Figure 4, respectively. These videos are captured from an object oriented programing in class. Figure 4 shows desktop capture software captured by the NVIDIA ShadowPlay. On the desktop video, textbook and Microsoft Visual Studio as integrated development environment puts on left and right side, respectively. Adobe Premiere Pro CC is used as extraction software both imag optical character recognition e and audio from the lecture video. Also, Adobe Premiere Pro CC is intended for character recognition. A text needs to be revised manually for insufficient explanation.

4 Conclusion

In this paper, we proposed multimedia contents to educate students with different from scholastic abilities by providing appropriate multimedia expressions. In these

contents, each media is switched to other types of media according to the scholastic abilities. Currently, we are investigating lecture video captured in class and text recognition software. In the future, we will construct and evaluate our proposed contents in real environment.

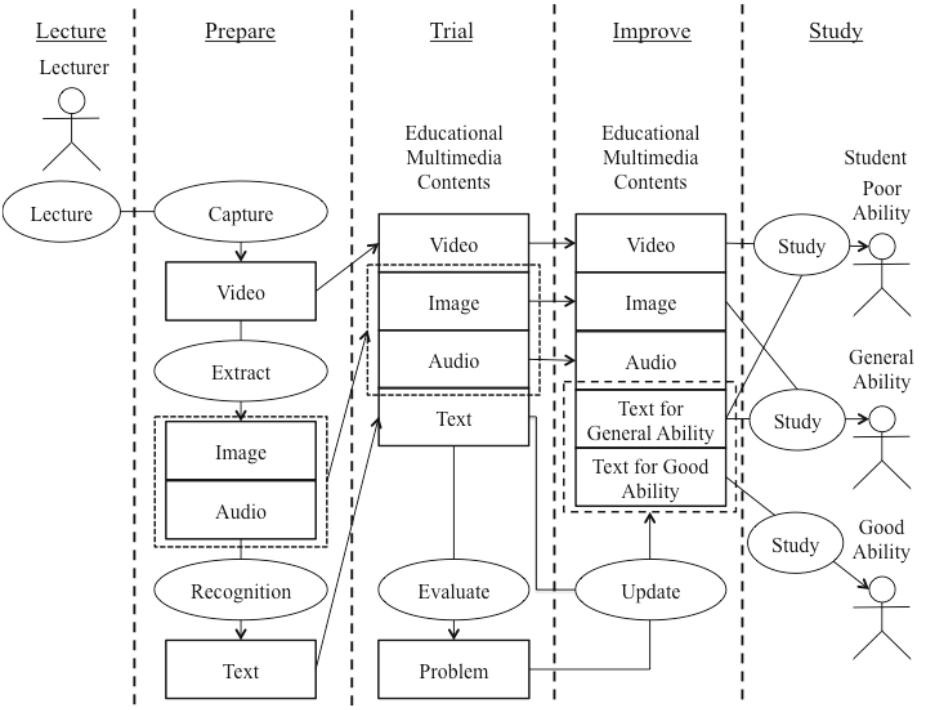


Fig. 2. Construction flow for multimedia educational contents.

Fig. 3. Lecture video captured from the rear of classroom.

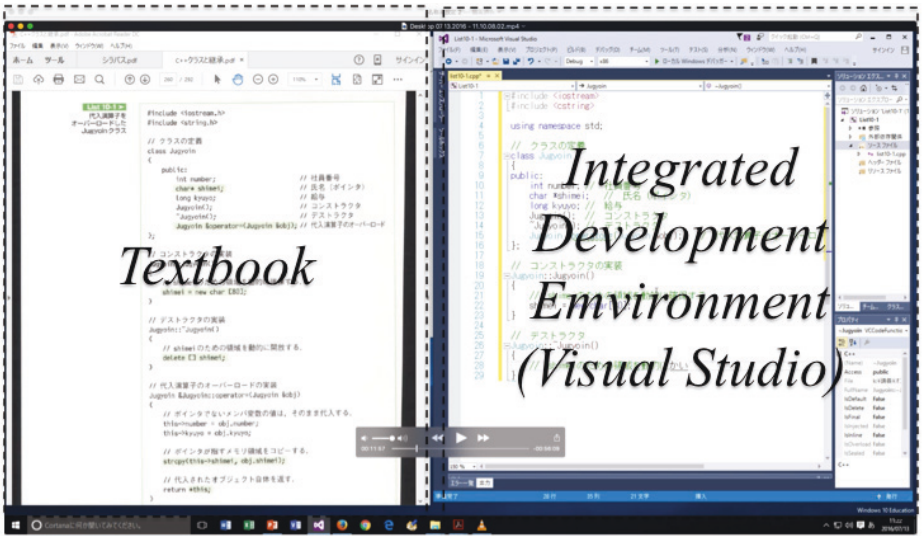


Fig. 4. Lecture video captured on desktop at object oriented programing exercise.

References

1. Latchman, H. A., Salzmann, C., Gillet, D., Bouzekri, H.: Information Technology Enhanced Learning in Distance and Conventional Education. IEEE Transantion on Education, Vol. 42, No. 4. (1999) 247-254
2. Sakamura, K., Koshizuka, N.: Ubiquitous Computing Technologies for Ubiquitous Learning. Proc. of the IEEE International Workshop on Wireless and Mobile Technologies in Education (2005) 11–20
3. Dagger, D., Connor, A. O., Lawless, S., Walsh, E., Wade, V. P.: Service-Oriented E-Learning Platforms. Proc. of the IEEE INTERNET COMPUTING (2007) 28-35
4. Khribi, M. K., Jemni, M.: Automatic Recommendations for E-Learning Personalization Based on Web Usage Mining Techniques and Information Retrieval. Proc. of the Eighth IEEE International Conference on Advanced Learning Technologies (2008) 241-245
5. Rapuano, S., Zoino, F.: A Learning Management System Including Laboratory Experiments on Measurement Instrumentation. IEEE Transaction on Instrumentation and measurement, Vol. 55, No. 5. (2006) 1757-1776
6. Graf, S., Kinshunk, Liu, T. C.: Identifying Learning Styles in Learning Management Systems by Using Indications from Students' Behaviour. Proc. of the Eighth IEEE International Conference on Advanced Learning Technologies (2008) 482-486
7. Stickic, M., Laerhoven, K. V., Schiele, B.: Exploring Semi-Supervised and Active Learning for Activity Recognition. Proc. of the 12th IEEE International Symposium on Wearable Computers (2008) 81-88
8. Striegel, A., Rover, D. T.: Problem-based learning in an introductory computer engineering course. Proc of the Frontiers in Education, Vol. 2 (2002) F1G-7 - F1G-12

9. Kaldoudi, E., Bamidis, P., Papaioakeim, M., Vargemezis, V.: Problem-Based Learning via Web 2.0 Technologies. Proc. of the 21st IEEE International Symposium on Computer-Based Medical Systems (2008) 391-396
10. Avery, Z., Castillo, M., Guo, H., Guo, J., Warter-Perez, N., Won, D. S., Dong, J.: Implementing Collaborative Project-Based Learning using the Tablet PC to enhance student learning in engineering and computer science courses. Proc. of the IEEE Frontiers in Education Conference (2010) F1E-1 - F1E-7
11. Mason, G.S., Shuman, T. R., Cool, K. E.: Comparing the Effectiveness of an Inverted Classroom to a Traditional Classroom in an Upper-Division Engineering Course. IEEE Transaction on Education, vol. 56, No. 4 (2013) 430-435
12. Deniz, D. Z., Karaca, C.: Pedagogically Enhanced Video-on-Demand Based Learning System. Proc. of the FIfth International Conference on Information Technology Based Higher Education and Training (2004) 415-420
13. Yousef, A. M. F., Chatti, M. A., Schroeder, U., Wosnitza, M.: What Drives a Successful MOOC? An Empirical Examination of Criteria to Assure Design Quality of MOOCs. Proc. of the IEEE 14th International Conference on Advanced Learning Technologies (2014) 44-48
14. Subbian, V.: Role of MOOCs in integrated STEM education: A learning perspective. Proc. of the IEEE Integrated STEM Education Conference (2013) 1-4
15. Claros, I., Garmendia, A., Echeverria, L., Cobos, R.: Towards a collaborative pedagogical model in MOOCs. Proc. of the IEEE Global Engineering Education Conference (2014) 905-911
16. Daradoumis, T., Bassi, R., Xhafa, F., Caballé, S.: A Review on Massive E-Learning (MOOC) Design, Delivery and Assessment. Proc. of the Eighth International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (2013) 28-30

Mobile Spam Filtering base on BTM Topic Model

Jialin Ma[1,2*], Yongjun Zhang[1,2], Lin Zhang[1]
[1] Huaiyin Institute of Technology, Huaian, China
majl@hyit.edu.cn, 13511543380@139.com, 11040041@hyit.edu.cn,
[2] College of Computer and Information, Hohai University, Nanjing, China

Abstract. At present, Short Message Service (SMS) is widespread in many countries. Many researchers usually use conventional text classifiers to filter SMS spam. In fact, the actual situation of SMS spam messages isn't consideration by most reseachers. Because the obvious characteristic is the content of SMS spam messages are miscellaneous, shorter and variant. Therefore, traditional classifiers aren't fit to use for SMS spam filtering directly. In this paper, we propose to utilize A Biterm Topic Model(BTM) to identify SMS spam. The BTM can effectively learn latent semantic features from SMS spam corpus. The experiments in our work show the BTM can learn higher quality of topic features from SMS spam corpus, and can more effective in the task of SMS spam filtering.

Keywords: SMS Spam, Topic model, LDA, BTM

1 Introduction

The definition of SMS spam messages is unsolicited and unwanted messages for users. With the rapid development of mobile phone, especially in last ten years, SMS message is one of a important way of information transmission in every day. Therefore, SMS attract criminals to send commercial and advertising information for profit. Because SMS message is cheap and convenient, the criminals abuse to defraud or send various kinds of advertisements [1-3]. The content of spam generally is various kinds, such as advertisements, frauds, erotic services, etc. SMS spam serious violates customers' rights and privacies, consumes the resource of mobile communication equipment, and result in the congestion of communication network [1, 4, 5]At present, SMS spam is overflowing in many countries, especially in India, Pakistan, and China [6]. It is reported that the Chinese SMS clients received 16 spam messages every week in average, 17% were encountered SMS bombing in the second half of 2014[7]. In the USA, more than 69% of the surveyed mobile users claimed to have received text spam in 2012[8].

The traditional classifiers, such as SVM, K-NN, Bayes [6], can't be suitable to class SMS spam. The obvious and fundamental reason is the SMS spam messages is miscellaneous and distinct from general class of text files. In this paper, we propose to utilize A Biterm Topic Model(BTM) to identify SMS spam. The BTM can effectively learn latent semantic features from SMS spam corpus. The BTM is based on the probability of topic model theory and first proposed by Yan, X. et al. in 2013. . We de-

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_63

rived from the work of Yan which is designed to model the generative process of the word co-occurrence patterns in short texts. By contrast, the BTM has the follow advantages: (a) the BTM remains the superiority of Latent Dirichlet Allocation (LDA)[10] , but more suitable for short texts; (b) BTM is very fit for identifying the miscellaneous SMS spam.

The paper is organized as follows: Section 2 presents our method in detail. Section 3 shows the experiments and discussion. Finally, we conclude and discuss further research in Section 4.

2 Our Work

In this section, we introduce the details of BTM[9] and our method of identifying SMS spam.

2.1 BTM

Conventional topic models, like PLSA[11] and LDA[10], reveal the latent topics within corpus by implicitly capturing the document-level word co-occurrence patterns. Therefore, directly applying these models on short texts will suffer from the severe data sparsity problem (i.e. the sparse word co-occurrence patterns in each short document)[9]. More specifically, (1) the occurrences of words in short document play less discriminative role when it is compared to lengthy documents where the model has enough word counts to know how words are related; (2) the limited contexts make it more difficult for topic models to identify the senses of ambiguous words in short documents[9]. They considered that a pair of words can much better reveal the topics then the occurrence of single word. They proposed a Biterm Topic Model(BTM), which can learn topics over short texts by directly modelling the generation of biterms in the whole corpus. The major advantages of BTM are that 1) BTM explicitly models the word co-occurrence patterns to enhance the topic learning; and 2) BTM uses the aggregated patterns in the whole corpus for learning topics to solve the problem of sparse word co-occurrence patterns at document-level[9].

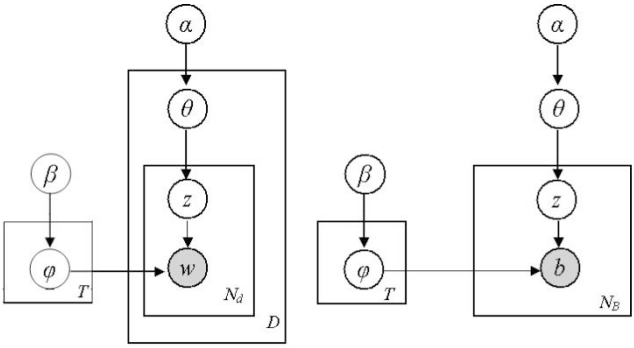


Fig. 1. Graphical models for (a)LDA,, (b) BTM

Fig. 1(a) shows the graphical model for the "standard topic model"(LDA). D is the number of documents in the corpus and document d has N_d words. The process includes two steps: first, assign a topic number from document-topic distribution θ; then, draw a word from topic-word distribution φ. All documents share T topics. Document-topic and topic-word distributions all obey the multinomial distributions, and each of them is governed by symmetric Dirichlet distribution. α and β are hyper-parameters of symmetric Dirichlet priors for θ and φ. Parameters θ and φ can be obtained through a Gibbs sampling.

Fig. 1(b) shows the graphical model for the "Biterm Topic Model"(BTM). N_B is the total number of biterm in the corpus, T is the topic number, b is one of a biterm which is a unordered term-pair (w_i, w_j) appeared in a short text. Different from LDA, topic assignment z is from the corpus-level topic distribution θ in BTM.

The probability of a biterm b is described as the follow[9]:

$$p(b) = \sum_{z} p(z) p(b \mid z) \tag{1}$$

The generative probability of the whole corpus is expressed as:

$$\begin{aligned} p(\vec{b}) &= \prod_{i=1}^{N_B} (\sum_{z} p(z) p(b \mid z)) \\ &= \prod_{i=1}^{N_B} (\sum_{z} \theta_z \varphi_{b/z}) \end{aligned} \tag{2}$$

The generative process of BTM is as follows:

1 Draw $\varphi \sim Dir(\beta)$, Draw$\theta \sim Dir(\alpha)$

2 For each topic t=1,2,…,T

Draw $\varphi \sim Dir(\beta)$

3 For each biterm b=1,2,...N_B

(I) Draw z~ $Multi(\theta)$

(Π) Draw $b \sim Multi(\varphi)$

2.2 Parameters inference

In the work of [9], because z is latent variables in BTM, it is difficult to estimate them exactly. Therefore, they adopt Gibbs sampling to estimate the parameters in the model. Gibbs sampling is an effectively and widely applicable Markov chain Monte Carlo algorithm for latent variable inference. In order to conduct Gibbs sampling, the first step is to initial states for the Markov chain randomly. Note that once a biterm b=(w_i,w_j) is assigned to the topic z, the two words w_i and w_j in it will be assigned to the topic simultaneously. Then we calculate the conditional distribution $P(z|z_{-b},B, \alpha, \beta)$ for each biterm b. Gibbs sampling equation are showed as the follow[9]:

$$P(z / \mathbf{z}_{-b}, B, \alpha, \beta) \propto (n_z + \alpha) \frac{(n_{w_i/z} + \beta)(n_{w_j/z} + \beta)}{\sum_w (n_{w/z} + M\beta)^2} \tag{3}$$

$$\varphi_{w/z} = \frac{n_{w/z} + \beta}{\sum_w n_{w/z} + M\beta} \tag{4}$$

$$\theta_z = \frac{n_z + \alpha}{|B| + K\alpha} \tag{5}$$

In Eq. (3) '-' denotes the current instance is removed, B is the total number of biterms in the corpus. $n_{w/z}$ is count of biterm b is assigned to topic z, M. is regulation parameter for β.

2.3 Classifying SMS message

In order to filtering SMS spam, we must be to identify an unknown SMS message. First we can train the BTM on SMS spam corpus, then infer topic for the unknown SMS via the BTM. The next step is to used JS(Jensen-Shannon)-divergence to measure the similarity between the unknown message and all spam message in the training corpus. Finally, we use presupposed threshold value μ to judge whether this unknown message is a spam. If D_{JS} value is more than the presupposed threshold value μ, we can judge the ms is spam message. The D_{JS} can be calculated as the follow formula[12]:

$$D_{JS}(p_t^c \,||\, p_t^{ms}) = \frac{1}{2}(D_{KL}(p_t^c \,||\, p) + D_{KL}(p_t^{ms} \,||\, p)) \tag{6}$$

In Eq. (6), $p = (p_t^c + p_t^{ms})/2$, and D_{KL} is the KL-divergence.

3 Experiments and discussion

In this following, we introduce about the experiments of SMS spam filtering using BTM, including experimental datasets, evaluation and comparsion. The experiments were conducted on a computer with 8GB memory and dual Intel processor with i5-4210U cores at 2.40 GHz.

3.1 Dataset

In order to evaluate the proposed method and acquire rich train data, we integrated two public SMS datasets: SMS Spam Collection v.1 and DIT SMS Spam Dataset. Th

e first dataset comes from UCI machine learning repository, which has been used in re levant researches [13, 14]. In particularly, the details about this corpus has been introd uced in the paper [2]. It can be downloading freely in web site: http://archive.ics.uci.e du/ml/datasets/SMS+Spam+Collection. Another SMS corpus: DIT SMS Spam Datase t (http://www.dit.ie/computing/research/resources/smsdata/) only has SMS spam data. Except for collecting messages from scraping two UK public consumer complaints w ebsites, DIT SMS Spam Dataset is also include 639 SMS spam from SMS Spam Coll ection v.1. More information about DIT dataset can be seen in the paper [6].The detail s about the two datasets are shown in Table 1.

Table 1. The details about the two datasets

Dataset name	Hams		Spams		Total
SMS Spam Collection v.1	4827	86.6%	747	13.4%	5574
DIT SMS Spam Dataset	0	0	1353	100%	1353

3.2 Evaluation Method

In oder to verify the effect of method using BTM for SMS filtering, we adopt traditional evaluation indicator F_1 and *Accuracy* and to evaluate the performances. According the classifiers' results and the true class, the samples can be divided into false positive(*FP*), true positive(*TP*), false negative(*FN*) and true negative(*TN*)[16]. It can describe by the confusion matrix as the table 2.:

Table 2. Classify resuls confusion matrix

Fact	Predict results	
	Positive sample	Negative sample
Positive sample	TP	FN
Negative sample	TN	FT

$$Precision(P) = \frac{TP}{TP + FP} \tag{7}$$

$$Recall(R) = \frac{TP}{TP + FN} \tag{8}$$

$$Accuracy(A) = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

$$F_1 = \frac{2RP}{R+P} \tag{10}$$

10-fold cross validation usually is adopted to reduce the contingency of the experiments. We bring in the follow indicators:

$$macro-P = \frac{1}{n}\sum_{i=1}^{n} P_i \quad macro-R = \frac{1}{n}\sum_{i=1}^{n} R_i \tag{11}$$

$$macro-F1 = \frac{2\times macro-P\times macro-R}{macro-P+macro-R} \tag{12}$$

$$micro-P = \frac{\overline{TP}}{\overline{TP}+\overline{FP}} \quad micro-P = \frac{\overline{TP}}{\overline{TP}+\overline{FN}} \tag{13}$$

$$micro-F1 = \frac{2\times micro-P\times micro-R}{micro-P+micro-R} \tag{14}$$

3.3 Parameters Determination

In all most topic models, topic number isn't uniform value. It is correlated with the individual corpus. In order to verify a matched topic number for our training corpus, we conducted a series of T (T=5,…, 30) values to test, and adopted macro and micro to estimate the parameter T of BTM. Other parameters refer to the experience of work [15], set α = 50/ T and β= 0.01.

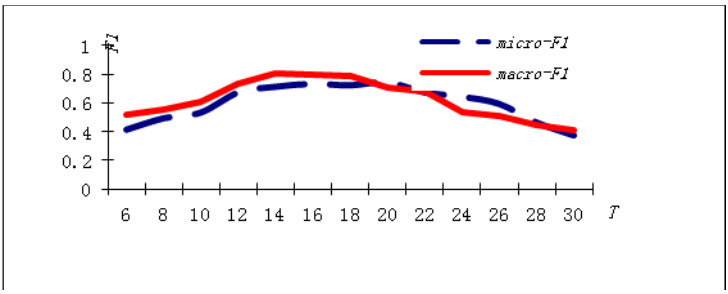


Fig. 2 *macro-F1* and *micro-F1* for different T

Fig. 2 reflects the classifying results of *micro-F1* and *macro-F1* in different T, from 6 to 30. The *micro-F1* and *macro-F1* are all varying alone with the increase of the number of topics, but they reaches crest when T is near 16. Therefore, we consider 16 is the appropriate parameter for T the experimental corpus.

3.4 Comparison

We conducted the experiments on the two models: LDA and BTM using the same classifying method we proposed. In additional, Naive Bayes(NB) as a classic classifor is used for comparison and baseline. We set T, α and β the same in BTM and LDA, and set μ=0.8. The experiments were run on the same hardware and software condition.

Table 3. Time consumption comparison

	BTM	LDA	NB
Times	21m25s	9m53s	7m31s

Table3 reflect the time gap of the different method are big. It is clear in table 2 that BTM is the longest, the next is LDA. In contrast, the NB is the most efficient. The main reason is BTM is consume most time on biterms, but the LDA works on words. If a corpus has N terms, it will have C_2^N biterms. This will increase the time consumption. Parameters inferring would consume more time in the processing of Gibbs sampling.

Furthermore, we compared the effect of filtering SMS spam on the three methods. *Accuracy* and F_1 were used to evaluate the performances.

Table 4. *micro-F1* and *macro-F1* comparsion

	BTM	LDA	NB
micro-F1	0.78	0.63	0.77
macro-F1	0.81	0.59	0.77

Table 4 shows the *micro-F1* and *macro-F1* of the three methods. It is clear that BTM is superior to LDA and NB. LDA is the most poor of them. The fundament reason is LDA isn't suitable to short text, such as SMS message analysis. By contrast, NB also have good performance. The BTM is a little better than BN. It is prove that BTM is more suitable to process short text. Therefore, it is effect to filtering SMS spam. This is also certified in Fig. 3 about the *accuracy*.

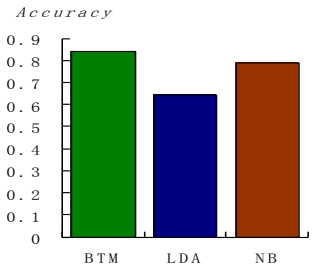


Fig. 3 *Accuracy* comparison of the three methods

4 Conclusions and future Work

At present, SMS spam has been overflowing in many countries. Many of the traditional classification algorithms isn't suitable for SMS spam filtering, because the content of SMS spam messages are miscellaneous and distinct from general class of text files. In this paper, we propose to use the model of BTM[9], which can automatically learn latent semantic features from SMS spam corpus using for the task of SMS spam filtering. The experiments the BTM is fit for identifying the miscellaneous SMS spam.Future work should be considered to improve the BTM. Because in the experiments, we found that the BTM was influenced by much noisy data. This hurt the quality of topics which acquired by BTM.

Acknowledgments. This work was supported by the Graduate Student Scientific Research and Innovation Project of Jiangsu Province, China (China Central University Basic Scientific Research Business Fund, Hohai university, Grant No.: 2015B38314), Science and Technology Projects of Huaian (Grant No.:HAS2015033), the University Science Research Project of Jiangsu Province(Grant No.:15KJB520004), and the Science and Technology Projects of Huaian (Grant No.:HAG2015060).

References

1. Wu, N., Wu, M., and Chen, S.,: Real-time monitoring and filtering system for mobile SMS. IEEE Conference on Industrial Electronics & Applications, 1319 (2008).
2. Almeida, T. A., Hidalgo, J. M. G., and Yamakami, A.,: Contributions to the study of SMS spam filtering: new collection and results. in Proceedings of the 11th ACM symposium on Document engineering, ACM, pp. 259 (2011).
3. Sohn, D.-N., Lee, J.-T., and Rim, H.-C.,: The contribution of stylistic information to content-based mobile spam filtering. in Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, Association for Computational Linguistics, pp. 321 (2009).
4. Sohn, D.-N., Lee, J.-T., Han, K.-S., and Rim, H.-C.,: Content-based mobile spam classification using stylistically motivated features. Pattern Recognition Letters, 33, 364 (2012).
5. Wadhawan, A., and Negi, N.,: A Novel Approach For Generating Rules For SMS Spam Filtering Using Rough Sets. INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME, 3, 5 (2014).
6. Delany, S. J., Buckley, M., and Greene, D.,: SMS spam filtering: Methods and data. Expert Systems with Applications, 39, 9899 (2012).
7. Center, C. N. S. R., (2014).
8. Jiang, N., Jin, Y., Skudlark, A., and Zhang, Z.-L.,: Understanding sms spam in a large cellular network: characteristics, strategies and defenses, in Research in Attacks, Intrusions, and Defenses. Springer, pp. 328 (2013).

9. Yan, X., Guo, J., Lan, Y., and Cheng, X.,: A biterm topic model for short texts. in Proceedings of the 22nd international conference on World Wide Web, International World Wide Web Conferences Steering Committee, pp. 1445 (2013).
10. Blei, D. M., Ng, A. Y., and Jordan, M. I.,: Latent dirichlet allocation. the Journal of machine Learning research, 3, 993 (2003).
11. Hofmann, T.,: Probabilistic latent semantic indexing. in Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, ACM, pp. 50 (1999).
12. Endres, D. M., and Schindelin, J. E.,: A new metric for probability distributions. IEEE Transactions on Information theory (2003).
13. Ho, T. P., Kang, H.-S., and Kim, S.-R.,: Graph-based KNN Algorithm for Spam SMS Detection. J. UCS, 19, 2404 (2013).
14. Ahmed, I., Ali, R., Guan, D., Lee, Y.-K., Lee, S., and Chung, T.,: Semi-supervised learning using frequent itemset and ensemble learning for SMS classification. Expert Systems with Applications, 42, 1065 (2015).
15. Heinrich, G., :Parameter estimation for text analysis, Technical Report (2004).
16. Ma J, Zhang Y, Wang Z, et al.: A Message Topic Model for Multi-Grain SMS Spam Filtering. International Journal of Technology and Human Interaction (IJTHI), 2016, 12(2): 83-95.

Construction of an Electronic Health Record System for supporting a Zoo Veterinarian

Tatsuya Oyanagi[1], Misaki Iyobe[1], Tomoyuki Ishida[1],
Noriki Uchida[2], Kaoru Sugita[2], Yoshitaka Shibata[3]
[1] Ibaraki University, Hitachi, Ibaraki 316-8511, Japan
13t4010y@vc.ibaraki.ac.jp, 16nm705a@vc.ibaraki.ac.jp,
tomoyuki.ishida.49@vc.ibaraki.ac.jp
[2] Fukuoka Institute of Technology, Fukuoka, Fukuoka 811-0295, Japan
n-uchida@fit.ac.jp, sugita@fit.ac.jp
[3] Iwate Prefectural University, Takizawa, Iwate 020-0693, Japan
shibata@iwate-pu.ac.jp

Abstract. We constructed the Electronic Health Record System for supporting a zoo veterinarian. This system supports creating of an animal health record for a zoo veterinarian. And, this system greatly reduces the creating time of the animal health record by digitalization of the health record. Moreover, a zoo veterinarian can easily grasp the past animal health record by referring to the Animal Health Record Management Database Server. On the other hand, this system corresponds to various terminal device by adopting the responsive design. Therefore, a zoo veterinarian can operate the animal health record at any arbitrary place through a tablet terminal.

1 Introduction

The zoo in Japan now has many problems to solve such as decrease of the number of visitors and decrease of entrance fees. The number of zoo visitors have been decreasing every year in the nation. Kamine zoo in Hitachi City, Ibaraki Prefecture is no exception. Due to diversification of recreational facilities, fewer children, decrepit premises, and so on, the number of Kamine zoo visitors also decreases from the peak. Therefore, Hitachi City maintained facilities such as "Forest of Chimpanzees" and "Pasture of Elephants" by Kamine zoo development project in order to restore the attractiveness and the prosperity of Kamine zoo. The number of visitors were about 460,000 visitors in 1970, but it decreased to about 260,000 visitors in 2008. Then it recovered to about 360,000 visitors in 2014 by maintenance in facilities [1][2]. Kamine zoo has to grope after a plan of an attractive event and exhibition method within the limited budget and staff to increase the number of visitors. Under such circumstances, as for internal operations of Kamine zoo, conventional analog operations often exist. These analog operations are obstruction of smoothness and

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_64

efficient work. The animal health record is no exception. Currently, a zoo veterinarian is managing the health record by paper medium and Excel data. Analog management of the health record causes routine work. We will carry out research in order to resolve this problem.

2 Purpose of Research

In this research, we construct the Electronic Health Record System for supporting a zoo veterinarian. This system covers all items of paper medium animal health record. We construct the input form design of this system with the same design as paper medium animal health record. Thus, a zoo veterinarian can reduce the learning cost to use the electronic health record system. And, when referring to a past animal health record, the paper medium animal health record needs much time and trouble. A zoo veterinarian can access a past animal health record by realization of the electronic health record system. Moreover, a zoo veterinarian can browse the animal health record on various terminals by digitalization. Therefore, a zoo veterinarian can share an efficient and effective information between the staff because they can browse the animal health record at any arbitrary place.

3 System Configuration

The system configuration of this research is shown in Fig. 1. The Electronic Health Record System consists of the Animal Health Record Management Agent, the Animal Health Record Edit Agent, the Animal Health Record Browse Agent, the Animal Health Record Management Application Server, and the Animal Health Record Management Database Server. The function of each agent and server is described below.

- The Animal Health Record Management Agent
  The Animal Health Record Management Agent has complete authority such as allocation of data registration, data compilation, and data deletion of the Electronic Health Record System. This agent assumes a zookeeper.
- The Animal Health Record Edit Agent
  The Animal Health Record Edit Agent authorized to edit of the Electronic Health Record System by the Animal Health Record Management Agent. This agent assumes a zoo veterinarian.
- The Animal Health Record Browse Agent
  The Animal Health Record Browse Agent authorized to browse of the Electronic Health Record System by the Animal Health Record Management Agent. This agent assumes zoo clerical staff and animal breeder staff.
- The Animal Health Record Management Application Server
  The Animal Health Record Management Application Server stores the animal health record information registered by the Animal Health Record Edit Agent in the Animal Health Record Management Database Server.

- The Animal Health Record Management Database Server
  The animal health record information stored in the Animal Health Record Management Database Server according to the data request from the Application Server.

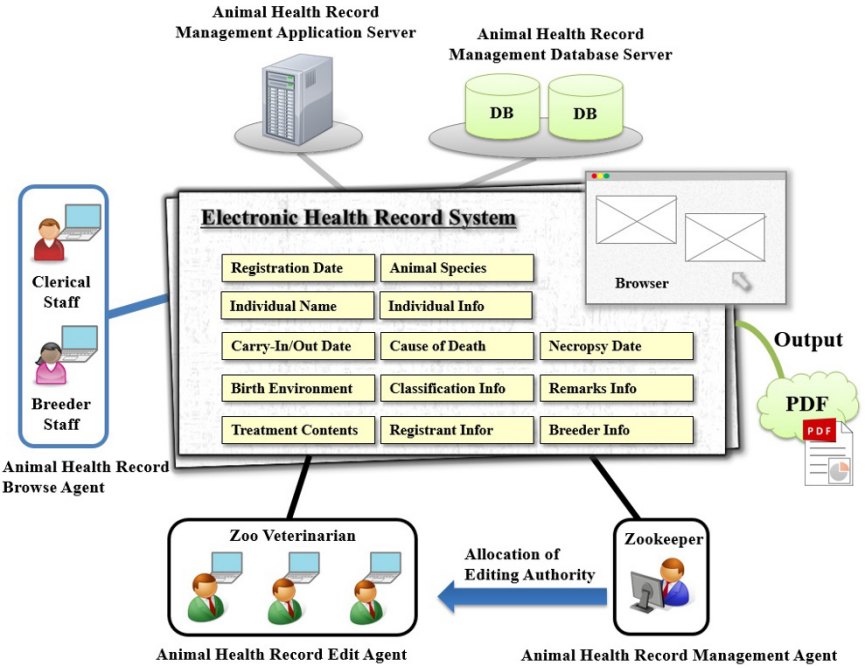


Fig. 1. System configuration of this research

4 System Architecture

The system architecture of this research is shown in Fig. 2. This application server consists of Animal Health Record and Animal Individual Information View Manager, Information Provide Manager, Information Index Manager, Information Registration Manager, Information Edit Manager, Information Validation Check Manager, Animal Health Record Management Controller, Information Retrieval Manager, Information Storage Manager, Information Update Manager, Information Delete Manager, Business Logic Processing Manager, Image Registration Manager, Information Query Processing Manager. Animal Health Record Management Controller plays an important role in the Animal Health Record Management Application Server. Animal Health Record Management Controller performs instruction processing to each module. This controller controls registration and edit processing of health record

information through Animal Health Record and Animal Individual Information View Manager.

The Animal Health Record Management Database Server consists of Animal Health Record Information Storage and Animal Individual Information Storage. The animal health record information is stored in Animal Health Record Information Storage, and the animal individual information is stored in Animal Individual Information Storage.

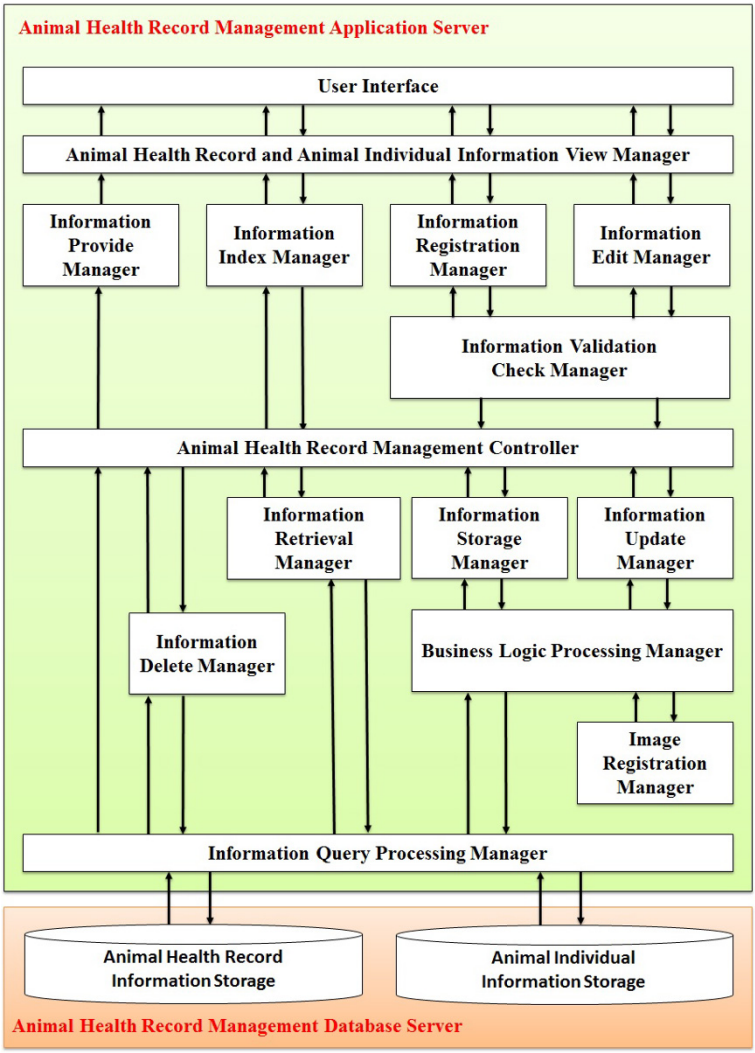


Fig. 2. System architecture of this research

5 Prototype System

The health record management screen is shown in Fig. 3. A zoo veterinarian can refer to all animal health records registered in the Animal Health Record Management Database Server through this health record management screen. The animal health record information registered in the database is listed on the bottom of the screen. The following information is included in a list.

- Animal Species Information
- Animal Individual Name Information
- Animal Number Information
- Microchip Number Information
- Registration Date Information
- Last Update Date Information
- Registrant Information

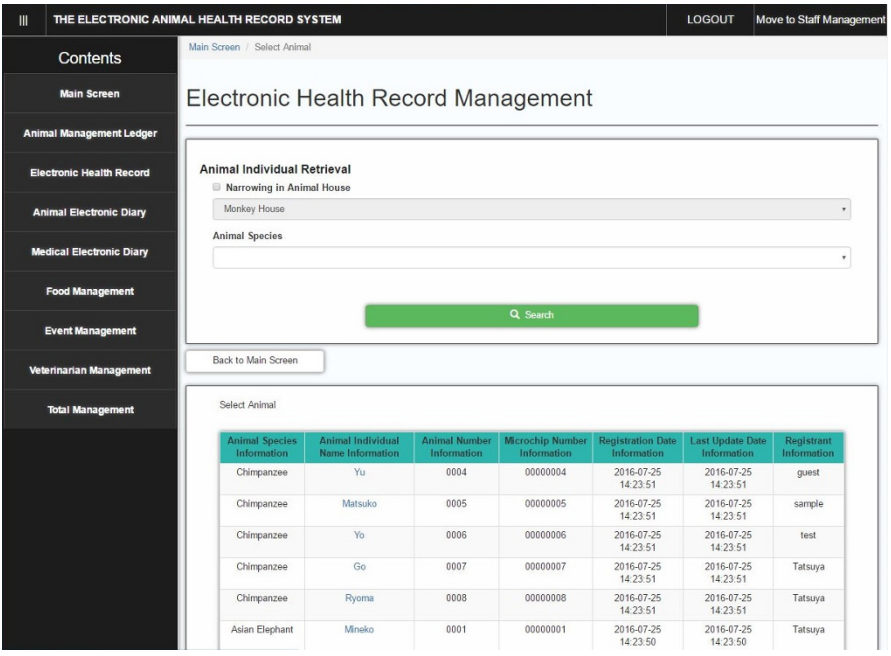


Fig. 3. Health Record Management Screen

The health record management screen transits in the animal individual health record management screen (Fig. 4) by a zoo veterinarian selecting an arbitrary animal individual name. The health record of an animal individual can be managed on the health record management screen. A zoo veterinarian can operate all health record information including the past animal health record data. Therefore, a zoo veterinarian can refer to the past animal health record information easily, and can search for the

animal health record in free words on the top of the screen. By this retrieval function, a zoo veterinarian can retrieve for the past animal health record information related to the selected animal individual. All animal health record information related an animal individual listed on the bottom of the screen. The following information is included in a list.

- Registration Date Information
- Symptom / State / Course
- Treatment Contents
- Registrant Information
- Edit or Delete

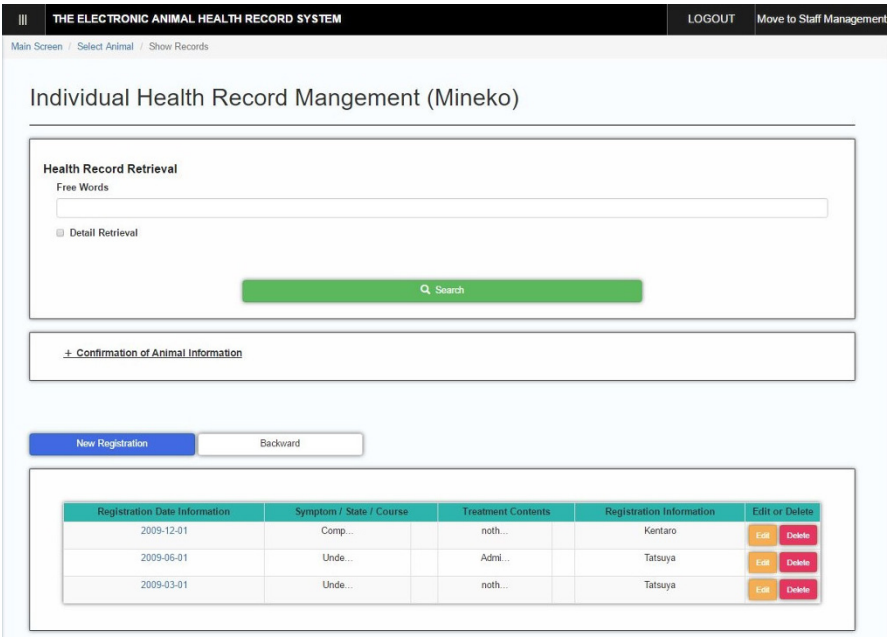


Fig. 4. Animal Individual Health Record Management Screen

The health record information can be browse by selecting a date in the list. Registered health record information can be edit and delete by selecting "Edit" and "Delete" button in the list. When registering the new animal health record information, the animal individual health record management screen transits in the health record registration screen (Fig. 5) by a zoo veterinarian selecting a "New Registration" button. The health record information can be registered on the health record registration screen. A register date of the health record information is displayed automatically to a "Date" column in the top of the screen. And, a registrant is also displayed automatically to a "Registrant" column from login information. A zoo veterinarian can input the symptom and the state related to disease and injury of animal to a "Symptom / State / Course" column freely, and can register a picture of

the state of injury. Moreover, a zoo veterinarian can input the treatment contents to a "Treatment Contents" column freely, and also can register a picture of the state of treatment.

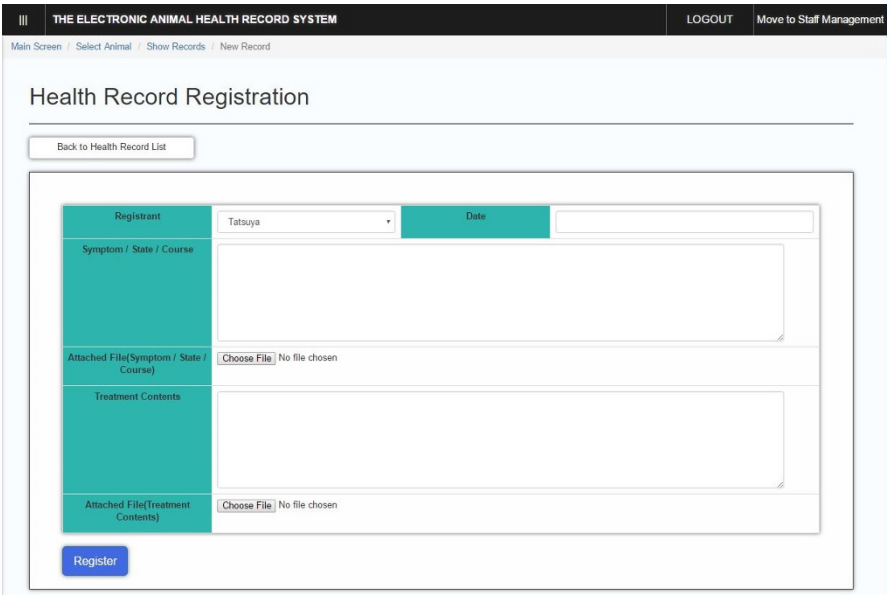


Fig. 5. Health Record Registration Screen

When a "Edit" button of the list is selected in the animal individual health record management screen (Fig. 4), this screen changes to the edit of health record information screen as shown in Fig. 6. A zoo veterinarian can edit the registered treatment contents, and can register a related picture additionally newly. On the other hand, when a "Delete" button of the list is selected in the animal individual health record management screen (Fig. 4), the confirmation dialogue is pop-up displayed as shown in Fig. 7. When an "OK" button of the confirmation dialogue is selected, registered health record information is deleted from the database.

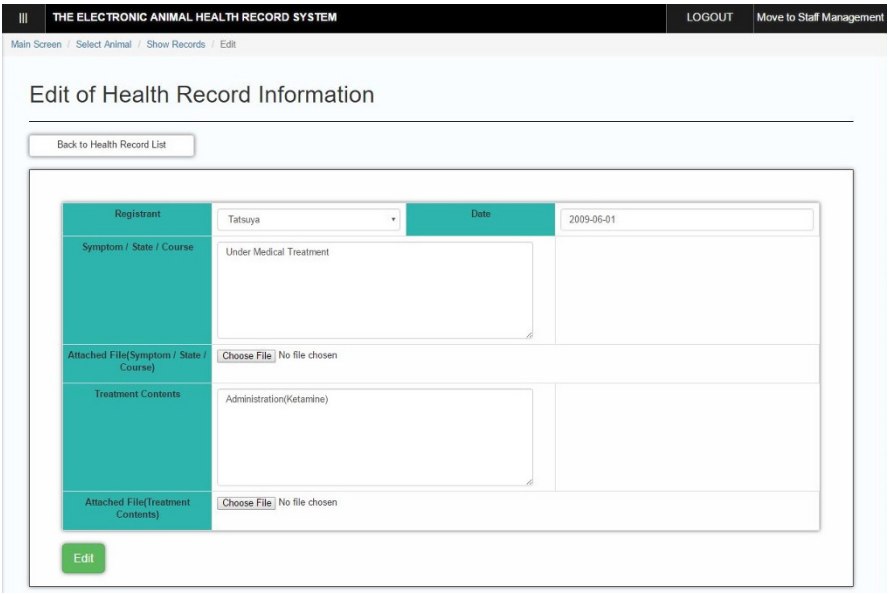


Fig. 6. Edit of Health Record Information

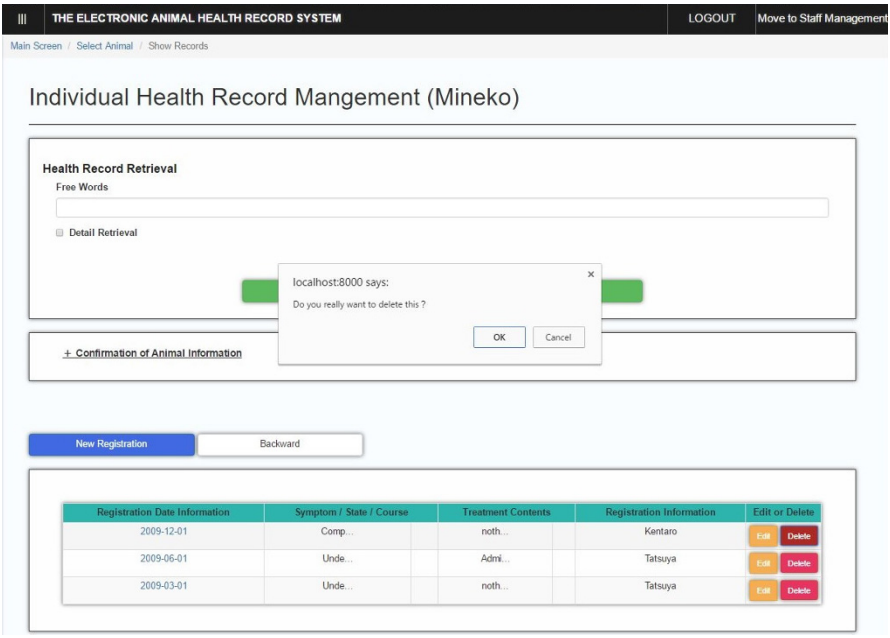


Fig. 7. Edit of Health Record Information

6 Conclusion

In this research, we constructed the Electronic Health Record System for supporting a zoo veterinarian. This system covers all items of paper medium animal health record, and we constructed the input form design of this system with the same design as paper medium animal health record. Thus, we succeeded to reduce the learning cost when a zoo veterinarian uses the electronic health record system. And, when referring to a past animal health record, the paper medium animal health record needs much time and trouble. However, a zoo veterinarian could access a past animal health record by realization of the electronic health record system. Moreover, a zoo veterinarian could browse the animal health record on various terminals by digitalization. Therefore, a zoo veterinarian could share an efficient and effective information.

7 Future Tasks

We are planning to advance research and development of the zoo business integrated management support system including the animal electronic diary management system, the animal management ledger system, and the feed management system from now on as shown in Fig. 8. The zoo business integrated management support system consists of the following system group.

- Animal Electronic Health Record System
- Animal Electronic Diary Management System
- Medical Treatment Electronic Diary System
- Feed Management System
- Animal Management Ledger System
- Zoo Veterinarian Management System
- Zoo Event Management System
- Registration Information Management System

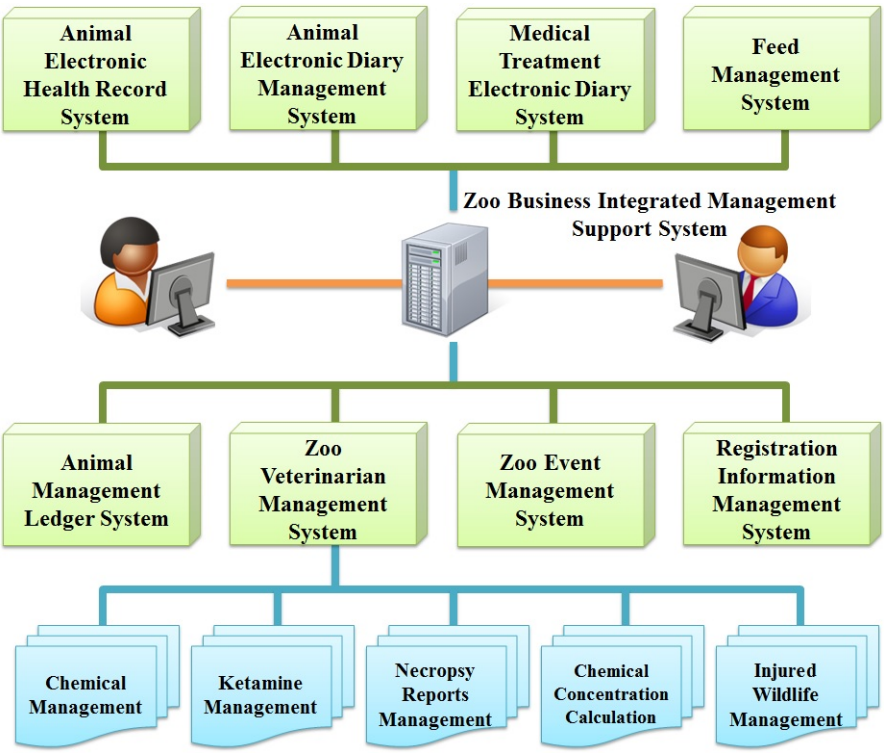


Fig. 8. Zoo Business Integrated Management Support System

Acknowledgments. The authors would like to thank N. Namae for total assistance with the system construction. We also thank Kamine zoo staff for fruitful discussions and valuable suggestions.

References

1. Kamine Zoo : Kamine Zoo Annual Report (2014).
2. Hitachi City Assembly, Hitachi-city Administrative Activities Case Report, http://hitachi-gr-giindan.jp/photonews/image/2010-hitachi.pdf, last viewed September 2016.

Part V
Workshop CADSA-2016: 4th International Workshop on Cloud and Distributed System Applications

An Architecture for processing of Heterogeneous Sources

Flora Amato, Giovanni Cozzolino, Antonino Mazzeo, Sara Romano

Abstract Different sources of information generate every day huge amount of data. For example, let us consider social networks: here the number of active users is impressive; they process and publish information in different formats and data are heterogeneous in their topics and in the published media (text, video, images, audio, etc.). In this work, we present a general framework for event detection in processing of heterogeneous data from social networks. The framework we propose, implements some techniques that users can exploit for malicious events detection on Twitter.

1 Introduction

Social media networks have emerged as powerful means of communication for people looking to share and exchange information on a wide variety of real-world events. These events range from popular, widely known ones (e.g., a concert by a popular music band) to smaller scale, local events (e.g., a local social gathering, a protest, or an accident). Short messages posted on social media sites such as Twitter can typically reflect these events as they happen. For this reason, the content of such social media sites is particularly useful for real-time identification of real-world events. Social networks have become a virtually unlimited source of knowledge, that can be used for scientific as well as commercial purposes. In fact, the analysis of on-line public information collected by social networks is becoming increasingly popular, not only for the detection of particular trend, close to the classic market research problem, but also to solve problems of different nature, such as the identification of fraudulent behaviour, optimization of web sites, tracking the geographical location of particular users or, more generally, to find meaningful pat-

Flora Amato,Giovanni Cozzolino, Antonino Mazzeo, Sara Romano
DIETI - Università degli Studi di Napoli "Federico II", via Claudio 21, Naples, e-mail: flora.amato,giovanni.cozzolino,mazzeo,sara.romano@unina.it

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_65

terns in a certain set of data. In particular, the rich knowledge that has accumulated in Twitter enables to catch the happening of real world events in real-time. These event messages can provide a set of unique perspectives, regardless of the event type [1, 2], reflecting the points of view of users who are interested or even participate in an event. In particular, for unplanned events (e.g., the Iran election protests, earthquakes), Twitter users sometimes spread news prior to the traditional news media [3]. Even for planned events, Twitter users often post messages in anticipation of the event, which can lead to early identification of interest in these events. Additionally, Twitter users often post information on local, community-specific events where traditional news coverage is low or nonexistent. Thus Twitter can be considered as a collector of real-time information that could be used by public authorities as an additional information source for obtaining warnings on event occurrence. In the last few years, particular interest has given to the extraction and analysis of information from social media by authorities and structures responsible for the protection of public order and safety. Increasingly, through the information posted on social media, public authorities are able to conduct a variety of activities, such as prevention of terrorism and bio-terrorism, prevention of public order problems and safety guarantee during demonstrations with large participation of people.

The heterogeneity of information and the huge scale of data makes the identification of events from Twitter a challenging problem. In fact, Twitter messages, or tweets, has a variety of content types, including personal updates, not related to any particular real-world event, information about event happenings, retweets of messages which are of interest for a user and so on [4]. As an additional challenge, Twitter messages contain little textual information, having by design the limit of 140 characters and often exhibit low quality [5]. Several research efforts have focused on identifying events in Twitter [6, 7]. For example, in recent years there has been a lot of research efforts in analyzing Tweets to enhance health related alerts[8], following a natural calamity or a bio-terrorist attack, which can urge a rapid response from health authority, as well as disease monitoring for prevention. The previous work in this area includes validating the timeliness of Twitter by correlating Tweets with the real-world statistics of disease activities, e.g., Influenza-like-Illness (ILI) rates [9, 10, 11], E-coli [12], cholera [13] or officially notified cases of dengue [14]. In this work, we present a general framework for event detection, starting from the heterogeneous processing of data coming from social networks systems. The proposed system process heterogeneous information in order to detect anomalies in Twitter stream. Event related anomalies are patterns in data that do not fit the pattern of the expected normal behavior [15]. Those anomalies might be induced in the data for malicious activity, as for example cyber-intrusion or terrorist activity. The proposed framework aims at analyzing tweets and automatically extract relevant event related information in order to raise a signal as soon as a malicious event activity is detected.

The reminder of the paper is structured as follows: in Section 2 we describe two examples that motivate our work; in Section 3 we present an overview of the framework along with the goals and requirements that are of importance for event detection from Twitter. Moreover in this Section we describe the main processing stages of the system. In the end, in Section 4 we present some conclusions and future work.

2 Motivating example

Social Network, as Twitter, have become a virtually unlimited source of knowledge, that can be used for scientific as well as commercial purposes. Twitter users spreads a huge amount of information ranging on a several "facts". The analysis of the information collected by social networks is becoming increasingly popular for the detection of trends, fraudulent behavior or, more generally, for finding meaningful patterns in a certain set of data. This process of extraction and analysis of large amounts of data is commonly defined data mining.

Particular attention has been paid to the analysis of the User-Generated Content (UGC) coming from Twitter, which is one of the most popular microblogging websites. As described in the Section 1, Twitter textual data (i.e., tweets) can be analyzed to discover user thoughts associated with specific events, as well as aspects characterizing events according to user perception. Tweets are short, user-generated, textual messages of at most 140 characters long and publicly visible by default. For each tweet a list of additional features (e.g., GPS coordinates, timestamp) on the context in which tweets have been posted is also available.

In this Section we show two examples of text mining on live tweets to monitor possible public security issues. In the first one we describe the topic relevance, in the second one we analyze the results, using some realistic data taken during a sport event.

In the last years, especially from the episodes following Sept. 11, the fear of bioterrorism is in the list of most urgent healthcare concerns, together with previous health issues - cancer, AIDS, smoking or alcohol/drug abuse, heart disease and the cost of healthcare and insurance. Gallup[1] has asked Americans since 2001 to name the most urgent health problem facing the U.S.. Cost of and access to healthcare have generally topped the list in last few years, while Americans most frequently mentioned diseases such as cancer and AIDS in the 1990s. Other health issues have appeared near the top of the list over the past 15 years, including 2014, when Ebola was listed among the top three health concerns. This was most likely due to multiple Ebola outbreaks in West Africa and a few confirmed cases all around the world, which prompted widespread media coverage of the disease. In 2015, less than 0.5 of Americans listed Ebola as the most urgent issue, as the threat of the virus has subsided. Similar short-lived spikes in the responses to this question have occurred regarding other recent health threats, including the H1N1/swine flu outbreak in 2009 and anthrax and bioterrorism attacks in 2001.

In our first scenario, we consider as application domain a global health treat like bioterrorism, so the system aim is to monitor the Twitter Stream detecting tweets

[1] http://www.gallup.com/

containing given keywords related to bioterrorism attacks, like “attack”, “terror”, “anthrax, “terrorism”, “jihadi”, etc.. In this scenario we focus our attention on gathering all the upcoming tweets about a particular event using Twitter’ streaming API. Depending on the search term, we can gather tons of tweets within a few minutes. This is especially true for live events with a world-wide coverage (World Cups, Academy Awards, Election Day, you name it). A working example that gathers all the new tweets with the #anthrax hashtag, or containing the word “anthrax”, shows that a large number of collected tweets is not related to an upcoming real treat. In our case, for example, the word “anthrax” is also related to a metal music band, so we need a technique to distinguish relevant tweets from not relevant ones.

As for the second example, we consider a given event as search domain to make a first selection of relevant tweets on which we adopt deeper textual analysis to detect the positive ones. A post hoc analysis of the information posted on Twitter, shows that, after the occurrence of a given event, the number of tweets related to that event is influenced by two main factors:

1. density of users in different geographic areas;
2. popularity of the event in the different geographic areas.

To argue this consideration we consider, as a simplified example, a sport event as given event; more specifically, we focus our attention to the soccer match between SSC Napoli and Legia Varsavia teams, played on 10 December 2015 during qualifying stages of Europa League competition. Before the match, tension, acts of vandalism and violence breaks out as opposite hooligans fight in city centre. The news was reported by main national newspapers, as well as main on-line information sites. Obviously this news was also reported by people and mass media on their Twitter profiles.

We analyse relevant tweet concerning this given event, to estimate user perception of the episodes of violence. Searching the string “Napoli Legia Varsavia” and filtering tweets by date between 9 December and 11 December, we collect the total amount of tweets concerning the given event. Refining the result with the words “arresti”, “scontri”, “polizia”, ect., we obtain the number of tweets related with the public security disorders, instead of the soccer match. In Table 1 we summary the results.

Table 1 Search results.

	Event related	Public security concern
Date of interest:	9 December – 11 December	
Query string:	Napoli Legia Varsavia	Napoli Legia Varsavia scontri, Napoli Legia Varsavia arresti, Napoli Legia Varsavia polizia
Number of tweets:	709	96

From this analysis we note that a considerable number of tweets is not related with the event itself, but rather with the occurrence of acts of vandalism and violence during the social event. Moreover, the proportion between total number of tweets

and the number of tweets concerning with public security issue depends by the severity of the episodes.

3 Framework description

In this work we present a general framework to be adopted for event detection from Twitter. In particular, we aim at detecting anomalies in Twitter stream related to malicious activity, as for example cyber-intrusion or terrorist activity. In this section

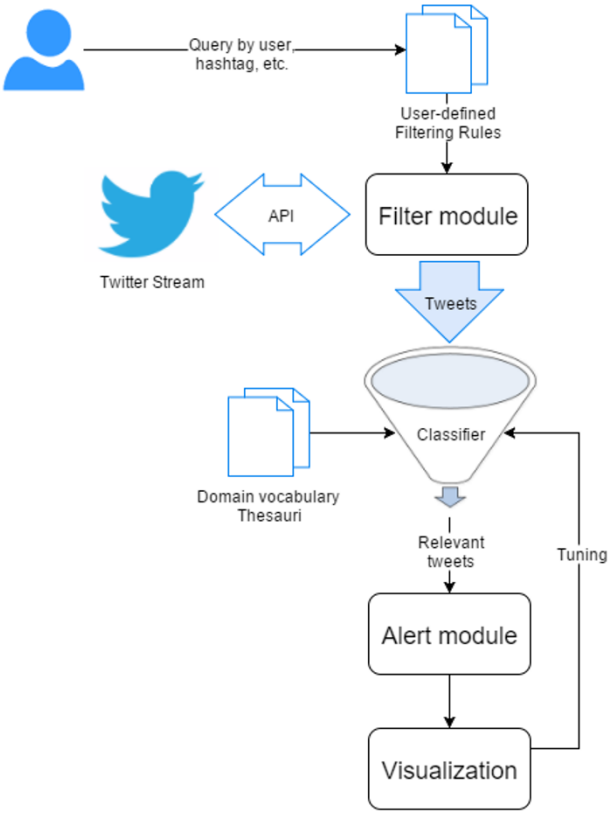


Fig. 1 Overview of the twitter stream processing framework for event detection and alerting.

we present an overview of the main processing stages of the framework along with the goals and requirements that are of importance for event detection from Twitter. Event detection from Twitter messages must efficiently and accurately filter relevant information about events of specific interest, which is hidden within a large amount of insignificant information. The proposed framework aims at analyzing tweets and automatically extract relevant event related information in order to raise a signal as

soon as a malicious event activity is detected. It is worth noting that in this work we consider an event occurrence in Twitter stream whenever an anomaly bursts in the analyzed data stream. The Twitter stream processing pipeline consists of three main stages, as illustrated in Figure 1. The system aims at detecting user defined phenomenon form the twitter stream in order to give alerts in case of event occurrence. In the following we describe the main stages of the proposed system.

3.1 Collection and Filtering

The first stage of the system is devoted to collect and filter tweets related to an event. It is based on Information extraction task. Information extraction is the process of automatically scanning text for information relevant to some interest, including extracting entities, relations, and, most challenging, events (something happened in particular place at particular time) [16]. It makes the information in the text more accessible for further processing. The increasing availability of on-line sources of information in the form of natural-language texts increased accessibility of textual information. The overwhelming quantity of available information has led to a strong interest in technology for processing text automatically in order to extract task-relevant information. Information extraction main task is to automatically extract structured information from unstructured and/or semi-structured documents exploiting different kinds of text analysis. Those are mostly related to techniques of Natural Language Processing (NLP) and to cross-disciplinary perspectives including Statistical and Computational Linguistics, whose objective is to study and analyze natural language and its functioning through computational tools and models. Moreover techniques of information extraction can be associated with text mining and semantic technologies activities in order to detect relevant concepts from textual data aiming at detecting events, indexing and retrieval of information as well as long term preservation issues [17],[18],[19]. Standard approaches used for implementing IE systems rely mostly on:

- Hand-written regular expressions. Hand-coded systems often rely on extensive lists of people, organizations, locations, and other entity types.
- Machine Learning (ML) based Systems. Hand annotated corpus is costly thus ML methods are used to automatically train an IE system to produce text annotation. Those systems are mostly based on supervised techniques to learn extraction patterns from plain or semi-structured texts. It is possible to distinguish two types of ML systems:
  - Classifier based. A part of manually annotated corpus is used to train the IE system in order to produce text annotation [20].
  - Active learning (or bootstrapping). In preparing for conventional supervised learning, one selects a corpus and annotates the entire corpus from beginning to end. The idea of active learning involves having the system select examples for the user to annotate which are likely to be informative

which are likely to improve the accuracy of the model [21]. Some examples of IE systems are [22].

In our work we are interested in detecting specified event which relies on specific information and features that are known about the event such as type and description, which are provided by the a domain expert of the event context. These features are exploited by adapting traditional information extraction techniques to the Twitter messages characteristics. In the first stage, the tweets are collected exploiting the REST API provided from Twitter that allow programmatic access to read data from the stream. Those API are costumed in order to crawl data exploiting user defined rules (keywords, hashtag, user profile, etc). The tweets are annotated, in the filter module, with locations and temporal expressions using a series of language processing tools for tokenization, part-of-speech tagging, temporal expression extraction, and tools for named entity recognition.

In order to access to Twitter data programmatically, it's necessary to register a trusted application, associated with an user account, that interacts with Twitter APIs. Registering a trusted application, Twitter provides the required credentials (consumer keys) that can be used to authenticate the REST calls via the OAuth protocol. To create and manage REST calls we integrate in the Filter Module a Python wrapper for the Twitter API, called Tweepy, that includes a set of class enabling the interaction between the system and Twitter Stream. For example, the API class provides access to the entire twitter RESTful API methods. Each method can accept various parameters and return responses.

To gather all the upcoming tweets about a particular event, we call the Streaming API extending Tweepy's StreamListener() class in order to customise the way we process the incoming data. In the Listing 1 we simply gather all the new tweets with the #anthrax hashtag:

Listing 1 Custom Listener to gather all the new tweets with the #python hashtag and save them in a file

```
class TwitterListener(StreamListener):

    def on_data(self, data):
        try:
            with open('anthrax.json', 'a') as f:
                f.write(data)
                return True
        except BaseException as e:
            print("Error on_data: %s" % str(e))
        return True

    def on_error(self, status):
        print(status)
        return True

twitter_stream = Stream(auth, TwitterListener())
```

```
twitter_stream.filter(track=['#anthrax'])
```

All gathered tweets will be furthermore filtered and then passed to the Classification module for deeper textual processing[23, 24, 25].

3.2 Classifying relevant information

The collected and annotated tweets are then classified, in the second stage, with respect to their relevance. As stated before in the paper, event detection from Twitter messages is a challenging task since the relevant information about a specific event is hidden within a large amount of insignificant messages. Thus a classifier must efficiently and accurately filter relevant information. The classifier is trained exploiting a domain dependent thesaurus that include a list of relevant terms for event detection. It is responsible for filtering out irrelevant messages.

3.3 Generating alerts

Once the set of irrelevant tweets are discarded, those remaining must be conveniently aggregated with respect to an event of interest. The third stage of the proposed system is aimed at detecting anomalies in the event related set of tweets. Anomalies are patterns in data that do not fit the pattern of the expected normal behavior. Anomalies might be induced in the data for a variety of reasons, such as malicious activity, as for example cyber-intrusion or terrorist activity, but all of the reasons have a common characteristic that they are interesting to the analyst. In literature there are several techniques exploited for the anomaly detection. Those range between several disciplines and approaches such as statistics, machine learning, data mining, information theory, spectral theory.

In our proposed system, the set of event related tweets are processed by the alert module that is responsible to raise an alert in case of anomaly within the collected tweets. The alert module implements a set of burst detection algorithms in order to generate alerts whenever an event occurs that means an anomaly in the analyzed data stream is detected. The detected events are then displayed to the end user exploiting the visualization module. An important aspect is that the analyzed data can be tuned by the end user for novelty detection which aims at detecting previously unobserved (emergent, novel) patterns in the data, such as a new topic. The novel patterns are novel patterns are typically incorporated into the normal model after being detected. Moreover with the tuning phase the parameters for the tweets classification module can be modified in case the messages that raised an alert are considered not relevant for the event occurrence.

4 Conclusions and Future Work

Social networking services are matter of interest for research activities in several fields and applications. The number of active users of social networking services like Twitter raised up to 320 million per month in 2015. The rich knowledge that has accumulated in the social sites enables to catch the reflection of real world events[26, 27].

In this work we proposed a general framework for event detection, starting from the processing of heterogeneous data coming from social networks systems. The framework aims at detecting malicious events in Twitter communities, exploiting techniques related to heterogeneous data analysis and information classification. Once the gathered data are classified as related to criminal activity, they are processed by the alert module that is responsible to raise an alert in case of anomaly within the collected information. The alert module implements a set of burst detection algorithms in order to generate alerts whenever an event occurs that means an anomaly in the analyzed data stream is detected.

As future work we intend to provide the implementation of our system tuned with a domain related to the bio-terrorism.

References

1. N. Diakopoulos, M. Naaman, and F. Kivran-Swaine, "Diamonds in the rough: Social media visual analytics for journalistic inquiry," in *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*, Oct 2010, pp. 115–122.
2. S. Yardi and danah boyd, "Tweeting from the town square: Measuring geographic local networks," in *International Conference on Weblogs and Social Media*. American Association for Artificial Intelligence, May 2010. [Online]. Available: http://research.microsoft.com/apps/pubs/default.aspx?id=122433
3. T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: Real-time event detection by social sensors," in *Proceedings of the 19th International Conference on World Wide Web*, ser. WWW '10. New York, NY, USA: ACM, 2010, pp. 851–860. [Online]. Available: http://doi.acm.org/10.1145/1772690.1772777
4. M. Naaman, J. Boase, and C.-H. Lai, "Is it really about me?: Message content in social awareness streams," in *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, ser. CSCW '10. New York, NY, USA: ACM, 2010, pp. 189–192.
5. H. Becker, M. Naaman, and L. Gravano, "Beyond trending topics: Real-world event identification on twitter," 2011.
6. ——, "Learning similarity metrics for event identification in social media," in *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, ser. WSDM '10. New York, NY, USA: ACM, 2010, pp. 291–300.
7. F. Atefeh and W. Khreich, "A survey of techniques for event detection in twitter," *Comput. Intell.*, vol. 31, no. 1, pp. 132–164, Feb. 2015. [Online]. Available: http://dx.doi.org/10.1111/coin.12017
8. K. Essmaeel, L. Gallo, E. Damiani, G. De Pietro, and A. Dipanda, "Comparative evaluation of methods for filtering kinect depth data," *Multimedia Tools and Applications*, vol. 74, no. 17, pp. 7331–7354, 2015.

9. E. Aramaki, S. Maskawa, and M. Morita, "Twitter catches the flu: Detecting influenza epidemics using twitter," in *Proceedings of Conference on Empirical Methods in Natural Language Processing*, 2011.
10. A. Lamb, M. J. Paul, and M. Dredze, "Separating fact from fear: Tracking flu infections on twitter." in *HLT-NAACL*, 2013, pp. 789–795.
11. J. Chon, R. Raymond, H. Wang, and F. Wang, "Modeling flu trends with real-time geo-tagged twitter data streams," in *Wireless Algorithms, Systems, and Applications*. Springer, 2015, pp. 60–69.
12. E. Diaz-Aviles and A. Stewart, "Tracking twitter for epidemic intelligence: Case study: Ehec/hus outbreak in germany, 2011," in *Proceedings of the 4th Annual ACM Web Science Conference*, ser. WebSci '12. New York, NY, USA: ACM, 2012, pp. 82–85. [Online]. Available: http://doi.acm.org/10.1145/2380718.2380730
13. R. Chunara, J. R. Andrews, and J. S. Brownstein, "Social and news media enable estimation of epidemiological patterns early in the 2010 haitian cholera outbreak," *The American Journal of Tropical Medicine and Hygiene*, vol. 86, no. 1, pp. 39–45, 2012. [Online]. Available: http://www.ajtmh.org/content/86/1/39.abstract
14. J. Gomide, A. Veloso, W. Meira, V. Almeida, F. Benevenuto, F. Ferraz, and M. Teixeira, "Dengue surveillance based on a computational model of spatio-temporal locality of twitter," in *Proceedings of ACM WebSci'2011*, 2011.
15. V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 15:1–15:58, Jul. 2009.
16. J. R. Hobbs and E. Riloff, "Information extraction," in *Handbook of Natural Language Processing, Second Edition*, N. Indurkhya and F. J. Damerau, Eds. Boca Raton, FL: CRC Press, Taylor and Francis Group, 2010.
17. F. Amato, A. Mazzeo, A. Penta, and A. Picariello, "Using NLP and Ontologies for Notary Document Management Systems," in *Proceedings of the 19th International Conference on Database and Expert Systems Application*, 2008.
18. F. Amato, V. Casola, N. Mazzocca, and S. Romano, "A semantic approach for fine-grain access control of e-health documents," *Logic Journal of the IGPL*, vol. 21, no. 4, pp. 692–701, 2013.
19. F. Amato, A. Fasolino, A. Mazzeo, V. Moscato, A. Picariello, S. Romano, and P. Tramontana, "Ensuring semantic interoperability for e-health applications," 2011, pp. 315–320.
20. E. Riloff, "Automatically constructing a dictionary for information extraction tasks," in *Proceedings of the eleventh national conference on Artificial intelligence*, 1993.
21. R. Grishman, "Information extraction: Capabilities and challenges," 2012.
22. P. Pantel and M. Pennacchiotti, "Espresso: leveraging generic patterns for automatically harvesting semantic relations," in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, 2006.
23. M. Barbareschi, E. Battista, N. Mazzocca, and S. Venkatesan, "A hardware accelerator for data classification within the sensing infrastructure," in *Information Reuse and Integration (IRI), 2014 IEEE 15th International Conference on*. IEEE, 2014, pp. 400–405.
24. M. Barbareschi, "Implementing hardware decision tree prediction: a scalable approach," in *2016 30th International Conference on Advanced Information Networking and Applications Workshops (WAINA)*. IEEE, 2016, pp. 87–92.
25. M. Barbareschi, A. De Benedictis, A. Mazzeo, and A. Vespoli, "Providing mobile traffic analysis as-a-service: Design of a service-based infrastructure to offer high-accuracy traffic classifiers based on hardware accelerators," *Journal of Digital Information Management*, vol. 13, no. 4, p. 257, 2015.
26. G. Sannino, P. Melillo, G. De Pietro, S. Stranges, and L. Pecchia, *To What Extent It Is Possible to Predict Falls due to Standing Hypotension by Using HRV and Wearable Devices? Study Design and Preliminary Results from a Proof-of-Concept Study*. Cham: Springer International Publishing, 2014, pp. 167–170.
27. N. Brancati, G. Caggianese, M. Frucci, L. Gallo, and P. Neroni, "Touchless target selection techniques for wearable augmented reality systems," in *Intelligent Interactive Multimedia Systems and Services*. Springer, 2015, pp. 1–9.

Modeling Approach for specialist domain

F. Amato, G. Cozzolino, F. Moscato, V. Moscato, A. Picariello

Abstract In the last two decades data analytics applications have been studied fairly extensively in the general context of analyzing business data, interactions between people through social networks analysis, marketing trends and so on. The increasing ability to capture and handle geographic data means that spatial analysis is occurring within increasingly data-rich environments. Although GIS provide platforms for managing these data, many of this applications require the integration or the cooperation between multiple information systems in order to satisfy the users' needs with accurate, not-ambiguous and not-redundant information. In this paper we present a methodology for modeling a domain of interest through semantic Web technologies, and we show a method to query the produced domain model in order to retrieve pertinent information.

1 Introduction

A tourist who visit a big metropolis have to face a number of difficulties, not only dictated by the chaos of the city, but above all by the enormous amount of places not to be missed, and therefore he would risk to lose many of them. In this work we present a system with the aim of creating an interactive and easy-to-use tourist guide to provide a reference point for anyone who is interested in visiting the main cultural sites of a big city. The intention is to offer indications for the user in a quick and intuitive way: therefore a set of basic services is guaranteed in order to facilitate the planning of the visit. Among these, we propose to indicate the main

Flora Amato, Giovanni Cozzolino, Vincenzo Moscato and Antonio Picariello
DIETI - Universit degli Studi di Napoli "Federico II", via Claudio 21, Naples, e-mail: flora.amato, giovanni.cozzolino, vmoscato, picus@unina.it

Francesco Moscato
Second University of Naples, Dipartimento di Scienze Politiche, Caserta, Italy, e-mail: francesco.moscato@unina2.it

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_66

eateries in a selected area and provide a suggestion of the nearest public transport in order to facilitate the travel from one point to another one. The project have been developed through the use of maps and utilities provided by the open source application QGIS, which allows the processing and manipulation of geometric geo-referenced data[1]. In this way the system is able to produce, manage and analyse spatial data, associating each geographical feature with images and text descriptions for each point of interest.

2 Goals

During the creation of our multimedia system we have set several goals to achieve and ensure:

- Support to the modelling of virtual objects: vector data (points, lines and polygons), saved as a shape-file are easily formable in order to compose a more faithful representation of the reality of interest.
- Support to the fruition of multimedia data and management of continuous media: through the actions implemented in QGIS, the user has to do just a click for to perform a web search or a consultation on TripAdvisor, a search on Wikipedia (in Italian or English language), watch videos on YouTube or even real-time webcam distributed in the city of Naples.
- High storage capacity: thanks to the support of a multimedia database located on a server, we can store a large number of data and images, without depending from the storage space of client machine.
- Information retrieval capability: the PostgreSQL spatial database allows you to make both simple query and spatial query to retrieve relevant information.
- High performance: the memorization of data into PostgreSQL ensures to load maps and shape-files faster than with a static load of QGIS. Besides, the choice of images in jpeg format, and then a smaller size than other formats, allows a faster fruition of multimedia data.
- Functionality for integration and presentation of multimedia data: data will be integrated and presented via website and Android app light and effective, it serves a continuous interfacing with the web site, through which they will be displayed immediately.

3 Quantum GIS

QGIS is an open source geographic information system easy to use, but it presents all the main functionalities of a GIS environment. Its basic characteristics, together with useful plug-in and the immediate interfacing with PostgreSQL and PostGIS, make it a cutting-edge tool that offers great potentialities.

3.1 Purpose and funtionalities

QGIS offers the possibility to create digital maps, creating spatial information to add to the map and manage and analyse all kinds of data associated to them, which are divided into vector and raster:

- **vector data**: stored as a series of coordinates (X, Y); They are used to represent points, lines and polygons.
- **raster data**: stored as orthogonal grid of points which forms a raster image.

The maps are composed by levels, or layers: in this way a different vector will be possible to associate for each layer, and then superimpose them to create a map with more substantial information. Its also very useful the function that allows you to hide or show all the features of that layer with a single mouse click. The peculiarity of GIS is to combine the information (non-geographic data) to places (geographical data). Therefore, each vector has an associated database containing descriptive data of each cell. The database that keeps all the vectors information is called shape-file. For this project we used the PostGIS vectors stored in a PostgreSQL database. PostGIS has several advantages: spatial indexing, filtering capabilities and very powerful queries. Besides using PostGIS, vector functions such as select and identify are more precise than with OGR vectors in QGIS.

4 Semantic Web

To make accessible for everyone the information stored in QGIS, we have exploited the potential of the semantic web technologies. This kind of data representation, due to the large amount of information, give a better solution than a classic representation in HTML pages to the problems of information retrieval and speed-up the extraction of relevant information. In fact, the information contained in web pages cant be automatically processed by the search engines like Yahoo, Google or Altervista but require human intervention.

In traditional web, the information is represented using a natural language and multimedia elements. Often when you do a search, you need to combine information from different sources or searching the desired information in a particular page. For a human these operations are achievable but not for the machines: computers are not able to use partial information or aggregate structured information in a different form. Semantic Web allows us to represent information in a manner that they are automatically processable by machines, using ”intelligent” techniques. It is an extension of the current Web in which information is structured with full meaning, thus making easier the interaction between man and computer. What differentiate it from current web is the dynamism of the information that become automated[2, 3]. The Semantic Web uses:

- Metadata to identify and extract information

- Ontologies to interpret them
- Logical Inferences to process them
- Agents for the automatic management of knowledge

4.1 Ontologies

Ontology, in Computer Science, is a formal representation, shared and explicit of a conceptualization relative to a domain of interest, through:

- Concepts, or object classes
- Relations between classes
- Properties between classes (domain and range)
- Restrictions on class and property claims on instances of classes

Ontologies are frequently used into information integration processes, because represent an effective solution to design a data integration layer in which several schemes are combined into a data structure containing all the relevant entities and the relationships between them in a specific domain. Such kind of ontologies, then, can be used by different modules of an information system for various purposes as inductive reasoning, classification and different problem solving techniques as well as to realize interoperation between different systems. The purpose of an ontology is to make understandable information both for the software and humans. Because an ontology to be valid must possess the following characteristics:

- Completeness: anticipate all key distinctions;
- Conciseness: Don't make assumptions implicit and redundant;
- Consistency: not contain definitions in contradiction. A definition is inconsistent when the expression of other definitions of ontology contradict it ;
- Coherence: allow all and sunny relations consistent with the definition of concepts;
- Modifiability;
- Re-usability/scalability: to extend the ontology without extending the definition interests already exist.

5 System description

To describe formally the domain of interest of our system we will provide an ontology based on Protg. Our knowledge base will also be available in a Prolog application. The method adopted to realize the proposed system aims to translate all the components that constitute the QGIS shape-files. All the layers constituting QGIS project map have been reported in Protégé as classes of our ontology, so in similar way, all the point or entities of each layer constitute the individuals of our ontology. During the development of the ontology, classes coming from layer translation

have been rearranged in a more effective hierarchy that considers all the possible relationship between the elements of our domain of interest: for example, museums, churches, attractions are all subclasses of a single class called sites of interest; restaurants and pizzerias are subclasses of ”eatery”, and so on. Other relationship in the QGIS project, not merely hierarchical one, are expressed by object properties, and additional attributes of each individual are listed by data properties.

5.1 Classes

Let’s present the ontology born from these premises:

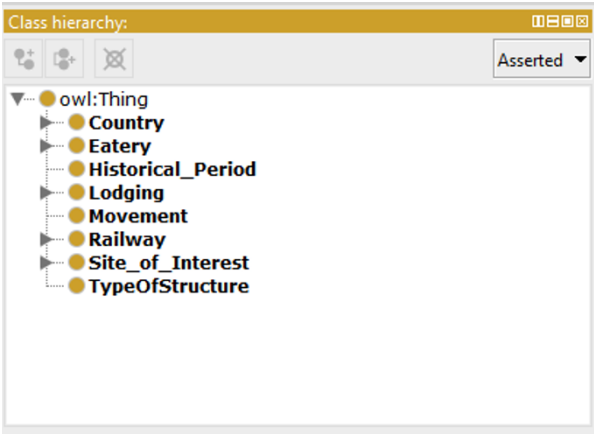
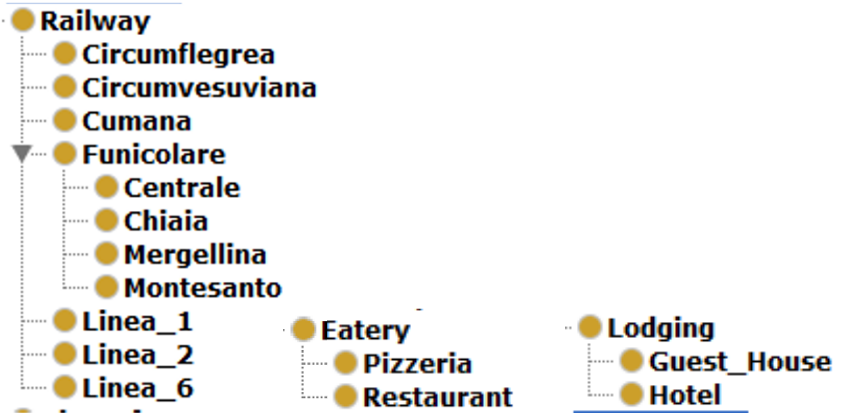


Fig. 1: Top level classes of the ontology.

Our classes are:

- SiteofInterest divided into subclasses containing various instances by categories such as Theatre, Museum, and Church,which are the backbone of our project. Each class contains the most important points of cultural interest of the city.
- Railway divided into subclasses representing the different railway lines in the city, each class contains as instances all the stops of the railway line which represents.
- Eatery divided into two subclasses, Restaurant and Pizzeria which contain as instances the most important restaurants and the most important pizzerias in the city.
- Lodging divided into Guest House and Hotel which represent the main housing of the city.
- Country/Region/District/City through which we will identify the location of our sites by specifying not only the city but also the district, region and the state to which they belong.



(a) Railway. (b) Eatery. (c) Lodging.

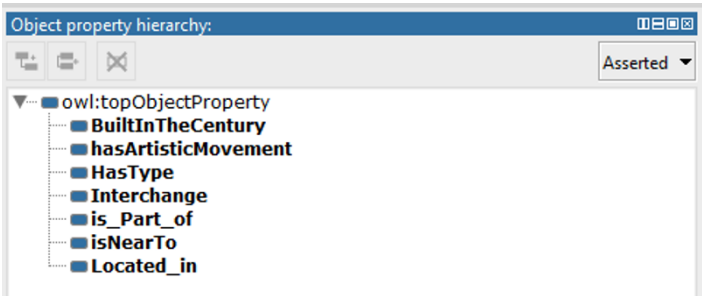
Fig. 2: Detail of top level classes taxonomy.

- HistoricalPeriod where we note the different centuries that will be used to indicate the period of construction of each Church.
- Movement class that represents the artistic styles of our Site of Interest .
- TypeOfStructure identifies the type of buildings that we will present.

5.2 Relations

In OWL the relations are properties and are divided into two types:

- Object property: the relationships (predicates) between two objects in an OWL ontology.



- Data property: are the relationships that link the attributes of individuals to the types allowed.

5.3 Individuals

For each class we have to insert the instances belonging to the class. Once created the instance we can associate the object and data property to it.

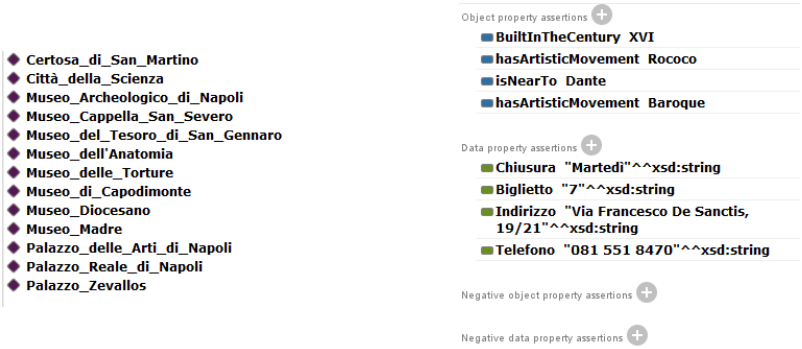


Fig. 3: Instances and Data Object Proprieties

6 Testing the ontology

The peculiarity of the Semantic Web is that the information it contains is understandable by applications as well as by humans. The Semantic Web in fact lends itself to questions usually are made in SPARQL. It is a query language based on RDF and has a syntax very similar to that of SQL. Extracts information from distributed knowledge bases on the web. Being triple-based RDF subject-predicate-object enables you to build queries based on the same triples. Such a query must have a well defined structure:

PREFIX assign a prefix to uri used in queries.

SELECT provides the list of variables to appear in the result (any variable must be preceded by the symbol ?

FROM indicates the data source to be queried.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT ?subject ?object
        WHERE { ?subject rdfs:subClassOf ?object }
```

WHERE defines the constraints on the triple to be extracted through a GRAPH Pattern to unify the graph being queried.

In Protégé we can make also DL Query, based on the Manchester OWL syntax is fundamentally based on collecting all information about a particular class, property, or individual into a single construct, called a frame.

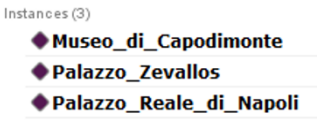
6.1 Query in Protégé

We tested the ontology performing two types of queries: DL and SPARQL query.

- **DL query:**
  Baroque and Renaissance Museum

```
(is_Part_of some Museum) and
   (          (hasArtisticMovement value Baroque) and
              (hasArtisticMovement value Renaissance)
)
```

Querying (making sure to tick Instances from the menu) we will have as a result the museums containing Baroque and Renaissance works. The answer of our ontology will be:



Nearest neighbor Query

```
(is_Part_of some Museum)and
   ( (isNearTo value Museo)
)
```

Through this query we can find the Museums near to underground station, we also carried out the same type of query in QGIS using buffers and range query. The result of this query will be

- **SPARQL query**
  Hotels with more than 3 stars
  Query that shows all the hotels that have more than 3 stars.

```
SELECT ?HOTEL
WHERE {
        ?HOTEL:Stelle ?Stelle.
FILTER                  (?Stelle > 3).
}
```

The answer is:

```
Grand_Hotel_Capodimonte
Hotel_Caravaggio
Grand_Hotel_Vesuvio
Partenope_Relais
Ramada_Naples
Grand_Hotel_Parker's
Hotel_Decumani
Holiday_Inn_Napoli
Royal_Continental_Hotel
Romeo_Hotel
Eurostars_Hotel_Excelsior
```

Pizzerias at Via dei Tribunali
Query that shows all the pizzerias that address Via dei Tribunali

The answer is:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX  : <http://www.semanticweb.org/claudio/ontologies/2016/7/ia#>

SELECT        ?Pizzerie
WHERE {
 ?Pizzerie :Indirizzo ?Indirizzo.
FILTER        regex( ?Indirizzo, "Tribunali")
  }
```

```
Antonio_e_Gigi_Sorbillo
Dal_Presidente
Di_Matteo
```

7 Conclusions and future works

Data analytics applications have been studied fairly extensively in the general context of analyzing business data, interactions between people through social networks analysis, marketing trends and so on. The increasing ability to capture and handle geographic data means that also spatial analysis is occurring within increasingly data-rich environments.

Many of data analytic applications require the integration of different sources of data, or the cooperation between multiple information systems in order to satisfy the users' needs with accurate, not-ambiguous and not-redundant information.

In this paper we presented a methodology for modeling a domain of interest through semantic Web technologies, and exploited a method to query the produced domain model in order to retrieve pertinent information.

Future works will be devoted to extend the model. We are going to design and implement a more user friendly system for managing the model and retrieving information from it.

References

1. S. Cuomo, P. De Michele, A. Galletti, F. Pane, and t. d. u. y. Giovanni Ponti, pages=337-343, in *DATA 2015 - Proceedings of 4th International Conference on Data Management Technologies and Applications, Colmar, Alsace, France, 20-22 July, 2015*.
2. F. Amato, A. Mazzeo, V. Moscato, and A. Picariello, "A framework for semantic interoperability over the cloud," 2013, pp. 1259–1264.
3. F. Amato, A. Mazzeo, A. Penta, and A. Picariello, "Building rdf ontologies from semi-structured legal documents," 2008, pp. 997–1002.

Designing a Service Oriented System for social analytics

Angelo Chianese , Paolo Benedusi and Francesco Piccialli

Abstract Event detection in social networks and, more in general, for large-scale Social Analytics, requires continuous monitoring and processing of huge volumes of data. In this paper we explored effective ways to leverage the opportunities coming from innovations and evolutions in computational power, storage, and infrastructures, with particular focus on modern architectures. In particular, we position a specific core technology and cloud platform adopted in our research with respect to micro-services. Furthermore, we demonstrate the application of some of its capabilities for performing data-intensive computations and implementing services in a real case of social analytics. A prototype of this system was experimented in the contest of a specific kind of social event, an art exhibition of sculptures, where the system collected and analyzed in real-time the tweets issued in an entire region, including exhibition sites, and continuously updated analytical dashboards placed in one of the exhibition rooms.

1 Introduction

Social Networking activities generate huge volumes of data, whose analysis is becoming more and more important for both profit and non-profit organizations; at the same time, social networks are used in proactive manners by organizations and even individuals for multiple objectives, including internal and external communication, coordination, information, recommendation, advertising, promotion campaigns, and organization of social events [6, 13]. The impact, efficacy and efficiency of these activities can be monitored and evaluated also by analysing the evolution of social network data, in combination with other sources of information on people

Angelo Chianese and Francesco Piccialli
University of Naples "Federico II", Italy, e-mail: {angchian,francesco.piccialli}@unina.it

Paolo Benedusi
In.Tel.Tec spa, Italy, e-mail: paolo.benedusi@inteltec.it

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_67

behaviours. In particular, the Twitter microblog features provide great opportunities and pose specific challenges for large-scale analytics, which stem from multiple aspects: real-time nature of messages, often based on volatile emotional reactions; fast viral propagation over the network; high peaks of interest concentrated in short timeframes; short lifetime of many hashtags; short and noisy text including non textual contents images, emoticons, other symbols, and so on. In this context, the need of advanced services requires evolutions in multiple aspects of Service Oriented Systems Engineering. One of the most important aspects is how to effectively leverage the opportunities coming from innovations and evolutions in computational power, storage, and infrastructures; our research is focused on modern architectures including in-memory database technology, in-database computation, MPP(Massive Parallel Processing), Open Data Services, scalability with multi-node clusters in Cloud infrastructures, with particular respect to the SAP HANA Cloud Platform(HCP). The main scenario of this research is supporting the objectives and activities of DATABENC, a High Tech Consortium for Cultural Assets in Italy, involving multiple Universities, cultural organizations and industries located in the Campania region. The rest of this article is organized as follows:

2 Related Work

In the last five years, the research in Service Oriented Computing has evolved in multiple directions and areas. In [1] a flexible service oriented architecture is proposed for planning and decision support in environmental crisis management, which includes real time middleware that is responsible for selecting the most appropriate method of the available geo-spatial content and service, in order to satisfy the QoS requirements of users and/or application. A conceptual model that supports evaluation of theory and methods for the management of service-oriented decision support systems is proposed in [2], which distinguishes Data-as-a-service (DaaS), Information-as-a-service (information on demand), Models-as-services (MaaS), Analytics-as-a-service (AaaS); this study emphasizes the importance of social media analytics, Continuous Analytics as a service, and the need of a shared infrastructure, with local processing power to perform real-time data analysis, in-memory grid computations and parallel transaction and event processing. In [3] an analytics service-oriented architecture (ASOA) is presented, based on ontology of business analytics. A recent study analyses the reasons behind low service reuse [4], highlighting that the main identified enabler for reusable services was the capability to define the services as a part of reusable business concepts, while the main reason for having overlapping services in the case enterprise was lagging migration of legacy services to use newer, reusable services. In [5] a survey on technologies for Web and cloud service interaction pinpoints that composition of RESTful services primarily takes place in Web mashups, and indicates the convergence of the Web-oriented view (which highlights the evolution of the Web and

technological advancements, e.g., AJAX, HTML5, syndication, and mashups) and the service-oriented view.

3 The Service System

3.1 An Overview

The S-InTime system (see 1) is conceived for fast integration and analysis of data coming from multiple domains, including catalogues and other semantic resources on cultural assets managed by both non-profit and profit organizations, their transactional data, as well as data gathered from Internet of Things [9, 14, 7, 8, 11] and Social Media.

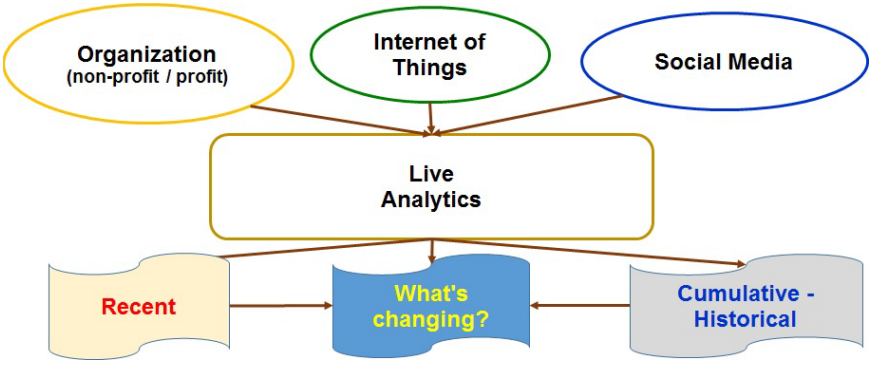


Fig. 1 System Overview.

The concept of live analytics entails not only the ability to support highly interactive analytical services with responsive user interfaces, but also real-time data processing with the capability of producing early-timely advices on new phenomena, which can be relevant for the objectives of one or more types of users. For this purpose, the system compares information obtained from cumulative-historical analysis over entire periods with information caught in recent / incoming data. In the following paragraphs, we explain how we addressed the main requirements of this kind of system by exploiting the opportunities and the peculiar architectural paradigms supported by HCP, with examples of Services for Social Networking data.

3.2 View-based Information Modeling and computing

For the real-time processing of huge quantities of data, especially for analytical purposes, it is necessary to exploit advanced features of the in-memory database system: non-materialized dynamic views. This is often to work around the fact that the platform and disk-based data access will bind I/O operations and simply prove negative performance implications when performing complex aggregations or queries across larger data sets. The HANA View-based Information Modeling and computing process utilizes SQLscript and stored procedures to perform business logic on the data models. From a design and continuous delivery perspective, the flexible data modeling on the fly liberates users from the traditional data-modeling bottleneck. These views enable the business to react more quickly to changing customer and market environments. HANA Views are subdivided into three main categories: Attribute, Analytical, and Calculation Views (see Figure 2).

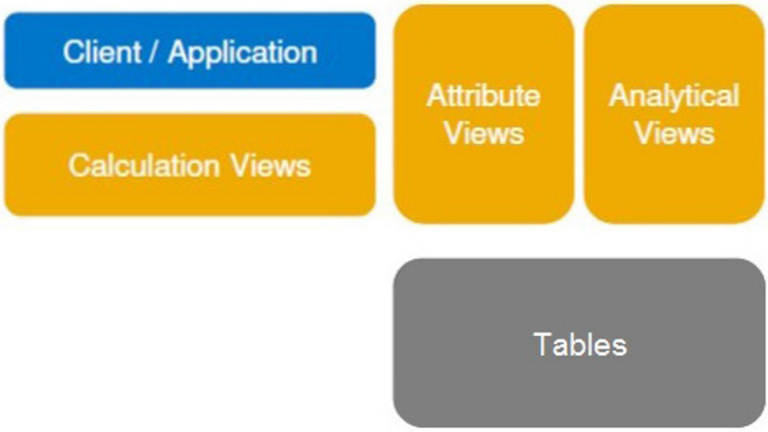


Fig. 2 Hana Views.

- Attribute views model an entity based upon one table or relationships of multiple source tables, and can contain columns, calculated columns, as well as hierarchies.
- Analytical Views leverage the computing power of SAP HANA to calculate aggregate data. Analytical views are the SAP HANA multidimensional modeling components that allow to define both measures and their relationships between one or more attributes, with the primary goal of defining a logical Star Schema.
- Calculation Views are used to create custom data sets to address complex business requirements, which cannot be obtained in the other kinds of views.

4 Preliminary Experimentation

We applied the architectural and programming principles of the S-InTime in the implementation of a set of Twitter Analytics services, which cover the following aspects:

- Analysis in Semantic Domains
  - Cultural Heritage Sensitivity
  - Presence of CH-related terms in messages
  - General CH-related terms vs Exhibition-specific terms
  - Sentiments in sensitive tweets

Fig. 3 Geo-referenced CH-related tweets

Figure 3 shows an example of geo-referencing CH-related tweets issued in the surroundings of the Castel Nuovo Smart Tour Exhibition (Last Hour snapshot). We created live demonstrators of the S-InTime system, using SAPUI5 to create a set of multiple Cloud-web-based dashboards acting as consumers of all the services in multiple types of clients and scenarios: mobile devices, desktop interactive, and unattended demo on a wide screen, where the various web pages were cyclically presented in a predefined sequence. This demonstrator was installed in a permanent exhibition located in Naples, Italy, entitled CastelNuovo Smart Tour [10]. This allowed a sort of stress social test: visitors had the opportunity of looking at the analytical dashboards in the context of the same room where sculptures were exposed, as well as to issue tweets in the context of their visits. The real-time performances of the HANA Cloud platform allowed to intercept cases when a visitor issued a message containing an unprecedented CH-related or Exhibition-related term, causing new entry popups and updates in the dashboard in such a short time, that the visitor could appreciate the impact of his/her social action on the screen before moving

elsewhere. The system focused the analysis on tweets issued in Campania Region, collecting more than 1,1 millions of messages from the twitter streaming API along a period of 5+ months. This sample included more than 50,000 distinct tweet users and 3,600 hashtags.

5 Conclusions

Finding effective ways to leverage the potential of platforms such as HCP, especially for microservices, is an open and active area of study and experimentation. For instance, Architecture Guidelines for Microservices Based on SAP HANA Cloud Platform are proposed in [16, 12], and an exemplary blueprint for a microservice architecture including SAP HANA, including sample code-base, is publicly available [15]. In our experience we can state that a microservice template can embrace the database, views, OData services and corresponding analytical objects of one simple web dashboard, based on the concatenation of the schemas. We exposed and experimented basic strategies for information partitioning and aggregation in views, and for mapping views to different types of OData Service endpoints.

References

1. V.Vescoukis, N.Doulamis, S. Karagiorgou, A service oriented architecture for decision support systems in environmental crisis management, Future Generation Computer Systems, Volume 28, Issue 3, March 2012, pp 593-604 doi:10.1016/j.future.2011.03.010
2. H. Demirkan, D. Delen, Leveraging the capabilities of service-oriented decision support systems: Putting analytics and big data in cloud, Decision Support Systems, Volume 55, Issue 1, April 2013, Pages 412-421, doi:10.1016/j.dss.2012.05.048
3. Z.Sun, K.D. Strang, J.Yearwood, Analytics service oriented architecture for enterprise information systems, iiWAS '14 Proceedings of the 16th International Conference on Information Integration and Web-based Applications & Services, Pages 508-516, ACM New York, NY, USA 2014, doi:10.1145/2684200.2684358
4. V.Alkkiomaki, K.Smolander, Anatomy of one service-oriented architecture implementation and reasons behind low service reuse, Service Oriented Computing and Applications June 2016, Volume 10, Issue 2, pp 207-220, DOI 10.1007/s11761-015-0181-2
5. H.Lampesberger, Technologies for Web and cloud service interaction: a survey, Service oriented computing and applications Springer, June 2016, Volume 10, Issue 2, pp 71-110, DOI 10.1007/s11761-015-0174-1
6. Duc T. Nguyen, Jai E. Jung, Real-time event detection for online behavioral analysis of big social data Future Generation Computer Systems, http://dx.doi.org/10.1016/j.future.2016.04.012, 2016
7. A. Chianese, F. Piccialli and I. Valente: Smart environments and Cultural Heritage: a novel approach to create intelligent cultural spaces Journal of Location Based Services, pp. 209-334, 2015
8. F. Amato, A. Chianese, A. Mazzeo, V. Moscato, A. Picariello, and F. Piccialli: The talking museum project, Procedia Computer Science, vol. 21, no. 0, pp. 114-121, 2013

9. Chianese, A., Piccialli, F., A smart system to manage the context evolution in the Cultural Heritage domain, Computers & Electrical Engineering, 2016, doi:10.1016/j.compeleceng.2016.02.008
10. S. Cuomo, P. De Michele, F. Piccialli, A. Galletti: A Cultural Heritage case study of visitor experiences shared on a Social Network Proc. of 10th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing, 3PGCIC, pp. 539-544, (2015)
11. Amato, Flora and De Pietro, Giuseppe and Esposito, Massimo and Mazzocca, Nicola An integrated framework for securing semi-structured health records Knowledge-Based Systems, vol. 79, pp. 99-117, 2015
12. Amato, Flora and Barbareschi, Mario and Casola, Valentina and Mazzeo, Antonino An fpga-based smart classifier for decision support systems Intelligent Distributed Computing VII, pp. 289-299, 2014
13. Amato, Flora and Moscato, Francesco A model driven approach to data privacy verification in E-Health systems Transactions on Data Privacy, vol. 8, pp. 273-296, 2015
14. Chianese, A., Piccialli, F. and Riccio, G, Designing a smart multisensor framework based on beaglebone black board, Lecture Notes in Electrical Engineering, Vol. 330, pp-391-397, 2015
15. Matthias Steiner, Microservices on SAP HANA Cloud Platform, http://scn.sap.com/community/developer-center/cloud-platform/blog/2015/11/24/microservices-on-hcp
16. DEV300: Architecture Guidelines for Microservices Based on SAP HANA Cloud Platform http://www.slideshare.net/saphcp/dev300-architecture-guidelines-for-microservices-based-on-sap-hana-cloud-platform, SAP TechEd Nov 24, 2015

GPU Profiling of Singular Value Decomposition in OLPCA Method for Image Denoising

Salvatore Cuomo, Pasquale De Michele, Francesco Maiorano and Livia Marcellino

Abstract We focus on the Graphic Processor Unit (GPU) profiling of the Singular Value Decomposition (SVD) that is a basic task of the Overcomplete Local Principal Component Analysis (OLPCA) method. More in detail, we investigate the impact of the SVD on the OLPCA algorithm for the Magnetic Resonance Imaging (MRI) denoising application. We have resorted several parallel approaches based on scientific libraries in order to investigate the heavy computational complexity of the algorithm. The GPU implementation is based on two specific libraries: NVIDIA cuBLAS and CULA, in order to compare them. Our results show how the GPU library based solution could be adopted for improving the performance of same tasks in a denoising algorithm.

1 Introduction

Numerous research problems, in order to be resolved, require huge computational resources [2, 3, 4]. Image denoising is a common expensive computational task in the research field of Magnetic Resonance Imaging (MRI). Here, well-known denoising algorithm are Non-local Means (NLM) [6, 10, 17], Adaptive Non-local Means (ANLM) [15], Principal Component Analysis (PCA) [7, 18] and Local PCA (LPCA) [16]. Respect to data managed, the format of Diffusion

Salvatore Cuomo, Pasquale De Michele and Francesco Maiorano
University of Naples "Federico II", Strada Vicinale Cupa Cintia 21, 80126, Naples, Italy,
e-mail: salvatore.cuomo@unina.it,
e-mail: pasquale.demichele@unina.it,
e-mail: frmaiorano@gmail.com

Livia Marcellino
Dept. of Science and Technology
University of Naples "Parthenope", Centro Direzionale Isola C3, 80143, Naples, Italy
e-mail: livia.marcellino@uniparthenope.it

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_68

Weighted (DW) images normally shows a low Signal to Noise Ratio (SNR) due to the presence of noise from the measurement process that complicates and biases the estimation of quantitative diffusion parameters [14].

In order to achieve smart parallel implementations, it is necessary to demand the heavy computational kernels of the OLPCA to optimized standard scientific libraries, see [11]. According to this, here we highlight that the more expensive linear algebra task is the Singular Value Decomposition (SVD). More in detail, SVD is a numerical technique to split a matrix into a set of linearly independent components [5]. On the other hand, the SVD is an attractive algebraic transform for image processing and it represents a foundation routine for many image processing applications. Because of its endless advantages The SVD is usually used in compression, as well as, in noise filtering [13, 19].

In a previous work [8], we resort to the Overcomplete LPCA (OLPCA) method discussed in [14], where for each point of an image, all the local estimates are combined in order to obtain the overlapping of little portions of the image. The SVD method is a very hard task for its high computational costs, but it is a crucial step for denoising process. The huge amount of computational resources demanded by this method has been recently addressed by using accelerated hardware as the Graphic Processor Units (GPUs) [11, 12, 9], with good performance in terms of execution time. GPUs are massively parallel architectures that efficiently work with impressive performance improvements. In this paper we choose to parallelize some basic computational tasks by using Compute Unified Device Architecture (CUDA) [1] on a GPU architecture. Here, our aim is to perform an SVD profiling in GPU-CUDA environment in order analyse how to speed up this computational task that is the bottleneck of the algorithm. We have investigated about using of the optimized routines of the CUBLAS and CULA library.

The plan of the paper is as follows. In Section 2 we give a short description of the OLPCA algorithm. Then, in Section 3, we provide implementation details. Moreover, the results of our work are reported and discussed in Section 4. Finally, in Section 5 we draw conclusions.

2 Preliminaries

The OLPCA method is used to perform the denoising of sequences of DW images (usually affected by Rician noise). They have a multi-directional structure generating 4D sequences along axes (x, y, z and k), where the last dimension displays the k directions of the microscopic motion for each slice. The OLPCA method takes advantage of the multi-directional redundancy of that multicomponent. Let x_p be a voxel of a three-dimensional DW image. In the OLPCA method, the 3D patches, surrounding x_p in any direction k, are reordered as a column vector of a matrix X

with $N \times K$ components, where N are the patch items and K the directions. The scheme of the LPCA method is listed in the Algorithm 1.

Algorithm 1 The LPCA scheme

1: **for** each X **do**
2: $C = X^T X$ //covariance matrix,
3: $C = WDW^T$ //singular value decomposition
4: $Y = XW$
5: $\hat{Y} = threshold(Y, \tau)$ //thresholding step
6: $\hat{X} = \hat{Y}W^T$ //denoised output
7: **end for**

More in detail:

- the matrix W contains the eigenvectors, while the diagonal matrix D contains the eigenvalues
- a SVD is computed for each matrix X of size $N \times K$,
- the threshold τ is equal to the local noise variance level $\gamma \cdot \sigma$

Finally, the Over Complete rule, for each voxel, combines all the estimates as follows:

$$\hat{x_i} = \sum_j \frac{1}{1 + ||\hat{Y}_j||_0}.$$

In this paper we are interested to analyse the step 3. In particular, our aim is to carry-out a SVD profiling in order to discuss the best strategy to give benefits on the performance of the overall Algorithm 1.

3 Parallel Approaches

As stated in the previous section in the OLPCA method, the SVD decomposition is performed on a covariance matrix associated to each voxel in a 3D patch, hence producing a number of small matrices that depends on the patch-size chosen. In our case, there are more than 650k small (14×14) matrices. In Algorithm 1, performance and speed-up can be increased by computing all SVD decompositions in a parallel fashion.

Two solutions could be taken into account:

i) a multi-threaded CPU version for each matrix X_i and performing a SVD of X_i, (as shown in Fig. 1, which would benefit of CPU caching for such a small problem size);
ii) a multi-streamed version by using CUDA streams, producing a pipeline of computations in a cascade manner (as shown in Fig. 2) where each new computation is done while moving previous results from device to host.

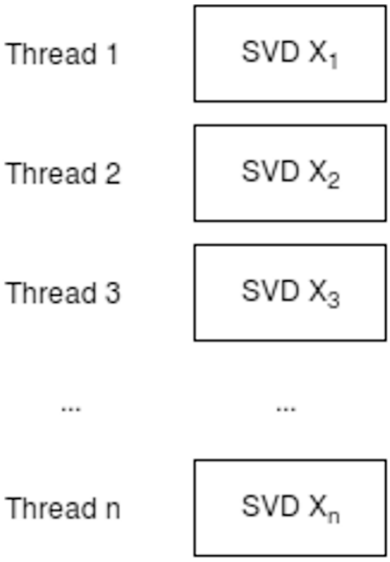


Fig. 1 A CPU multi-threaded approach on a set of small matrices. Each thread performs a SVD computation on a X_i matrix of small size. Depending on the CPU caching capabilities, performance can be enhanced for small problem sizes.

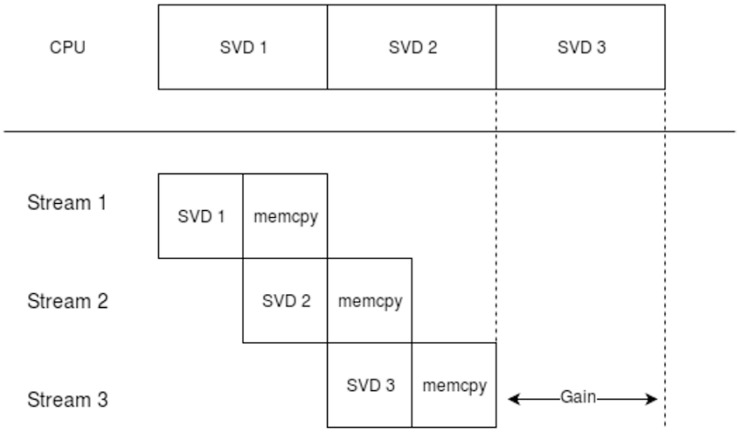


Fig. 2 CUDA streams approach to multiple SVD problems. Each stream performs a SVD decomposition on a small matrix. A pipeline is executed using multiple streams. In this example, 3 streams are used to compute 3 SVD decomposition. The upper side of the image shows a CPU executing 3 SVDs. The bottom side shows 3 streams working together on the same problem.

Listing 1 Calculate size for pre-allocated buffer

```
cusolverStatus_t cuSolverStatus;
cusolverDnHandle_t cuSolverHandle;
cusolverDnCreate(&cuSolverHandle);

cuSolverStatus = cusolverDnSgesvd_bufferSize(cuSolverHandle, M, N,
&Lwork);

float *Work;
cudaMalloc(&Work, WorkSize*sizeof(*Work));
float *rwork;
cudaMalloc(&rwork, M*M*sizeof(*rwork));
```

However, it is necessary to be careful in choosing a parallel approach to tackle this problem. Although it is possible to achieve good performance and speed-ups parallelizing the SVD computation, for instance by using a custom kernel or a library of choice (cuBLAS, CULA or other), we stress that in this case the size of all matrices X_i is very small (in the order of 10 to 14 squared matrices). Performing a SVD decomposition, which in the truncated case has a computational complexity of $O(min(mn^2, m^2n))$, of such small matrices on a CPU requires only a hundred of FLOPS thus allowing to compute a solution in a range of milliseconds.

The process of initiating a GPU and then perform a SVD decomposition requires more time than just performing it on a CPU. For instance, the theoretical number of threads that could be used for a 5×5 matrix is 25. In practice, the number of simultaneous threads per matrix is 1 or 5, meaning many thousands of simultaneous matrices are needed to obtain a sufficient number of threads. In order to justify a GPU computation, more operations should be done for which thousands of simultaneous threads are preferable to achieve a noticeable difference in overall performance.

Theoretically, one CUDA stream per problem could be used as a solution for one problem at a time, as depicted in Fig. 2. However, two issues arise:

- the number of threads per block would be very low. Typically, no less than 32 threads per block (the warp size, effectively the SIMD width, of all current CUDA-capable GPUs is 32 threads) are preferred to fully exploit the potential of a GPU card, and this is not practical, as previously mentioned;
- the prepare-lauch code would be more expensive than just performing the operation on a CPU and the resulting overhead would be unacceptable.

NVIDIA cuBLAS has introduced a batch-mode feature but this is only available for multiply, inverse routines and QR decomposition. It is intended for solving a specific class of problems by allowing a user to request a solution for a group of identical problems (a SIMD operation in this case). cuBLAS batch interface is intended for specific circumstances with small problems (matrix size < 32) and very

Listing 2 An example of a cuBLAS SVD routine call. Profiling events registration is included.

```
cudaEventCreate(&start);
cudaEventCreate(&stop);
cudaEventRecord(start, 0);

cuSolverStatus = cusolverDnSgesvd(cuSolverHandle, 'A', 'A', M, N,
devA, M, devS, devU, M, devV, N, Work, Lwork, NULL, devInfo);

cudaEventRecord(stop, 0);
cudaEventSynchronize(stop);
cudaEventElapsedTime(&executionTime, start, stop);
cudaEventDestroy(start);
cudaEventDestroy(stop);
```

Listing 3 An example of a SVD routine call using CULA. Profiling event registration is included.

```
cudaEventCreate(&start);
cudaEventCreate(&stop);
cudaEventRecord(start, 0);

status = culaSgesvd('A', 'A', M, N, A, lda, S, U, ldu, V, ldvt);

cudaEventRecord(stop, 0);
cudaEventSynchronize(stop);
cudaEventElapsedTime(&executionTime, start, stop);
cudaEventDestroy(start);
cudaEventDestroy(stop);
```

large batch sizes (number of matrices > 1000). Our case study falls in this category but a SVD decomposition should be derived and implemented using the batch QR decomposition. This goes beyond the scope of this case study. Also, using QR iterations to achieve a SVD decomposition would result in a slower algorithm than LAPACK optimized SVD.

The GPU implementation is based on two libraries: NVIDIA cuBLAS and CULA. Here we report a sample code for each library. In block 1 and 2, we exploit the cuSOLVER APIs. They represent a high-level package based on cuBLAS and cuSPARSE, combining three separate libraries. Here we make use of the first part of cuSOLVER, called cuSolverDN. It deals with dense matrix factorization and solve routines such as LU, QR and SVD. We use the S data type for real valued single precision inputs.

In Listing 1, we use a helper function to calculate the sizes needed for pre-allocated buffer *Lwork*. Moreover, in Listing 2, the *cusolverDnSgesvd* function computes the SVD of a $m \times n$ matrix *devA* (the prefix *dev* accounts for a matrix placed on the

device) and corresponding left and/or right singular vectors. The working space needed is pointed by the *Work* parameter, declared as a vector of float, while the size of the working space is specified in *Lwork*, as calculated in Listing 1.

Finally, in Listing 3, we report another test code based on the CULA library. Here, the SVD functionality is implemented in the *gesvd* routine. Initialization and actual routine call is simplified as opposed to cuBLAS and cuSOLVER libraries, as it is not necessary to perform any kind of preallocation. A matrix A is passed to the *culaSgesvd* routine and the result is stored in the S, U and V matrices. Here we use the S data type for real valued single precision inputs. CULA also provides a version of its routines to specifically work on the device. In this case, *culaDeviceSgesvd* would be used instead of *culaSgesvd* and all input and output matrices need to be properly moved from host to device and vice-versa.

4 Performance Analysis and Profiling

To support our analysis, we report a test for a group of SVD decompositions. In Fig. 3 we have compared the execution times of a SVD decomposition performed on a CPU (Intel Core i7 860 @ 2.80Ghz) and a GPU (NVIDIA Quadro K5000 4Gb). In the GPU version, we perform a SVD decomposition on a random test matrix and compare the libraries: CULA and cuBLAS. GPU performance gains are noticeable when the size of a squared input matrix increases. In fact, the CPU version performs better when the problem size is smaller than 256×256 (Fig. 4). These results show how the GPU overhead has an important impact on performance for very small matrices. Moreover, it naturally follows that calling a SVD routine, of a chosen library, several thousands of times affects the overall performance. In our tests, issuing several SVD decompositions using CULA or cuBLAS resulted in a stalling computation.

The CPU version we use in this comparative study is based on the SVD implementation provided by the OpenCV library. We use the *compute::SVD* routine which stores the results in user-provided matrices, similarly to the cuBLAS and CULA routines. This choice is mainly motivated by the intention of testing state of the art and/or production-ready routines used in many application of image processing and to investigate the impact each module, CPU or GPU -based, gives to algorithms that specifically need to manipulate a considerable amount of small matrices.

5 Conclusion

In many image processing methods, heavy computational kernels are implemented by using suitable methods and scientific computing strategies. We propose a solution based on optimized standard scientific libraries in order to improve the performance

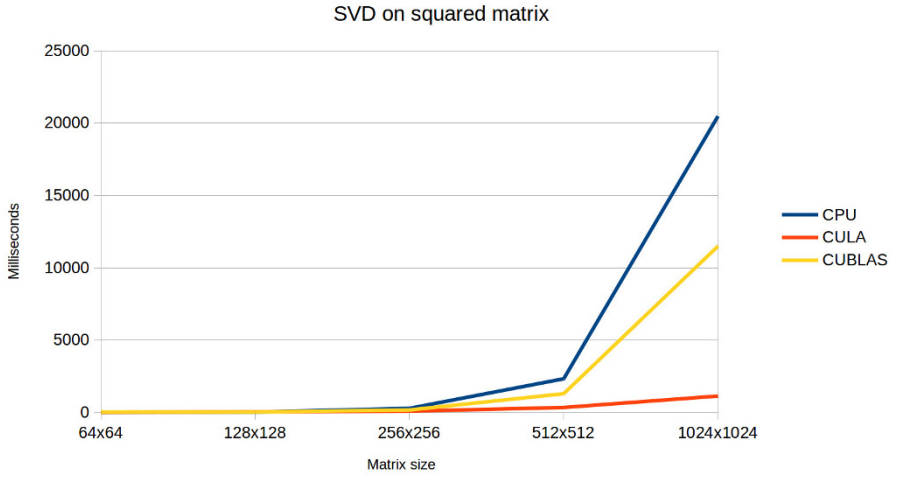


Fig. 3 CPU and GPU Execution time comparison for squared matrices.

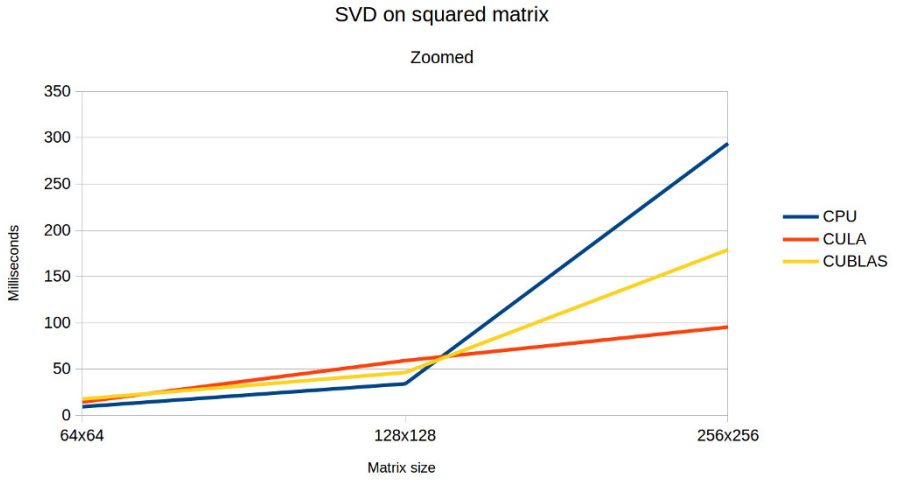


Fig. 4 CPU and GPU execution time comparison for squared matrices. A section of Fig. 3 in order to highlight the performance scatter for small problem sizes. As shown, CPU performance is still preferable for small problem sizes.

of same tasks in the denoising OLPCA algorithm. In a future work we will develop a more optimized parallel software for this algorithm.

Acknowledgment

The work is partially founded thanks to Big4H (Big data analyics for e-Health applications) Project, Bando Sportello dell'Innovazione - Progetti di trasferimento tecnologico cooperativi e di prima industrializzazione per le imprese innovative ad alto potenziale - della Regione Campania.

References

1. Nvidia cuda programming guide. [online] (2012). Tech. Rep. available at http://www.nvidia.com/content/cudazone/download/OpenCL/NVIDIAOpenCLProgrammingGuide.pdf
2. Amato, F., Barbareschi, M., Casola, V., Mazzeo, A.: An fpga-based smart classifier for decision support systems. In: Intelligent Distributed Computing VII, pp. 289–299. Springer (2014)
3. Amato, F., De Pietro, G., Esposito, M., Mazzocca, N.: An integrated framework for securing semi-structured health records. Knowledge-Based Systems **79**, 99–117 (2015)
4. Amato, F., Moscato, F.: A model driven approach to data privacy verification in e-health systems. Transactions on Data Privacy **8**(3), 273–296 (2015)
5. Andrews, H.C., Patterson, C.L.: Singular value decompositions and digital image processing. IEEE Trans. on Acoustics, Speech, and Signal Processing **ASSP-24**, 26–53 (1976). URL http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=1162766
6. Buades, A., Coll, B., Morel, J.: A review of image denoising algorithms, with a new one. Multiscale Modeling & Simulation **4**(2), 490–530 (2005). DOI 10.1137/040616024. URL http://dx.doi.org/10.1137/040616024
7. Bydder, M., Du, J.: Noise reduction in multiple-echo data sets using singular value decomposition. Magnetic Resonance Imaging **24**(7), 849 – 856 (2006). DOI http://dx.doi.org/10.1016/j.mri.2006.03.006. URL http://www.sciencedirect.com/science/article/pii/S0730725X06001317
8. Cuomo, S., De Michele, P., Galletti, A., Marcellino, L.: A gpu parallel implementation of the local principal component analysis overcomplete method for dw image denoising. In: 2016 IEEE Symposium on Computers and Communication (ISCC), pp. 26–31 (2016). DOI 10.1109/ISCC.2016.7543709. URL http://dx.doi.org/10.1109/ISCC.2016.7543709. The Twenty-First IEEE Symposium on Computers and Communication, 27-30 June 2016, Messina, Italy
9. Cuomo, S., De Michele, P., Piccialli, F.: A regularized mri image reconstruction based on hessian penalty term on cpu/gpu systems. pp. 2643–2646 (2013). DOI 10.1016/j.procs.2013.06.001. URL http://www.scopus.com/inward/record.url?eid=2-s2.0-84892506892&partnerID=40&md5=cc785a43da0426b134b5a4e05bc3ad5e
10. Cuomo, S., De Michele, P., Piccialli, F.: 3d data denoising via nonlocal means filter by using parallel GPU strategies. Comp. Math. Methods in Medicine **2014**, 523,862:1–523,862:14 (2014). DOI 10.1155/2014/523862. URL http://dx.doi.org/10.1155/2014/523862

11. Cuomo, S., Galletti, A., Giunta, G., Marcellino, L.: Toward a multi-level parallel framework on gpu cluster with petsc-cuda for pde-based optical flow computation. pp. 170–179 (2015). DOI 10.1016/j.procs.2015.05.220. URL http://www.scopus.com/inward/record.url?eid=2-s2.0-84939155665&partnerID=40&md5=ddcb2162cbc29925e582fc9498463059
12. Cuomo, S., Galletti, A., Giunta, G., Starace, A.: Surface reconstruction from scattered point via rbf interpolation on gpu. pp. 433–440 (2013). URL http://www.scopus.com/inward/record.url?eid=2-s2.0-84892530536&partnerID=40&md5=517ef890d27db116781907864f6861fd
13. Konstantinides, K., Natarajan, B., Yovanof, G.S.: Noise estimation and filtering using block-based singular value decomposition. IEEE Transactions on Image Processing **6**(3), 479–483 (1997). DOI 10.1109/83.557359
14. Manjón, J., Coupé, P., Concha, L., Buades, A., Collins, D., Robles, M.: Diffusion weighted image denoising using overcomplete local pca. PLoS ONE **8**(9) (2013). DOI 10.1371/journal.pone.0073021. URL http://www.scopus.com/inward/record.url?eid=2-s2.0-84883366803&partnerID=40&md5=467a3af41b50d17486ab1385ccf8e816
15. Manjón, J.V., Coupé, P., Martí-Bonmatí, L., Collins, D.L., Robles, M.: Adaptive non-local means denoising of MR images with spatially varying noise levels. Journal of Magnetic Resonance Imaging **31**(1), 192–203 (2010). DOI 10.1002/jmri.22003. URL http://www.hal.inserm.fr/inserm-00454564
16. Muresan, D.D., Parks, T.W.: Orthogonal, exactly periodic subspace decomposition. IEEE Transactions on Signal Processing **51**(9), 2270–2279 (2003). DOI 10.1109/TSP.2003.815381. URL http://dx.doi.org/10.1109/TSP.2003.815381
17. Palma, G., Piccialli, F., Michele, P.D., Cuomo, S., Comerci, M., Borrelli, P., Alfano, B.: 3d non-local means denoising via multi-gpu. In: Proceedings of the 2013 Federated Conference on Computer Science and Information Systems, Kraków, Poland, September 8-11, 2013., pp. 495–498 (2013). URL http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6644045
18. Poon, P., Wei-Ren, N., Sridharan, V.: Image denoising with singular value decompositon and principal component analysis. http://www.u.arizona.edu/~ppoon/ImageDenoisingWithSVD.pdf (2009)
19. Yang, J.F., Lu, C.L.: Combined techniques of singular value decomposition and vector quantization for image coding. IEEE Transactions on Image Processing **4**(8), 1141–1146 (1995). DOI 10.1109/83.403419

A machine learning approach for predictive maintenance for mobile phones service providers

A. Corazza, F. Isgrò, L. Longobardo, R. Prevete

Abstract The problem of predictive maintenance is a very crucial one for every technological company. This is particularly true for mobile phones service providers, as mobile phone networks require continuous monitoring. The ability of previewing malfunctions is crucial to reduce maintenance costs and loss of customers. In this paper we describe a preliminary study in predicting failures in a mobile phones networks based on the analysis of real data. A ridge regression classifier has been adopted as machine learning engine, and interesting and promising conclusion were drawn from the experimental data.

1 Introduction

A large portion of the total operating costs of any industry or service provider is devoted to keep their machinery and instruments up to a good level, aiming to ensure a minimal disruption in the production line. It has been estimated that the costs of maintenance is the range 15-60% of the costs of good produced [14]. Moreover about one third of the maintenance costs is spent in not necessary maintenance; just as an example, for the U.S. industry only this is a $60 billion each year spent in unnecessary work. On the other hand an ineffective maintenance can cause further loss in the production line, when a failure presents itself.

Anna Corazza
DIETI, Università di Napoli Federico II e-mail: anna.corazza@unina.it

Francesco Isgrò
DIETI, Università di Napoli Federico II e-mail: francesco.isgro@unina.it

Luca Longobardo
DIETI, Università di Napoli Federico II e-mail: luc.longobardo@studenti.unina.it

Roberto Prevete
DIETI, Università di Napoli Federico II e-mail: roberto.prevete@unina.it

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_69

Predictive maintenance [14, 10] attempts to minimise the costs due to failure via a regular monitoring of the conditions of the machinery and instruments. The observation will return a set of features from which it is possible in some way to infer if the apparatus are likely to fail in the near future. The nature of the feature depend, of course, on the apparatus that is being inspected. The amount of time in the future that the failure will arise also depends on the problem, although we can state, as a general rule, that the sooner a failure can be predicted, the better is in terms of effective maintenance.

In general the prediction is based on some empirical rule [23, 17, 19], but over the last decade there has been some work devoted to apply machine learning [6, 22, 5] techniques to the task predicting the possible failure of the apparatus. For instance, a Bayesian network has been adopted in [9] for a prototype system designed for the predictive maintenance of non-critical apparatus (e.g., elevators). In [12] different kind of analysis for dimensionality reduction and support vector machines [7] have been applied to rail networks. Time series analysis has been adopted in [13] for link quality prediction in wireless networks. In a recent work the use of multiple classifiers for providing different performance estimates has been proposed in [18].

An area where disruption of service can have a huge impact on the company sales and/or the customer satisfaction is the one of mobile phone service providers [8, 4]. The context considered in this work is a the predictive maintenance of national mobile phone network, that is being able to foresee well in advance if a cell of the network is going to fail. This is very important as the failure of a cell can have a huge impact on the users' quality of experience [11], and to prevent them makes less likely that the user decides to change service provider.

In this paper we present a preliminary analysis on the use of a machine learning paradigm for the prediction of a failure on a cell of a mobile phones network. The aim is to predict the failure such in advance that no disruption in the service will occur, lets say, at least a few hours in advance. A failure is reported among a set of features that are measured every quarter of an hour. The task is then to predict the status of the feature reporting the failure within a certain amount of time. As for many other predictive maintenance problems given we are dealing with a very large amount of sensors [15].

The paper is organised as follows. Next section describes the data we used and reports some interesting properties of the data that have been helpful in designing the machine learning engine. The failure prediction model proposed is discussed in Section 3, together with some experimental results. Section 4 is left to some final remarks.

2 Data analysis

To predict failure, considered data is obtained by monitoring the state of base transceiver stations (also known as cells) in a telecommunication network, during a 1 month (31 days) time-span. A cell represents the unit for a telecommunication net-

work in the tackled case study. Cells are grouped into antennas, so that one antenna can contain several cells. The goal for the problem is to predict a malfunctioning (pointed out by an alarm signal originated from cells) in a cell.

Furthermore, information about the geographical location of the cell can be relevant. When the Italian peninsula is considered, the total number of cells amounts to nearly $52,000$. For instance, when considering the total number of measurements, we get more than 150 millions of tuples.

Several kinds of statistical analysis were implemented to explore the data, and some interesting key-points and critical issues emerged from this analysis.

First of all, more than the 60% of the cells did not show any alarm signal. This is a quite usual behavior, as the system works smoothly for most of the time. Even when such cases are excluded, the average number of alarms per cell is only 3 in a month. In order to obtain a data set which is meaningful enough for a statistical analysis, only cells with at least 6 alarms have been kept: in this way the number of cells is further reduced to less than $2,000$. Moreover, among the remaining cells the proportion between alarm tuples and non-alarm tuples still remains high, as the former represent barely 1% of the total. However, we considered it acceptable, as malfunctioning must be considered unlikely to happen. In the end, this unbalance strongly influence the pool size of useful input data, and must faced with an adequate strategy.

Another critical issue regards the presence of several non-numeric values spread among the tuples. There are four different undefined values, among whose, `INF` values are the most frequent. Indeed, `INF` is the second most frequent value among all fields. All in all, discarding tuples containing undefined values in some of the fields would cut out 80% of data, leaving us with a too small dataset. We therefore had to find a different strategy to face the problem.

As already stated, another problematic issue regards the temporal dimension of data. Time-span is only one month, which on a time series related problem is not very much, to begin with the very basic problem of properly splitting data into training and test sets.

Another key-point regarding the data was found by looking at scatter plot diagrams between pairs of features. These diagrams highlighted two aspects: the first one is that alarm occurrences seem related to the values of some specific features. The second one is that there are two identical features. Since we don't have information about the meaning of the various features, we can't tell if this is supposed to be an error on the data supplied.

In addition to these, other statistical analysis were performed, focusing mainly on the values assumed by the features. Average values are summarized, along with standard deviations in Figure 1.

Inspecting Figure 1 we can see that FEATURE_7 and FEATURE_9 show a significant difference in term of average value between alarm and non-alarm events and thus, can trace a good starting point for a machine learning approach. Moreover, we can split features in three different groups.

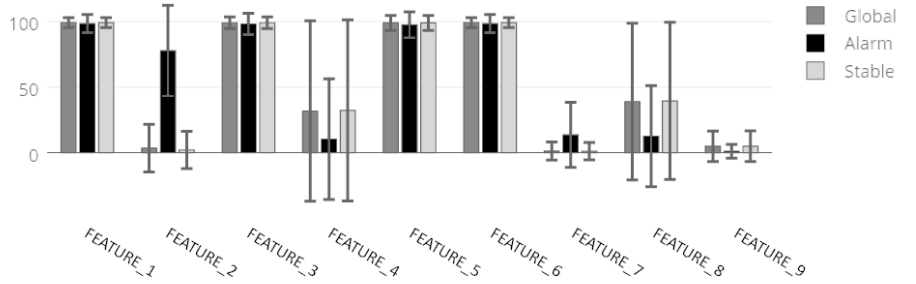


Fig. 1 Average values for the features in a stable or alarm situation. The line on every bar represent the standard deviation

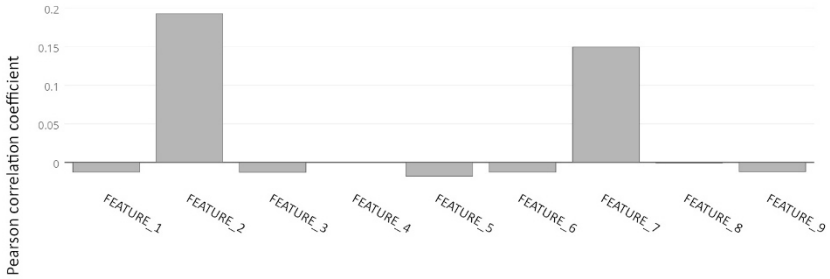


Fig. 2 Pearson correlation coefficient between features and alarm indicator

1. The first group is composed of: FEATURE_4, FEATURE_8, and FEATURE_9. These features have constant values in all the three conditions but also a relatively high standard deviation.
2. The second group is composed of: FEATURE_6, FEATURE_3, FEATURE_5, and FEATURE_1. Also in this case, the features tend to have constant values in all three situations, but with a relatively lower standard deviation.
3. The third group is composed of: FEATURE_7 and FEATURE_2. These features show a large difference in terms of both average value and standard deviation between alarm and non-alarm situations.

To better analyse the differences between alarm and non-alarm situations, Pearson correlation coefficients have been calculated between each feature and the alarm indicator. Results are shown in Figure 2, and confirm that FEATURE_2 and FEATURE_7 appear to be more related with an alarm occurrence, although the correlation value is always lower than 0.2.

Last but not least, the alarm propagation effect has been analyzed, to check if an alarm occurring in a cell is correlated to alarms in nearby cells. The results in Figure 3 show that this is the case only for cells belonging to the same antenna. In general, when the distance increases and cells of different antennas are considered, the probability of cooccurrent alarms drops close to 0. We can therefore conclude that, according to our data, there is no propagation effect.

3 Failure prediction

Alarm prediction is approached as a binary classification in a vector space model [6]. In simple words, the input to the classifier is given by a vector of features measured at times $t, t+1, \ldots, t+d$ in a given cell, while the output is positive or negative, depending whether an alarm in the same cell is foreseen or not for the time $t+\Delta$, with $\Delta > 0$. Such an approach is very general, and can be ported to similar problems in different domains.

More specifically, we chose a classification model based on Tikhonov regularization [21] (also known as Ridge regression). This is a direct classification model, based on the assignment of binary label to a tuple of features in accordance with the presence or the absence of an alarm signal after a chosen amount of time.

The representation of the features we use for classification is important for the system performance. In fact, we have a time series and how the temporal information is represented is a crucial point. Our approach consists of the following two steps:

1. Feature expansion;
2. Feature selection.

Feature expansion is obtained by calculating new features derived from the existing ones. An example of this is the application of some aggregation measures like means, standard deviations and/or variances. Another used technique for increasing the dimensionality is related to the use of convolution filters on the variables. More specifically, we used a set of filters belonging to the family of wavelets. One single filter was used multiple times with different parameters. The output for this step are tuples with a much larger dimensionality. We switched, in some extreme cases, from a 9 dimensions problem to a 1200 dimensions problem. Such number of variables

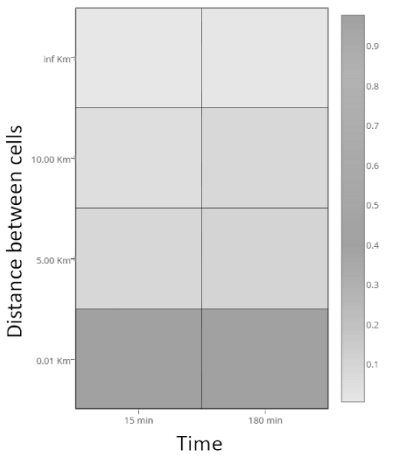


Fig. 3 Illustration of the probabilities of getting an alarm for the cells close to a cell signaling an alarm, after 15 and 180 minutes

is vastly overabundant. For this reason a further step for feature selection becomes necessary.

A process of automatic *feature selection* was chosen to increase portability and maintain a data-oriented approach. In particular we used an algorithm for L1L2 regularization implemented in the "l1l2py"[1] Python package. This algorithm combines the classic shrinkage methods given by Ridge Regression and Lasso Regression [20].

We consider a regression problem where the output y is reconstructed from features $x_i, i \in [1,p]$ by combining them with coefficients $\beta = \{\beta_i\}$. Ridge Regression uses a L2 norm in order to force a constraint on the regression coefficients size by reducing their absolute value:

$$\hat{\beta}_{ridge} = \arg\min_{\beta} \left\{ \left\| \mathbf{y} - \sum_{j=1}^{p} x_j \beta_j \right\|^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \tag{1}$$

On the other hand, Lasso Regression instead uses a L1 norm forcing sparsity in data and the annulment of some of the coefficients:

$$\hat{\beta}_{lasso} = \arg\min_{\beta} \left\{ \left\| \mathbf{y} - \sum_{j=1}^{p} x_j \beta_j \right\|^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{2}$$

The final step is the effective experimental assessment of the classifier. First of all, we have to decide how to solve the critical issues emerged from data analysis and pointed out in the preceding section: how to mandage undefined values and how to split data into training and test set while reducing unbalancing of positive and negative examples.

With relation to the issue of undefined values we decided to operate a fixed substitution of the most frequent of such values (`INF`) based on the average value assumed by the considered feature, according to the scheme in Table 1. The number

Table 1 Substitution of undefined values in the features which assume such value.

Feature	Substitution value
FEATURE_6	120
FEATURE_3	120
FEATURE_5	120
FEATURE_1	120
FEATURE_7	-10

of occurrences of the other undefined values is relatively negligible and the tuples containing those values were simply dropped.

[1] http://slipguru.disi.unige.it/Software/L1L2Py/

In order to fix the balance between positive and negative samples we kept all the available positive samples, which were the ones with a minor number of occurrences N_p, and randomly choose $N_n = 4N_p$ negative examples.

The splitting of data into training and test set has been solved by a temporal based partitioning: we selected the first $2/3$ of the month for the training, and the remainder of the data was used as test set. We could act like that because positive examples have a nearly uniform distribution in data. Therefore, even if we applied the split without considering the frequencies of positive and negative examples, we obtained an acceptable balance for both sets. Furthermore, we want to underline how it is fundamental to operate an accurate sampling of data composing the training set, because including tuples related in some way with the occurrence of an alarm results in a hike of performance.

In the experiments different time shifts Δ have been considered. An analysis of the results showed some few interesting points.

First of all, we tested both generic and location-based models. The former does not consider geographical information, while the latter is a location-based model. Classification results have shown how the geographical information is crucial to the classification, while a single generic model for the whole area fails to catch the different variety of underlying key factors specific to each geographic subarea. One example is illustrated in Figure 4, where we compare the results, in terms of ROC curve, from a sample of such two models. Training strongly geo-localized models resulted, in some of the best performance, with AUC values (the area under ROC curve) of $0.7 - 0.8$.

Another point regards the inverse proportionality between classification performance and the time shift: performance decreases while the time shift between observations and alarm increases. In fact, a regular loss in performance can be observed when the time shift raises from a quarter of an hour up to $6 - 7$ hours; after that performance fundamentally go close to a random guess.

Last, we run some tests to analyse how performance changes in relation to the introduction of automatic feature selection. Models built directly using all the features produced by the feature expansion phase and models where we an *l1l2py* step of feature selection was applied have been compared. The performance of the system with feature selection shows a constant (although relatively low) improvement with respect to the one without it. One example of this is showed in Figure 5.

We noticed that the set of features chosen by the feature selection step changes depending on the location of the considered cell. However, the final features are always correlated with the two which showed the largest coefficient of linear correlation with the output, that is: FEATURE_7 and FEATURE_2. Such analysis can also help the service provider to analyse which are the most likely causes of malfunctioning.

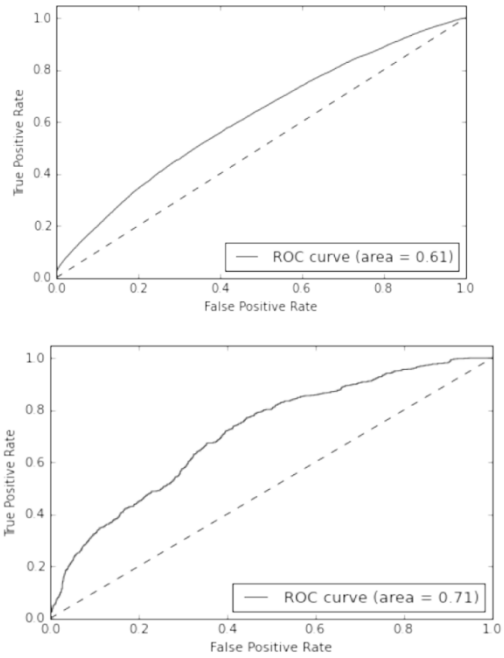


Fig. 4 Comparison between ROC curves generated on test set by a model trained and tested on the whole Italian area (top) and another specific to a single location (bottom). Time shift for prediction is 3 hours.

4 Conclusions and future work

In this paper we described a possible strategy to tackle a problem of alarm prediction in a domain where time series of features are available. From a first analysis of the dataset some issues have been raised, including the problem of undefined values. The alarm prediction has been defined as a classification problem which was solved by ridge regression. From the experimental results, we can conclude that geographical localization is important for the performance and that it is only possible to preview alarms occurring in a few hours.

Future directions of this analysis should explore models able to exploit the mutual position among cells, or, in general, to better exploit data. Among these, one of the most promising is the *cascade based model*. This approach is not focused on a single defined model to use, but on the chaining of various models for data classification which are applied sequentially. It aims to reduce, through several steps, the range of variables and dataset size. In a way which reminds of the decision tree approach, every stage is built to reduce the pool of data that is given as input to the next stage or the dimensionality of the problem.

The idea had origin from the implementation of AdaBoost [16] algorithm, which consists of a sequence of steps aiming to reduce the number of elements to classify.

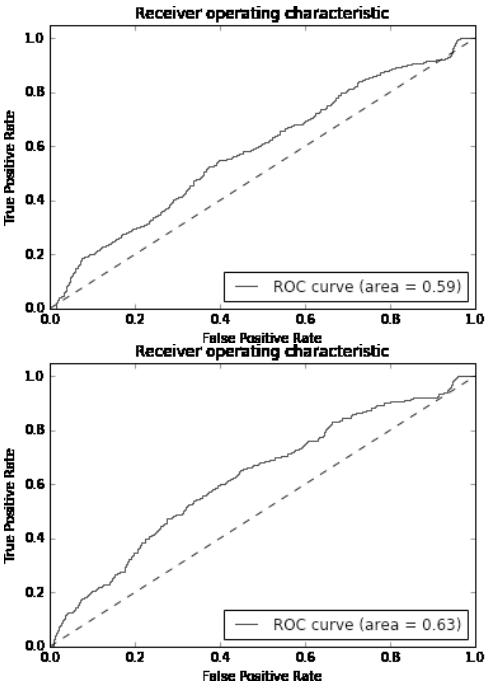


Fig. 5 Comparison between ROC curves obtained from models in a case without using feature selection (top) and in a case with using it (bottom)

The main key-point is to use a sequence of models increasing in complexity in order to cut out as many samples as possible in each stage. The early stages are typically the ones where the biggest cut is operated, while the latest stages are reserved to more refined models which have to take the hardest decision on the most difficult data.

In our case, this approach could not be applied because of the relativly small size of the sample dataset which remained after removing all non-relevant cells and balancing data. In fact, such multi-stage system is effective when it can exploit a really large dataset. On the other hand, in actual cases, a nearly unbounded quantity of data can be collected and therefore this approach could express its potentialities.

Acknowledgements

The research presented in this paper was partially supported by the national projects CHIS - Cultural Heritage Information System (PON), and BIG4H - Big Data Analytics for E-Health Applications (POR).

References

1. Amato, F., De Pietro, G., Esposito, M., Mazzocca, N.: An integrated framework for securing semi-structured health records. Knowledge-Based Systems **79**, 99–117 (2015)
2. Amato, F., Moscato, F.: A model driven approach to data privacy verification in e-health systems. Transactions on Data Privacy **8**(3), 273–296 (2015)
3. Amato, F., Moscato, F.: Exploiting cloud and workflow patterns for the analysis of composite cloud services. Future Generation Computer Systems (2016)
4. Asghar, M.Z., Fehlmann, R., Ristaniemi, T.: Correlation-Based Cell Degradation Detection for Operational Fault Detection in Cellular Wireless Base-Stations, pp. 83–93. Springer International Publishing, Cham (2013)
5. Barber, D.: Bayesian Reasoning and Machine Learning. Cambridge University Press (2012)
6. Bishop, C.M.: Pattern recognition and machine learning. Springer (2006)
7. Cortes, C., Vapnik, V.: Support-vector networks. Machine learning **20**(3), 273–297 (1995)
8. Damasio, C., Frölich, P., Nejdl, W., Pereira, L., Schroeder, M.: Using extended logic programming for alarm-correlation in cellular phone networks. Applied Intelligence **17**(2), 187–202 (2002)
9. Gilabert, E., Arnaiz, A.: Intelligent automation systems for predictive maintenance: A case study. Robotics and Computer-Integrated Manufacturing **22**(5), 543–549 (2006)
10. Grall, A., Dieulle, L., Berenguer, C., Roussignol, M.: Continuous-time predictive-maintenance scheduling for a deteriorating system. IEEE Transactions on Reliability **51**(2), 141–150 (2002)
11. Jain, R.: Quality of experience. IEEE MultiMedia **11**(1), 96–95 (2004)
12. Li, H., Parikh, D., He, Q., Qian, B., Li, Z., Fang, D., Hampapur, A.: Improving rail network velocity: A machine learning approach to predictive maintenance. Transportation Research Part C: Emerging Technologies **45**, 17 – 26 (2014). Advances in Computing and Communications and their Impact on Transportation Science and Technologies
13. Millan, P., Molina, C., Medina, E., Vega, D., Meseguer, R., Braem, B., Blondia, C.: Tracking and predicting link quality in wireless community networks. In: 2014 IEEE 10th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), pp. 239–244 (2014)
14. Mobley, R.K.: An introduction to predictive maintenance, 2nd edn. Butterworth-Heinemann (2002)
15. Patwardhan, A., Verma, A.K., Kumar, U.: A Survey on Predictive Maintenance Through Big Data, pp. 437–445. Springer International Publishing, Cham (2016)
16. Schapire, R.E.: Explaining AdaBoost, pp. 37–52. Springer Berlin Heidelberg, Berlin, Heidelberg (2013)
17. Scheffer, C., Girdhar, P.: Practical machinery vibration analysis and predictive maintenance. Elsevier (2004)
18. Susto, G.A., Schirru, A., Pampuri, S., McLoone, S., Beghi, A.: Machine learning for predictive maintenance: A multiple classifier approach. IEEE Transactions on Industrial Informatics **11**(3), 812–820 (2015)
19. Swanson, D.C.: A general prognostic tracking algorithm for predictive maintenance. In: Aerospace Conference, 2001, IEEE Proceedings., vol. 6, pp. 2971–2977. IEEE (2001)
20. Tibshiriani, R.: Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society. Series B **58**(1), 267–288 (1996)
21. Tychonoff, A., Arsenin, V.: Solution of ill-posed problems. Winston & Sons, Washington (1977)
22. Vapnik, V.: The Nature of Statistical Learning Theory. Springer, New York (1995)
23. Zhou, X., Xi, L., Lee, J.: Reliability-centered predictive maintenance scheduling for a continuously monitored system subject to degradation. Reliability Engineering & System Safety **92**(4), 530–534 (2007)

A Cloud-Based Approach for Analyzing Viral Propagation of Linguistic Deviations by Social Networking: Current Challenges and Pitfalls for Text Analysis Tools

Fiammetta Marulli[1], Alessandro Nardaggio[2], Adriano Racioppi[2], Luca Vallifuoco[2]
[1] DIETI, University of Naples Federico II, Via Claudio 21, Naples, Italy
fiammetta.marulli@unina.it
[2]In.tel.tec. S.p.a. – Centro Direzionale, Is. F2, Naples, Italy
{alessandro.nardaggio, adriano.racioppi, luca.vallifuoco}@inteltec.it}

Abstract. Social Networks activities offer rooms a non-trivial testbed for linguistic analysis, introducing significant revolutions into the creation of textual content. Linguistic solecisms, blunders, and generally speaking deviations from standard linguistic norms, are becoming the rule rather than the exception. Social Networks instantly and virally propagate deviations among users, who are increasingly moving away from standard language usage. Performing text analysis on deviated textual documents is a challenging and hard task. In this work, we propose an approach supporting text analysis tasks against a set of deviated textual documents. It exploits a "linguistic blundersonomy", a taxonomy of linguistic deviations, progressively built by processing textual Big Data provided by social network, in a Cloud-Based environment (SAP-HANA). A preliminary case study for Italian language is presented, showing how the exploitation of a linguistic blundersonomy could improve the precision of a sentiment and opinion mining process, and more generally, of a text analysis process.

1 Introduction

In in the past 15-20 years natural language engineering has made giant progress that has made several applications practical and widespread, such as machine translation, information retrieval and extraction, text summarization and generation, and sentiment and opinion mining. The underlying premises for successful NLP applications lie both in finer grained capabilities of linguistic analysis and in processing robustness, features which become more and more relevant vis-à-vis the need of operating in fully multilingual domains, as requested by a worldwide user community. Proven tools and methods from linguistics, statistics and machine learning have been leveraged for this purpose.

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_70

In this paper we show how they can be usefully combined with methods coming from social network analysis by presenting a data-driven approach aiming to exploit information from social networks so as to take into account the dynamics of language change in the age of Web 2.0.

The key point here is that Social Networks and more generally the Web 2.0 paradigm have progressively blurred the distinction between spoken and written language. Indeed, users tend to generate contents in these contexts in the same relaxed way as if they were carrying conversations at home or at their local bar, by writing very frequently long "streams of consciousness" that tend to disregard strict language rules. Hence, linguistic mistakes and blunders, or, more neutrally said, deviations from standard grammar and vocabulary, pour down very copiously. Furthermore, such variations are virally propagated among users by the very pervasive mechanisms of information sharing supported by Social Networks, thus greatly accelerating the dynamics of language change.

As a consequence, also the NLP well assessed approaches risk to fail, engaged as they are with the processing of snippets of text containing unclassified words and/or deviated language constructs.

Finally, the data-driven approach and the case study for Italian language here presented are parts of a wider research activity focusing on Social Networks users' behaviors, sentiments and opinions, performed within the CHIS (Cultural Heritage Information System) [11] project, promoted by the DATABENC [12] district.

2 Related Works

Most of NLP tasks, as referred in [3], have at their core a subtask of extracting the dependencies—who did what to whom—from natural language sentences. This task can be understood as the inverse of the problem solved in different ways by diverse human languages, namely, how to indicate the relationship between different parts of a sentence. Understanding how languages solve the problem can be extremely useful in both feature design and error analysis in the application of machine learning to NLP. Likewise, understanding cross-linguistic variation can be important for the design of Machine Translation systems and other multilingual applications.

When used in the analysis of Social Network activities, such as the automated interpretation of on-line product reviews, text analysis and automatic extraction of semantic orientation can be found under different umbrella terms, among which the most popular are sentiment analysis and opinion mining, which can be extremely helpful in marketing, measures of popularity and success, and compiling reviews.

In [4] authors propose a lexicon-based approach to extracting sentiment from text, exploiting dictionaries of words annotated with their semantic orientation (polarity and strength), and incorporating intensification and negation. Semantic orientation (SO) is a measure of subjectivity and opinion in text. It usually captures an evaluative factor (positive or negative) and potency or strength (degree to which the word, phrase, sentence, or document in question is positive or negative) towards a subject topic, person, or idea. There exist two main approaches to the problem of extracting sentiment automatically. A lexicon-based approach involves calculating orientation

for a document from the semantic orientation of words or phrases contained in the document [5]. On the other hand, the text classification approach, involves building classifiers from labeled instances of texts or sentences [6], essentially a supervised classification task. The latter approach could also be described as a statistical or machine-learning approach. We follow the first method, in which we use dictionaries of words annotated with the word's semantic orientation, or polarity.

Dictionaries for lexicon-based approaches can be created manually, as described in [4], or automatically, using seed words to expand the list of words [7].

Social resource sharing systems are also central elements of the Web 2.0 and use all the same kind of lightweight knowledge representation, called folksonomy [8][9]. As these systems are easy to use, they attract huge masses of users. Data Mining provides methods to analyze data and to learn models which can be used to support users.

Differently by sentiment and opinion mining, whose main purpose could be summarized into categorizing users' opinions and assigning to these ones an affectivity polarity, the folksonomies are taxonomy of resources, explicitly categorized by Web users', by exploiting semantic annotations (tags) to label them (e.g. multimedia objects). Folksonomies support information extraction and retrieval queries expansion, by adopting labels and terms which are closer and more familiar to users' language.

Furthermore, folksonomies attempt to catch users' idea about a resource and its belonging category, such providing also a folk dictionary of users' familiar (usually adopted) terms, very probably adopted in their researches.

In this perspective, folksonomies shorten the distance between official vocabularies and users' ones, such taking into account, even if in a blurred way, linguistic deviations introduced by users. As converse, much of the lexicon-based techniques focuses on using terms and corresponding semantic values extracted from official language dictionaries [10]. Following linguistic deviations and users' attitude to blur bounds between spoken and written language, particularly when they act in the social networks ground, is a very hard task for automatic text analyzer, also if supported by robust and intelligent NLP engines. Mistakes, blunders and any linguistic deviation, when is recurrent, could be a problem for a correct sense disambiguation or text analysis.

So, in this work we propose a data-driven approach that operate at a deep level of analysis, incorporating more frequent linguistic deviations, extracted by analyzing Italian social networks activities for a 6 months period, yet do not aim at a full linguistic analysis, although further work in that direction is possible.

3 The Motivating Example

Social Networks activities offer rooms a non-trivial testbed for linguistic analysis, introducing significant revolutions into the creation of textual content. Linguistic solecisms, blunders, and generally speaking deviations from standard linguistic norms, are becoming the rule rather than the exception.

Social Networks instantly and virally propagate deviations among users, who are increasingly moving away from standard language usage.

As the motivating example for our study, we can consider the Social Network analysis performed as a part of the CHIS [11] project research activities.

Among the aiming goals of this project, there was a quantitative estimation of the appreciation level of users towards the Cultural Heritage domain. Such type of social analysis was performed mainly by analyzing and mining sentiments and opinions expressed by users in their textual messages exchanged in Twitter.

A preliminary case study was performed by exploiting a Java based algorithm and an associative in-memory engine, as described in [13].

Results, in terms of precision, accuracy and processing performances were appreciable but not fully satisfying, thus motivating authors to improve their approach and to adopt more robust and performing technological solution, in order to deal with Big Textual Data.

From the authors' further investigations, a new type of analysis was performed. By exploiting the cloud based calculation service platform of the SAP HANA solution, a more robust and performing Text Analysis Process was designed.

Italian language was established as the reference language for this preliminary research step and case study. A massive analysis on textual Big Data was performed over a period of about 6 months; about 2 millions of messages published in the Italian Twitter platform were stored and textually mined.

The standard text analysis process natively powered by HANA platform was enriched with more specific domain vocabularies, in order to filter, at first, only the messages related to cultural events and topics. After this pre-filtering stage, sentiment and opinion mining processes were applied to the filtered messages, in order to discover and estimate, quantitatively, the distribution and the typology of sentiment of users toward cultural heritage domain. Even in this step, the sentiment analysis tools provided by SAP HANA Cloud were exploited.

Finally, the classification of sentiments and opinions, automatically obtained, was compared and analyzed by a human-made oracle, to estimate the achieved level of accuracy and precision.

The scores reached in the precision and recall evaluation metrics amounted about at 60% for the recall value and 65% for the precision value. It could be considered as appreciable values, observing that it was produced by a totally automatic process.

Authors further investigated over some possible causes and the feasible improvements for this process.

In order to improve the process, we introduced the human check in the control loop; we further clustered the whole set of tweets into smaller sets of fifty units, to manually control automatically obtained results. This reprocessing and benchmarking activities highlighted that the 75% of the messages wrongly classified or not classified (when the automatic process was unable to disambiguate), were caused by grammar mistakes or not exiting terms, freely introduced by users when writing and publishing their messages on the web.

Standing, at one side, from this "strongly variable" element and, at the other one, from the high level accuracy provided with grammatically correct messages by the available text analysis tools, we considered the opportunity of enriching the

performed text analysis loop by a further step, the linguistic blunders evaluation step, as shown in Figure 1.

This approach was also motivated because, after highlighting grammar mistakes (unofficial neologism, solecism, blunders (occurring typically as the wrong translation of a foreign term), etc.), some statistics were performed in order to quantify the frequency of these mistakes. After this evaluation, grammar blunders overpassing the 50 % threshold of occurrence, were classified, as regular mistakes, almost fully transparent to users' attention. This value for the systematic errors threshold was established by consulting linguistic experts and trusted research in the linguistic field [14]. The reference language considered in these studies was the Italian language.

So, linguistic blunders and mistakes spread over the web communities, were mined by performing a "reverse" text analysis and then collected in a taxonomy, called "blundersonomy" (an Italian Blundersonomy, in this case study). This particular type of vocabulary, in which each blunder was also associated with the grammatically correct form, was inserted into the text analysis loop, in order to improve the accuracy performances of sentiment and opinion mining automatic process.

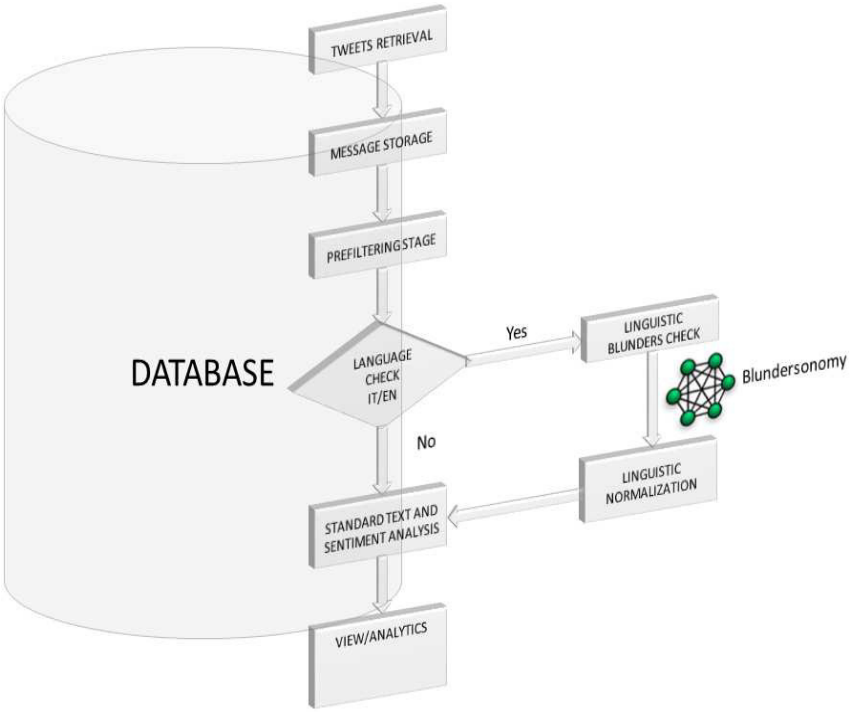


Fig. 1. The Text Analysis Process enriched by the introduction of the Blundersonomy Check in the standard linguistic loop.

4 The Italian Blundersonomy

The linguistic bludersonomy was introduced mainly for supporting standardized text analysis loops against a set of deviated textual documents.
Such lexical resource was organized as a taxonomy of linguistic deviations progressively extracted by textual messages and snippets, produced by social network users during their web activities.

Figure 2 shows the the XML Schema representing the composition for a typical blundersonomy.

It is organized as a tree, with seven first level branches, each representing a grammar element: articles, adverbs, adjectives, nouns, pronouns, prepositions and verbs.

Each branch holds a grammar hierarchy, wherever it makes sense (nouns, verbs, adjectives, adverbs) to make easy the blunder or mistake detection, indicating a more specific typology of element (e.g. an adverb or an adjective specifying a place, a time, a qualify, a quantity, etc..). Each child node populating a branch is made of two element: the deviated lemma (the blunder) and the corresponding correct form for that lemma. As an example, extracted from Italian language analysis, we can consider the following ones.

A very spread and common grammar mistake, mainly among, but not limited to, Italian young people is the adoption in the written language of the lemma “ke”, a not exiting word in the official Italian Vocabulary. This lemma is employed instead of the correct form “che”, standing for the pronoun “that”; as a consequence, also the lemma “xké” is abusively and brutally used instead of “perché” (used in Italian to mean both “why” and “because”). Another example of a very spread solecism among Italian people, involves the phoneme “hanno”, whose meaning is the verb “to have”, declined to the third plural person in the simple present tense (that is “they have”); this lemma has the same phonetic for the different one “anno”, (whose meaning is “year”). This is a very difficult case to detect in Italian language, because either the strings exist in the Italian vocabulary and only by the context disambiguation is possible to decide to which one the text is referring to.

More “funny” but desolating examples, extracted by our Italian Blundersonomy, showing of how “tragically” some linguistic dynamics are evolving across the web, are collected in Figure 3.

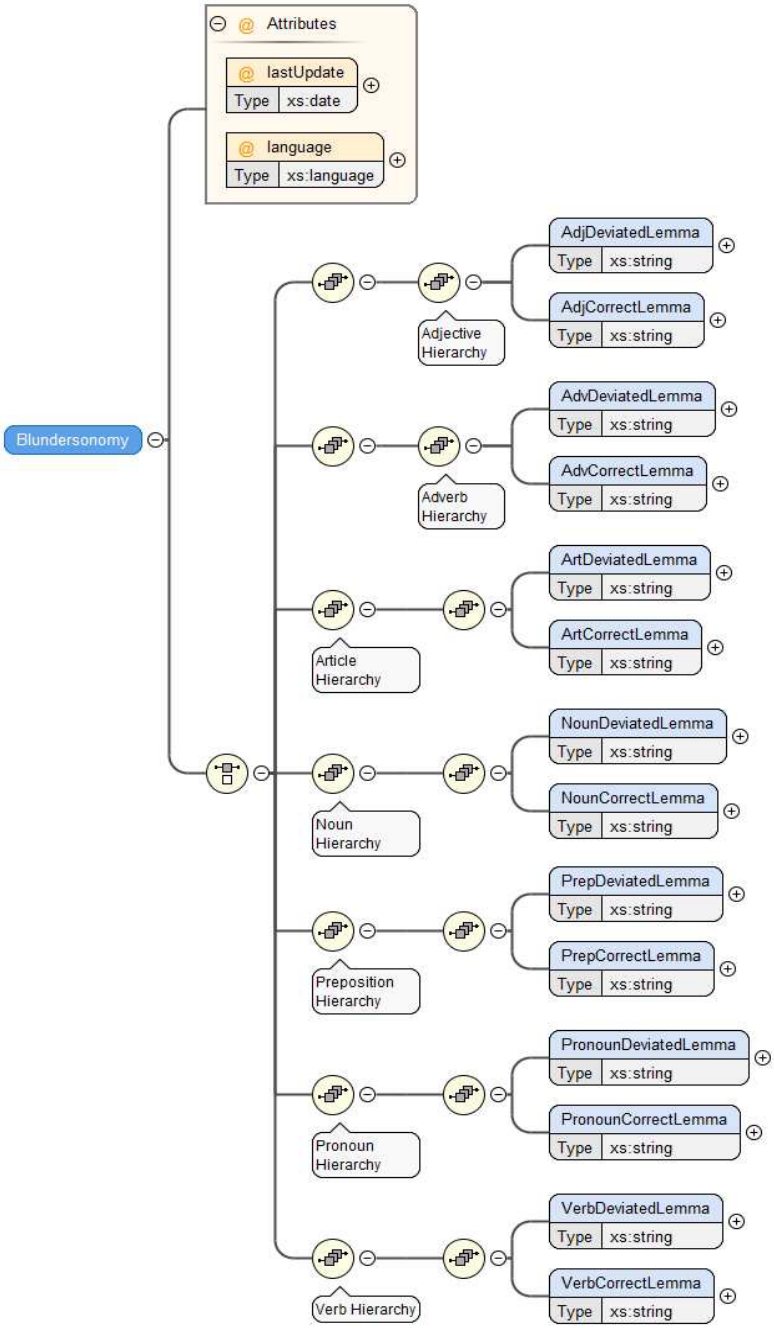


Fig. 2. XML-Schema for the Blundersonomy structure.

	version="1.0" encoding="UTF-8"	
Blundersonomy	@xmlns	.blundersonomy
	@xmlns:xsi	http://www.w3.org/2001/XMLSchema-instance
	@xsi:schemaL...	.blundersonomy file:/C:/BlundersonomySchema.xsd
	@lastUpdate	2016-06-22
	@language	en-US
	NounDeviated...	conoscienza
	NounCorrectL...	conoscenza
	NounDeviated...	ingegnieria
	NounCorrectL...	ingegneria
	NounDeviated...	QUORE
	NounCorrectL...	CUORE
	VerbDeviated...	AI
	VerbCorrectL...	HAI
	VerbDeviated...	anno
	VerbCorrectL...	hanno
	VerbDeviated...	travasare
	VerbCorrectL...	travisare
	VerbDeviated...	accellerare
	VerbCorrectL...	accelerare

Fig. 3. Some examples of Italian linguistic common mistakes from the Blundersonomy.

5 The Pilot Case of Study and Preliminary Results

The preliminary case study was performed on an Italian sample of Twitter social networks users. A set of about 2 millions (1.978.000 messages) of tweets exchanged in Italy over the early half of the 2016 were analyzed.

Extracted textual snippets from Twitter messages were processed via the SAP HANA Cloud Platform [1], a powerful system supporting basic and advanced text analysis as well as sentiment and opinion mining. In order to provide a robust structure to our approach, we first monitored users' activities for a time period of about six months, in order to collect and classify, in terms of occurrence criteria, the most frequent language deviations. This task was semi-automated and composed of two processing phases. The first one consists in filtering and extracting, from the input textual snippets, a list of candidate deviated terms and ambiguous expressions, exploiting official and trusted Italian linguistic resources, among which the semantic network MultiWordnet [2] and the Base Vocabulary provided by the Text Analysis Services provided by SAP HANA Cloud Platform. The second phase consists in mapping the language deviations to the most likely and supposedly correct form, and

was performed manually through human experts' intervention, in order to create a Knowledge Base, structured as a taxonomy of linguistic deviations, that we baptized as "Blundersonomy".

After building the Blundersonomy, it was integrated in the SAP HANA Text Analysis Knowledge Base (by extending the system with specific vocabularies and dictionaries). So, we performed a double check test, by giving as input original textual snippets from social networks activities, and in turn disabling and enabling the exploitation of our Blundersonomy. At a glance, we noticed that the tolerance and the capability of the Text Analysis system improved its performance of an overall value of about the 20% (about at the 20% higher than before, in the precision and recall values) after enabling the Blundersonomy exploitation (see Table 1).
Precision and Recall metrics were compared performing two parallel text analysis processing on the same data set, by employing and after by not employing the blundersonomy in Text Analysis loop.

Currently, a collection made of about 1 thousand of Italian deviated terms, was built. Each linguistic deviation was classified as systematic and not as sporadic according a frequency threshold, established at the value of 50 % of the occurrence over the whole set of lemma and words extracted and analyzed from the social network activity and other web resources (blogs and other blogging platforms).

Table 1. Precision and Recall values before and after introducing a Linguistic Blundersonomyin the Text Analysis Loop.

Evaluation Metric	Before introducing the Blundersonomy	After introducing the Blundersonomy
PRECISION	65%	85%
RECALL	60%	80%

6 Conclusions and Future Work

Social Networks activities offer rooms a non-trivial testbed for linguistic analysis. User Generated Contents and Web 2.0 introduced significant revolutions into the creation of textual content, thus effectively stimulating everyone to write and read almost about everything. In this scenario, linguistic solecisms and blunders, and generally speaking deviations from standard linguistic norms, are becoming the rule rather than the exception. Social Networks instantly and virally propagate deviations among users, who are increasingly moving away from standard language usage. Performing text analysis on deviated textual documents is a challenging and very hard task. In this work, the authors propose an approach for supporting text analysis tasks against a set of deviated textual documents. It was based on the creation and exploitation of a "linguistic blundersonomy", a particular taxonomy that collects and organizes linguistic deviations.

Sensible improvements, amounting to the 20% in the precision and recall values, were obtained, thus motivating the authors to go further in this direction.

Further refinements have to be applied to our approach, in order to support an effective NLP application. According to authors' opinion, the main contribution of this work consists in outlining and paving the way for introducing a data-driven approach, based on the exploitation of the Big Data provided from Social Networks and web activities, to improve textual and linguistic processing. Furthermore, this approach stimulate to monitor and take into account the change dynamics that natural languages are going through in consequence of the viral linguistic activities of users of social networks.

References

1. SAP HANA Cloud Platform, http://www.saphana.ondemand.com, (2016).
2. MultiWordNet, http://multiwordnet.fbk.eu/, (2015).
3. Bender, E.: Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax, Morgan & Claypool (Synthesis Lectures on Human Language Technologies, edited by Graeme Hirst, volume 20), (2013).
4. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-Based Methods for Sentiment Analysis, in Computational Linguistics, June 2011, Vol. 37, No. 2, Pages 267-307: 267-307, (2011).
5. Turney, P.: Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In Proceedings of 40th Meeting of the Association for Computational Linguistics, pages 417–424, Philadelphia, PA., (2002).
6. Pang, B., Lee, L., Vaithyanathan, S.: 2002. Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the Conference on Empirical Methods in NLP, pages 79–86, Philadelphia, PA, (2002).
7. Turney, P., Littman, M.: 2003. Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems, 21(4):315–346, (2003).
8. Giovanni Semeraro, Pasquale Lops, Marco De Gemmis, Cataldo Musto, and Fedelucio Narducci. 2012. A folksonomy-based recommender system for personalized access to digital artworks. J. Comput. Cult. Herit. 5, 3, Article 11 (October 2012), 22 pages.
9. Iván Cantador, Ioannis Konstas, and Joemon M. Jose. 2011. Categorising social tags to improve folksonomy-based recommendations. Web Semant. 9, 1 (March 2011), 1-15.
10. WordNet, a lexical database for English, https://wordnet.princeton.edu/
11. CHIS Project, http://www.databenc.it/progetti/chis-cultural-heritage-information-system/
12. DATABENC District, http://www.databenc.it
13. Benedusi, P., Marulli, F., Racioppi, A., Ungaro, L.: What's the matter with Cultural Heritage tweets? An Ontology–based approach for CH Sensitivity Estimation in Social Network Activities. In Proc. of IEEE SITIS 2015 International Conference, IEEE, 2015.
14. ACL, Association for Computational Linguistics, https://www.aclweb.org/

A Logic-Based Clustering Approach for Cooperative Traffic Control Systems

Walter Balzano, Maria Rosaria Del Sorbo, Aniello Murano, Silvia Stranieri
Dip. Ing. Elettrica e Tecnologie dell'Informazione, Università di Napoli Federico II, Italy
wbalzano@unina.it, marodel@gmail.com, aniello.murano@unina.it, s.stranieri@icloud.com

Abstract. Cooperative traffic control-systems have helped in the last years to extend the electronic horizon of vehicles in such a way that drivers can be informed quite in advance about any unexpected event occurring along a road. In this area, Vehicular Ad Hoc Networks (VANETs) perform information exchanges between vehicles by exploiting decentralized structures, where each vehicle is equipped with On Board Units (OBU) that allow to broadcast signals. In this paper, we revisit this communication framework by using a centralized approach, i.e., we consider a central unit for the data processing. This approach, on the one hand, facilitates the clustering process, thus obtaining the mitigation of the "broadcast storm" phenomenon (typically affecting VANETs); on the other hand, it allows the collection, management, and storage of the most relevant data concerning roads. Along the paper we evaluate the benefits of using our model, as well as we discuss on a smartphone integration.

1 Introduction

Road accidents, happening everyday all around the world, have spurred the interest in Vehicular Ad Hoc Networks (VANETs) [9]. Information exchange between vehicles is aimed to warning the drivers against exceptional events occurring on the road, so obtaining safety improvements among the drivers. In this context, it is very important to realize an attenuation of a bothersome phenomenon, often affecting the VANETs, known as "broadcast storm". A possible solution, as suggested by this work, is to minimize any redundancy in the transmissions among vehicles and introduce an innovative management of VANETs: a centralized structure where a central server stores the full knowledge and the features of the full network. Using our model it is possible to optimize the clustering process by choosing a cluster-head responsible for messages broadcasting, without using any directional antenna or wait for a green light communication. A centralized approach provides many other benefits: it allows collection, management, and storage of data observed by vehicles along their path.

The model we propose improves the one we presented in [5]. Along this paper we also consider an integration of the proposed model with smartphones. To this aim, we make use of the likely assumption that in each vehicle there is at least one of such a device due to its low cost and large spread. We exploit this integration in order to improve the performances of our logical model and so to extend its potentialities.

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_71

The rest of the paper is organized as follows. Section II gives the state of the art. Section III shows the centralized VANET model. Section IV describes a nonexclusive clustering algorithm, starting from the known DBSCAN one. Section V introduces the integration of smartphones. Section VI evaluates the time working complexity. Finally, Section VII gives some conclusions and future work directions.

2 Related Works

VANETs are a significant part of the Intelligent Transportation Systems (ITS) [11]. Nowadays they represent an important topic of research in scientific contexts, due to the high number of roads deaths recorded every year in the world [9]. A significant feature in the most common systems for inter-vehicular communication [10] is the interaction between the nodes of the network, which however is performed without any central unit for data management. In this work, we suggest instead a centralized model: a central server receives, manages, and stores data collected from all the detected events. In this way, we can observe many benefits. In particular, this allows to mitigate the transmission redundancy on using clustering processes. As explained in [8], the clustering can reduce some of the discomforts in communication between vehicles. For this reason they create clusters according to the vehicular mobility. Similarly, in [12], the clusters creation is based on the information about the average velocity and acceleration of each vehicle. Additionally, in [13] they use some metrics, based on the location or direction of travel, in order to organize vehicles into different clusters. We believe that the creation of the cluster is a task more appropriate for a central unit, as we exploit in this paper. This allows to store a full knowledge of the network topology and therefore allows to optimally choose a cluster-head, without using any directional antenna to address a message (as proposed to [14]), or without waiting a "*defertime*" (as instead shown in [4] and [5]).

We start from the model presented in [5]. This work introduces a logical multilayer model, whose aim is to represent the most common exceptional road events using a Petri Net, which creates a cause/effect relationship between events, so that one event can generate another in a predictive manner. In this work we improve and enrich that model by integrating it with electronic devices. In particular, we exploit the use of smartphones belonging to drivers or pedestrians.

Among other works on closed related topics we mention [1] in which a set of applications in Android platform has been proposed in order to avoid crashes, manage and monitor the traffic, obtain weather information, and optimize the public transportation. Similarly, in [15] a VANET formed by Android smartphones is presented and showed that an "information dissemination service can entirely rely on smartphone-based communications". Instead, in [3], it is highlighted a low performance using a VANET exclusively composed by phones and, for this reason, we propose an integration of these devices rather than a replace of the existing technologies.

3 Centralized Model

Typically, a VANET represents an ad hoc network, where each vehicle constitutes a node that is able to broadcast alarm messages to any other vehicle in the network by using on-board units, that is wireless devices placed within each vehicle.

In this work, we propose an innovative approach to this model network, by considering a centralized arrangement. The idea is to customize the existing model so that it can be equipped with a central server used both in the data management and storage tasks. Such a structure provides numerous advantages, all analyzed below in details, which can be schematically listed as follows:

- clustering process;
- full knowledge of the network topology;
- stronger Experience layer [6];
- data storage and statistics.

3.1 Clustering

In a VANET, and in information transmissions in general, the most frequent problem is known as "broadcast storm": it occurs when more vehicles are trying to transmit their information at the same time. This generates many discomforts in communication, due to interruptions or wireless channel saturation [7]. It is necessary to reduce the redundancy in information transmission in order to mitigate this phenomenon [18]. Thanks to a centralized model, this can be possible using a clustering process. When a vehicle detects a road event, it generates a cluster with all the nodes belonging to its communication range and sends an alarm message to each member of this cluster. According to the VANET operation, the receiver vehicles, in turn, broadcast this message to their adjacent nodes: it is possible, especially in a busy traffic condition, that two or more vehicles are really near to each other and, thus, the set of neighbors of the first vehicle contains some nodes belonging to the set of neighbors of the second vehicle. Let us suppose that a situation as the one shown in Fig. 1 occurs, where one car has an accident and this information is communicated to other vehicles. As we can see, the vehicles grouped in A and B have three and two nodes in their respective lists of neighbors, but one of them belongs to both the first and the second list. In particular, if we indicate with N_A and N_B respectively the lists of neighbors of A and B, we can say that:

$$N_A \cap N_B \neq \phi \tag{1}$$

This determines that the vehicle resulting from this intersection receives the same alarm message two times (from the ones directed to both the group A and B) and therefore he has a redundancy in transmission.

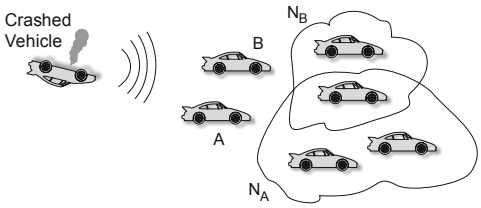


Fig. 1. Clustering to mitigate redundancy in transmission.

In order to avoid these repetitions in broadcasting, the central server could choose a cluster-head, interested in transmitting the received information externally. In order to explain how such a cluster-head works, we have to distinguish between two different kinds of road: one-way roads and two-way roads.

In a one-way road, the alarm message propagation only occurs in the opposite of the driving direction since (*i*) the vehicles, following the crashed vehicle, can detect the event and use it to avoid accidents; (*ii*) the other vehicles preceding the crashed vehicle have already passed the event location and so they are not interested about it. For this reason, the central server has to choose only a single cluster-head between the nodes belonging to the same cluster of the crashed vehicle. In order to cover a bigger zone, the cluster-head must be the farthest node with respect to the crashed vehicle.

Conversely, in a two-way road, the information broadcast must occur in both the driving direction and its opposite. Indeed, in such a kind of roads both vehicles that precede and follow the crashed one must be informed about it. For this purpose, the central server has to choose two different cluster-heads (rather than a single one): (*i*) the first one is the farthest node with respect to the crashed vehicle in the direction of traveling (tail cluster-head); (*ii*) the second one is the farthest node in the opposite direction (head cluster-head). In this way, a vehicle that receives an alarm message can in turn broadcast it only if it is elected as a cluster-head from the central server.

In order to know if a certain road is one way or two way, we suppose to associate a feature vector to it to keep track of the most important aspects of the road itself.

3.2 Clustering

The presence of a central server in a system for inter-vehicular communication allows to have an entity with a full knowledge of the network topology. This fact cannot happen in a decentralized structure and brings some benefits. For example, in this kind of centralized structure it is possible to choose the optimal routing protocol, according to the topology features.

In [2] the authors distinguish among three kinds of routing approaches (broadcast, multicast, and unicast), which are based only on the number of senders and receivers. Conversely, in [19] the authors identify two groups of routing protocols (proactive and reactive) according to the network topology. In addition, in [20] the authors add a third group, represented by *hybrid routing*, and they say that, in the topology-based routing, some routing choices are made according to the information of the communication links. In this kind of routing approach, the Link State and Distance Vector algorithms are the most used, due to their suitability to vehicular contexts.

3.3 *Experience layer & Storage and Statistic*

An important layer of the multi-layer structure described in [6] is the Experience layer, which stores all the data of the events observed by a specific vehicle: it can be useful when some events occur in a certain area, without understanding their causes, according to a predictive model. Thanks to this layer, we have some additional information and we can consider the cases not covered by the Petri Net model we have previously mentioned. In a centralized structure, this layer has a stronger role, since the surveys do not affect a single vehicle, but all the vehicles going across a certain road. This layer results to be very useful, on the one hand, for the purpose of a correct forecast of exceptional road events and, on the other hand, for the purpose of creating some statistics data concerning a certain road. For example, if on a street it is frequentlyp1a detected the event "punctures", it probably means that such a street is in a bad condition and this suggests to contact some public entities for the road maintenance. Thus, we can say that the Experience layer supports the data storage by collecting all the observations made by each vehicle concerning a certain road. In this way, it is possible to have a more complete knowledge of traffic conditions and also to increase the safety of the roads.

4 Clustering Algorithm

The clustering algorithm we present in this section is a non-exclusive clustering as each vehicle could belong to more than one cluster. We were inspired by a known density-based algorithm, the DBSCAN [17], in order to create a customized algorithm for the cluster creation in a centralized model. For simplicity, let us assume the case of a one-way road. All the reasoning below can be easily expanded to deal with two-way roads. The control flow cab be summarized in the following steps:

- *starting_node = crashed vehicle;*
- Repeat:
  - *new_cluster* = $N\varepsilon$(*starting_node*);
  - *cluster-head* = $n \in$ *new_cluster*:
    $d(starting_node,n) = \max\{d(starting_node,i)\}$,
    $\forall i \in$ *new_cluster*;
  - *starting_node* = *cluster-head*;
- Until time and space limitations are satisfied.

Initially, the *starting_node* is represented by the crashed vehicle; given a value ε, used in the DBSCAN algorithm to create the neighborhood of a node, it creates a new cluster according to the set of neighbors of the starting node; then, the central server chooses a cluster-head as the farthest node from the *starting_node* within the current cluster; this cluster-head becomes the new *starting node* in order to repeat the previous operations until the time and space limitations are satisfied (Fig.2).

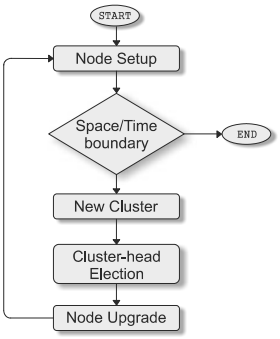


Fig. 2. Control flow for clusters creation.

Assume that each alarm message has four different fields as follows:

- the identifier (**ID**), associated to the transmitter vehicle;
- the time (**T**), when the event occurred;
- the location (**L**), where the event occurred;
- a message (**M**), which explains the event.

ID identifies the vehicle that observed the event; *T* allows to take into consideration the temporal limit, introduced in [6] and thus to broadcast an alarm message only within the limit imposed by the event average duration; *L* allows to respect the spatial limit, explained in [6] and thus to broadcast a message only to vehicles that, traveling with a certain speed, could reach the event location before it expires; *M* is a textual message containing information about the detected event.

When an event is observed, only the vehicle which detects its presence has to broadcast an alarm message without additional checks. Any other receiving vehicle does not transmit the received message automatically, but only if it is chosen as a cluster-head from the central server.

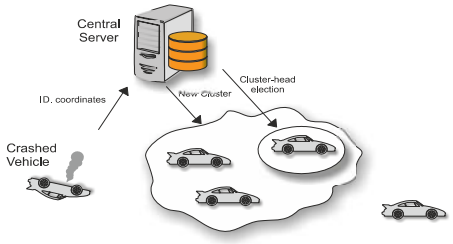


Fig. 3. Creation of a cluster in a one way road.

The vehicle that detects the event transmits its *ID* and its coordinates to the central server. The latter uses the received information in order to create clusters and elects one (or two) cluster-head (Fig. 3). The coordinates of a vehicle are represented by a pair (*longitude*, *latitude*). Let us suppose that in our Prolog database there is a *fact* that indicates the location (longitude and latitude) of a vehicle, identified by its *ID*:

```
coordinates(ID,LON,LAT)
```

When the first vehicle transmits the coordinates of the event location, the central server calculates a maximum longitude (MAX_LONG) and a maximum latitude (MAX_LAT) according to the maximum distance covered by the transmission protocol. These two values represent the ε in the DBSCAN [17] algorithm and they allow to detect the neighbors of a certain vehicle. By using the received coordinates, the central server is able to obtain information about the way (one or two ways) of the road in which the event occurred, thanks to the feature vector associated to each road. Now, the central server can create a cluster (see lines 1-5 in Algorithm 1).

Algorithm 1 Cluster-head selection.

Input: starting network configuration.
Output: clustered network.

```
1:   sameCluster(LON,LAT,WAY) :-
2:         coordinates(V,X>LONG,X<MAX_LONG,
3:                                Y>LAT,Y<MAX_LAT),
4:         member(V,List),
5:         electClusterHead(List,WAY).

6:    electClusterHead(List,WAY) :-
7:          WAY==1,!,farthestSameLane(List).

8:    electClusterHead(List,WAY) :-
9:         farthestSameLane(List),
10:        farthestOppositeLane(List).

11:  return cluster-head
```

Thanks to the *coordinates* procedure, we can select the group of vehicles V that satisfy the described properties. As shown in Fig. 4, we create a clustering tree [16], by separating a group of vehicle from another one according to their geographic coordinates and their distance from the event location.

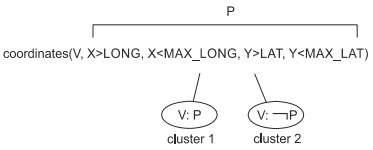


Fig. 4. Clustering tree.

After the cluster creation, the central server elects one or two cluster-head, according to the retrieved information concerning the way of the road (see lines 6-10 in Algorithm 1).

In case of a one-way road, the central server chooses a single cluster-head by choosing it as the farthest from the transmitting vehicle between the nodes of the cluster represented by *List.* Conversely, if the road is a two-way road, the central server selects two different cluster-heads: the first one is the farthest node in the *List* cluster in the same driving direction of the crashed vehicle, while the second one is the farthest node in the opposite direction.

5 Smartphones Integration

The logical model presented in [6] can be improved by also integrating the use of commercial smartphones, which in our model are supposed to belong to drivers or pedestrians along its working. Thanks to the low costs of these devices and their large spread, it is easy to assume that nowadays every vehicle is equipped with at least one smartphone, i.e. which we consider to belong to the driver.

To see the usefulness of considering smarthpones, assume that an accident occurs in a certain road and the driver loses consciousness. It would be very dangerous to wait until someone realizes that there was an accident and calls an ambulance. The smartphones connected to the Wi-Fi of the vehicles could instead automatically warn an ambulance, so that relief efforts arrive as soon as possible. Such a protocol would clear increase the road safety level.

Similarly, our model would prevent pedestrian collisions and help in case it happens. This is another important topic [21] largely investigated in the litterature. Among the others, we mention here [22] in which the authors try to calculate the location of a pedestrian through his geographic longitude and latitude, his class-speed, and other metrics. Similarly, in [23] the authors exploit some parameters such as speed and position in order to predict possible collisions. Smartphones and any other electronic device belonging to pedestrians can be used in order to predict and avoid pedestrian collisions, permitting the communication between these devices and the on-board units placed within each vehicle for the exchange of alarm messages.

6 Evaluation

In this section, we evaluate the performance of our model, by giving a closer look at the Prolog code we have provided. Precisely, we calculate the time complexity of our algorithm with respect to the cardinality of the set of neighbors of the crashed vehicle, or the first vehicle detecting the event.

The first step for the operation of the cluster-head election is the cluster creation. The idea is that the central server receives the main information about the vehicle V that observed the road exceptional event. In this way it can create a cluster containing some adjacent nodes to V, according to the distance covered by the transmission protocol. Thus, by calling A the cardinality of the list of adjacent vehicles to the vehicle V, we can say that the operation of the creation of the cluster (lines 2-4) has a time complexity $\Theta(A)$, since it consists of inserting the adjacent nodes into a list.

Then, the next operation consists of electing one or two cluster-heads for the new created cluster, represented by *List*. The cluster-head choice happens by selecting the node in *List* which is the farthest with respect to the vehicle V, according to the coordinates of each node in the cluster:

- in a one-way road, we need to elect a single cluster-head;
- in a two-way road, we need two different cluster-heads in the two opposite lane.

In both cases, a comparison between the coordinates of the nodes in the cluster is made, so the time complexity required is $\Theta(|List|)$, and thus $\Theta(A)$ (lines 5-10). Hence, a generic execution of the presented code requires a time complexity $T(A)$:

$$T(A) = \Theta(A) + \Theta(A) = \Theta(A). \tag{2}$$

So, the time complexity is linear with respect to the cardinality of *List*, and thus the cardinality of the cluster.

7 Conclusion and Future Work

The field of vehicular inter-networking technologies (VANET), with its vehicle-to-vehicle infrastructure communication, has witnessed in recent years an increasing attention from the scientific community, due to its ability of providing traffic safety and transport efficiency. VANETs are often used in decentralized structures, based on ad hoc networks, where each vehicle has the role of a node that is able to receive some information and, in turn, to broadcast them to other vehicles. The main innovative aspect of this work is the introduction of VANET in a centralized model, by introducing a central server. As we have discussed in this paper, this novelty brings some benefits such as (*i*) the possibility of using a clustering algorithm in an optimal way, since the central server can properly choose a cluster-head according to the vehicles location within the network; (*ii*) the full knowledge of the network topology and, thus, the opportunity of using the best routing protocol possible; (*iii*) the creation of an Experience Layer that is more robust than the one described in our previous model [5]; (*iv*) the chance of collecting, analyzing, and storing information about road exceptional event, which could be very useful in constructing some statistics concerning the frequency of certain events in a specific road.

We explain how the central server selects the cluster-head correctly, by exploiting the knowledge of the topology and by choosing, as cluster-head, the node that is the farthest from the source of information transmission. This is a valid technique in order to minimize the redundancy in information exchange and, thus, reduce the "broadcast storm" phenomenon. Thanks to a centralized model, this is possible without using any directional antennas or waiting time, as proposed in other works [4], [14], and [15].

We have reported on the usefulness of a smartphone integration in the system. Smartphones can improve the road safety conditions, by calling the adequate road authorities, according to the road exceptional event occurred. Smartphones can also contribute to reduce pedestrian collisions, by locating pedestrians on the road, and so predicting possible collisions before they happen.

This work can be seen as a starting point for the implementation and practical evaluation of a centralized structure for VANETs, more flexible than the decentralized one. As future work, we plan to further evaluate the effective centralized interoperability of the system model, in particular while using smartphones. To formally prove its correct behavior, we will also make use of semantical base data models [24].

Acknowledgments. This research was partially supported by Big4H (Big Data for e-Healthapplications) Research Project founded by MIUR. n.496/Ric. of 07/05/2008.

References

1. P. Caballero-Gil, C. Caballero-Gil, J. Molina-Gil. "Design and implementation of an application for deploying vehicular networks with smartphones". International Journal of Distributed Sensor Networks, 2013.
2. W. Vandenberghe, I. Moerman, P. Demeester. "On the feasibility of utilizing smartphones for vehicular ad hoc networking". ITST'11 (pp.246-251). IEEE 2011.

3. D. Li, H. Huang, X. Li, M. Li, F. Tang. “A distance-based directional broadcast protocol for urban vehicular ad hoc network”. In WiCom'07 (pp. 1520-1523). IEEE, 2007.
4. A. Benslimane. “Optimized dissemination of alarm messages in vehicular ad-hoc networks (VANET)”. In HSMC’04, (pp. 655-666). Springer Berlin Heidelberg, 2004.
5. W. Balzano, M. R, Del Sorbo, S. Stranieri. “A Logic Framework for C2C Network Management”. IEEE WAINA workshop, 2016.
6. N. Wisitpongphan, O. K. Tonguz, J. S. Parikh, P. Mudalige, F. Bai, V. Sadekar. “Broadcast storm mitigation techniques in vehicular ad hoc networks”. IEEE Wireless Commun., vol. 14, no. 6, pp. 84-94, 2007.
7. C. Shea, B. Hassanabadi, S. Valaee. “Mobility-based clustering in VANETs using affinity propagation”. In GLOBECOM 2009. (pp. 1-6). IEEE, 2009.
8. World Health Organization, “Road Traffic Injuries” and “Risk factors and what can be done to address them”, www.who.int/mediacentre/factsheets/fs358/en/, October 2015.
9. H. Hartenstein. “VANET: vehicular applications and inter-networking technologies”. Vol. 1. Chichester, UK: Wiley, 2010.
10. Research Planning, “Intelligent Transportation Systems (ITS) Standards Program Strategic Plan for 2011–2014” and “A Vision of Intelligent Transportation”, http://www.its.dot.gov/standards_strategic_plan/, 2011.
11. W. Fan, Y. Shi, S. Chen, L. Zou. “A mobility metrics based dynamic clustering algorithm for VANETs”. In ICCTA 2011, (pp. 752-756). IET, 2011.
12. A. Benslimane, T. Taleb, R. Sivaraj. “Dynamic clustering-based adaptive mobile gateway management in integrated VANET—3G heterogeneous wireless networks”. Selected Areas in Communications, IEEE Journal on, 29(3), 559-570, 2011.
13. G. Zhang, Y. Xu, X. Wang, X. Tian, J. Liu, X. Gan, L. Qian. “Multicast capacity for VANETs with directional antenna and delay constraint”. Selected Areas in Communications, IEEE Journal on, 30(4), 818-833, 2012.
14. S. Busanelli, F. Rebecchi, M. Picone, N. Iotti, G. Ferrari. “Cross-network information dissemination in vehicular ad hoc networks (VANETs): experimental results from a smartphone-based testbed”. Future Internet, 5(3), 398-428, 2013.
15. H. Blockeel, L. De Raedt, J. Ramon. “Top-down induction of clustering trees”. arXiv preprint cs/0011032, 2000.
16. H. P. Kriegel, P. Kröger, J. Sander, A. Zimek. “Density-based clustering”. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1(3), 231-240, 2011.
17. O. K. Tonguz, N. Wisitpongphan, J. S. Parikh, V. K. Sadekar. “On the broadcast storm problem in ad hoc wireless networks”. In BROADNETS'06, (pp. 1-11). IEEE, 2006.
18. J. Toutouh, J. García-Nieto, E. Alba. “Intelligent OLSR routing protocol optimization for VANETs”. IEEE Transactions on Vehicular Technology, 61(4), 1884-1894, 2012.
19. M. N. Rajkumar, M. Nithya, P. Hemalatha, R. Rajeswari. “Topology Based Routing Protocols in Vehicular Ad-Hoc Networks (VANET)”. In Avinashilingam, Two Day National Conference on Internet of Things , 200-207, 2016.
20. C. T. Barba, M. A. Mateos, P. R. Soto, A. M. Mezher, M. A. Igartua. “Smart city for VANETs using warning messages, traffic statistics and intelligent traffic lights”. In Intelligent Vehicles Symposium (IV), 2012 IEEE (pp. 902-907). IEEE, 2012.
21. M. A. Abid, O. Chakroun, S. Cherkaoui. “Pedestrian collision avoidance in vehicular networks”. In ICC'13, (pp. 2928-2932). IEEE, 2013.
22. H. Hayashi, R. Inomata, R. Fujishiro, Y. Ouchi, K. Suzuki, T. Nanami. “Development of pre-crash safety system with pedestrian collision avoidance assist”. ESV’13 (p1-5), 2013.
23. O. K. Tonguz, W. Viriyasitavat. “Cars as roadside units: a self-organizing network solution”. In Communications Magazine, IEEE 51.12: 112-120, 2013.
24. F. Amato, V. Casola, A. Gaglione, A. Mazzeo. “A semantic enriched data model for sensor network interoperability”. Simulation Modelling Practice and Theory 19(8): 1745-1757, 2011.

Chain-of-Trust for Microcontrollers using SRAM PUFs: the Linux Case Study

Domenico Amelino, Mario Barbareschi, Antonino Mazzeo

Abstract Many security challenges have emerged from what is defined as Internet of Things (IoT), due to the inherent permanent connection of devices involved in networks. Furthermore, IoT devices are often deployed in unattended working environment and, hence, they are prone to physical attacks. Attackers take advantages of such weaknesses to clone devices, tamper the software installed on them and extract cryptographic keys. In this paper, we propose a technique to exploit Static Random Access Memory based Phisical Unclonable Functions to have available a chain-of-trust on a microcontroller device. We prove its effectiveness in terms of reliability and required overhead by introducing a case study based on the STM32F7 device running the Linux operating system.

1 Introduction

The Internet of Things (IoT) identifies a network of several digital components equipped with sensors and actuators [26, 1]. IoT devices interact and cooperate each other for creating smarter environments/spaces for several field of applications, such as transportation, cities, health-care, energy, tourism, and industry. They are spreading out thanks to technological availability of advanced devices and smarter tools for application developing [12, 13]. These devices are permanently connected to the network, exchanging data over public channels, typically wireless. As anyone may be able to inject altered data or attack communication protocols, it is imperative for critical applications to adopt security mechanisms to avoid exchanged of compromised information. In particular, devices have to guarantee confidentiality, authenticity and integrity of the data [25]. To protect their secrecy and guaranteeing reliability for the applications, cryptography is a powerful mean its strength relies on key(s) generation and management [14, 3, 2]. Indeed, keys involved in encryp-

DIETI - Department of Electrical Engineering and Information Technologies
University of Naples Federico II e-mail: firstname.lastname}@unina.it

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_72

tion/decryption schemes have to be protected against (physical) tampering attacks and unwanted accesses, hence they have to reside in a secure perimeter [15, 8]. Such a perimeter can be considered as a hardware region which protects key material from the information leak and from physical attacks.

Moreover, an attacker which has physical access to the device clone the software, copying it on other identical devices. This way he/she will be able to introduce a malicious node into the network.

In this scenario, literature offers several solutions to improve the security of IoT devices [26, 7]. Since the limited amount of resources available, introducing new trustworthy and lightweight solutions becomes a significant challenge.

A novel approach, proposed in the scientific literature, is based on the generation of security keys from device intrinsic physical characteristics, rather than storing this information in an NVM. Manufacturing process of Integrated Circuits (ICs) inherently imprints variations that are uncontrollable, random and unique. Such imperfections can be exploited to define Physical(ly) Unclonable Functions (PUFs) [18, 23], which measure physical flaws to extract binary responses, so as to be used as hardware fingerprints. They are able to guarantee the trustworthiness of a software execution environment providing key material, key generation and storage features. The great advantage of intrinsic physical characteristics is their strict coupling with the device.

In the embedded domain, the spread of the PUF finds some obstacles due to its noisy nature [21]. Indeed, PUF responses are not stable, requiring the employing of post-processing operations, which can introduce a huge overhead for their implementation [17].

A promising PUF architecture is represented by the static random access memory (SRAM) cell, since is available in many ICs as technology for volatile memory, such as in microcontrollers [19]. The SRAM PUF makes use of the binary pattern generated by the initial states of memory cells.

SRAM technology is used for implementing volatile memories for processors and its content is lost when they are powered-off. An SRAM cell is designed as two cross-coupled symmetric halves and each one contains an inverter. During power-up, the SRAM cell is characterized by an unstable state, but due to mismatches between two halves, it is forced to a logic-0 or logic-1 state by the inherently positive feedback, which amplifies even small voltages variations. The mismatch is imprinted by the manufacturing process and the initial preferred state of each cell cannot be predicted. Therefore, the generated pattern can be used to retrieve unique bit-sequence responses, even though SRAM needs to be in a pristine state, i.e. no writing operations should be performed on it, until the pattern is someway extracted. SRAM PUF guarantees random, unclonable and unpredictable responses and these properties fulfill what is required for keys eligible to be used in cryptographic schemes. Though, in order to exploit these benefits, the start-up pattern needs to be processed such that the PUF provides stable values [9].

In this paper we propose a key generation mechanism based on exploiting the SRAM PUF security properties for the IoT domain and detailing the process to generate a secret and trusted key from PUF responses. The key extraction is based on

lightweight error correction code and crypto-engine, suitable for embedded devices equipped with scarce computational resources. As case-study, we illustrate a working prototype of chain-of-trust which makes use of the PUF in order to run secured software and services over a Linux kernel.

2 Chain-of-Trust on Microcontrollers

In computer security, the verification and validation of hardware and software components can be fulfilled by employing a particular cryptographic scheme, which is known as chain-of-trust. This technique is based on a step-wise verification, which starts from a trusted component, typically implemented on the hardware, named root-of-trust, and goes up for a different number of layers. From bottom up, the chain-of-trust is established by running only software components which are digitally signed with a key and each of them is the verifier of the next one in the chain. The final software can be trusted to have specific security properties, because if any of the previous layer have been tampered, one of the verified signature would be invalid and previous layers would not have executed it. Otherwise, the last verified layer can be trusted as it would not have been executed if its signature has been invalid [5].

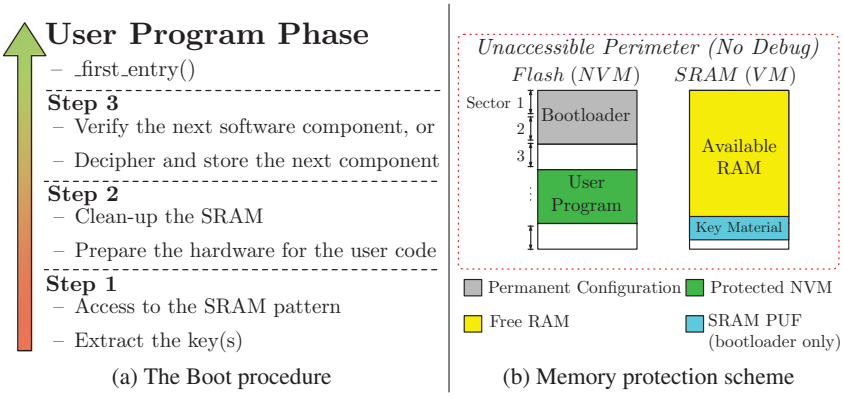
The approach of running and verifying software in sequence allows to have a unique anchor, the root-of-trust, and allows also to have keys spread in each layer instead of storing involved keys on hardware NVM. Moreover, the approach guarantees flexibility in replacing any layer of the sequence with another one, without go all the way back to the root.

From previous considerations, it follows that the critical component is the root-of-trust, as it must guarantee trustworthiness and security for all the layers in the chain. Such features are typically implemented by a trustworthy component, called Trusted Platform Module (TPM), which can be inserted into a system design to offer hardware security primitives. In particular, TPM allows to build a chain-of-trust in which each component of the system, once validated, can validate itself another component, continuing the trusted sequence offering secrecy and integrity of the code running on a device for security applications.

TPMs have been designed for high-level resourced architectures, such as Desktop PC or Servers, but they require a technological effort and additional hardware resources such that are not very common on embedded architectures.

Generally, a microcontroller contains a processor core, programmable input/output peripherals, storage memory and a limited amount of volatile memory, usually implemented adopting SRAM technology. Minimal capabilities required by the chain-of-trust scheme are a secure perimeter in which have available cryptographic keys and a cryptographic algorithm suitable for being ran on the target.

With regards to the secure storage, the PUF is an ideal component since, as illustrated in the previous Section, it guarantees very attractive properties, such as tamper-evidence, uniqueness, randomness, and so forth. Moreover, the key gener-



(a) The Boot procedure (b) Memory protection scheme

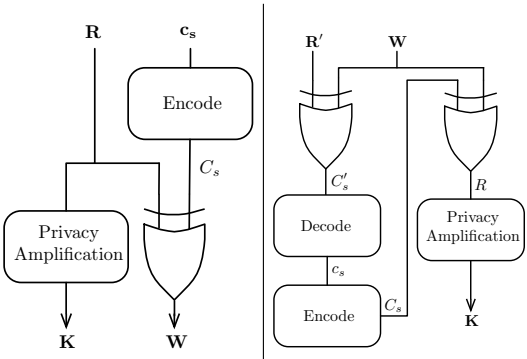
Fig. 1: The PUF-based Boot Sequence and the Memory protection scheme

ated by a PUF is not stored in an NVM, since it is generated from an physical properties of a device, and hence, when the device is switched off, none is able to extract information about it. As we previously stated, the volatile memory of the processor is typically realized with the SRAM technology, which is eligible to be a PUF. The implementation of key extraction algorithm from SRAM PUF has to be suitable for being executed on an embedded system, taking into account resource constrains.

The building of chain of trust on microcontrollers can be realized implementing a secure boot procedure in which the first executed software, that is inherently the root-of-trust, component is secured against modification by third-parties and the key is securely stored in a secure perimeter which is unaccessible from the outside [7].

2.1 Fuzzy Extractor

SRAM PUF guarantees random, unclonable and unpredictable responses and these properties fulfill what is required for keys eligible to be used in cryptographic schemes. Though, in order to exploit these benefits, the start-up pattern needs to be processed such that the PUF provides stable values [9]. Among available post-processing techniques introduced in the literature, the fuzzy extractor is the most effective in recovering stable values from noisy responses [20]. It employs an error correction code (ECC) to realize two main primitives: enrollment and information reconciliation. These two are described in Figure 2. The enrollment procedure (Figure 2a) is executed before the device issuing and aims to extract a key K and produces the helper data W. The key is obtained by the PUF response through the privacy amplification function, that is a one-way function, while the helper data is obtained by masking the response with a codeword C_s. The codeword must be generated by a random decoded symbol c_s and kept secret.



(a) Enrollment phase. (b) Information reconciliation.

Fig. 2: The two primitives of the fuzzy extraction mechanism.

Once the device is issued, it can execute the information reconciliation (Figure 2b) which provides the same key K, previously extracted, but, this time, from a noisy version of the PUF response, namely R'. The helper data is masked with R' to get back a noisy version of the codeword C_s (indicated as C'_s), which can be recovered by using the ECC. Once C_s is available, the device is able to obtain, by masking W with C_s, the response originally extracted during the enrollment phase. From the previous scheme, it is clear that W is not a critical data and can be publicly exchanged and stored.

2.2 Implementing the chain-of-trust by employing PUFs

In a chain-of-trust scheme, the root-of-trust is certainly the component which realizes the boot of the system. In case of the usage of PUF for extracting keys, the bootloader is responsible for such activity. This is due the fact that the SRAM has to be in a pristine state in order to properly run the key extraction. Moreover, a portion of the available SRAM has to be reserved, i.e. to be protected by potential overwrites, until keys are produced. The fulfillment of this constraint is mandatory for our scope and limits the availability of volatile memory resources during the boot phase. Furthermore, the boot code has to be stored in a protected area, as it must be not possible to alter or bypass its instructions [4].

The boot procedure has to operate only with volatile data residing in the SRAM, hence it must be stateless, that is each time it be executed by the target does not consider information about previous executions. This is due to the fact that it has to not save any data in the NVM for avoiding any leak of some critical information about the key extraction procedure, such as intermediary variables of a fuzzy extraction scheme (cf. subsection 2.1). Furthermore, the SRAM has to be cleaned in order to remove any information about the previous state of the memory. Nonetheless, it is

worth noting that this requirement may be not fulfilled if the key is used in a public scheme, as its disclosure does not involve any security issue, for instance using the key as identification bitstring.

Figure 1a shows main steps for realizing the chain-of-trust over a microcontroller by designing a secure bootloader and adopting a PUF. The first phase consists in the extraction of the key by means of a post-processing technique. During the second phase, the boot loader has to prepare the hardware in order to execute the first component in the chain; additionally the SRAM has to be cleared. Finally, during the third phase, the bootloader runs the verification step, which can additionally involve a decryption of the software component. At the end, a jump to the entry point of the user code is performed. Figure 1b shows a solution for a scheme of memory protection. The NVM and SRAM resides within a secure perimeter, which is part of the microcontroller. An image of the bootloader is permanently stored on the flash memory to be the first entry at start-up, which is protected as none can access it from outside, except the processor. The flash memory is logically split in sectors and each of them has to be protected by a hardware mechanism to prevent any write operation.

The bootloader itself do not contain the key used to start the verification process of the chain-of-trust, but the key is physically generated from the SRAM start-up pattern. Another protected section of the flash can store the user program, which can be verified by the bootloader and, for instance, stored into the SRAM of the processor for being executed. Preventing any debug operation the SRAM, and in particular on flash sectors, the user program can be still updated under the control of the user code itself which is validated during the booting procedure.

3 Chain of Trust on STM32F7: The Linux Case-Study

In this Section we describe a real scenario in which we implement a secure bootloader as root of trust using an STM32F746NGH6, a microcontroller which includes a 216 MHz ARM Cortex-M7, 1 Mbytes of flash memory and 340 Kbytes of RAM.

As previously stated, the design of a chain-of-trust using PUFs needs an implementation of fuzzy extractor algorithm which must be as lightweight as possible. First of all, in order to extract a key of a specific length from the SRAM, we need to define the amount of source bits needed by the fuzzy extractor scheme. Such calculation mainly depends on the entropy which characterizes the SRAM cells of the memory circuit, but also on the bit error probability and on the adopted ECC. In particular, it is important to introduce an effective entropy estimator to verify the unpredictability, or randomness, or guessing probability of a bit sequence, which is significantly important as we intend to extract keys from the SRAM pattern. Obviously, in order to use SRAM bits as key material, its entropy should be as high as possible.

A good practice to estimate the entropy of a source of bits is calculating the min-entropy, that is a lower bound for the entropy and represents the worst-case measure

of uncertainty for a random variable [16]. We refer to the NIST specification 800-90 for min-entropy calculation [10]. For a binary source, the possible outcomes are 0 and 1, which are characterized by occurrence probabilities of p_0 and p_1 respectively. We consider p_{max} as the maximum value of these two probabilities, min-entropy can be calculated as $H_{min} = -log_2(p_{max})$.

Assuming that each bit of SRAM is independent from others [6], the total min-entropy of N bits can be calculated as a summation of the min-entropy of each individual bit, as it follows:

$$(H_{min})_{total} = \sum_{i=1}^{N} -log_2 (p_{imax}) \tag{1}$$

The value p_{imax} can be derived by sampling SRAM bits from different devices. Indeed, sampled bits are bitwise concatenated, and the Hamming weight (HW) for each bit, i.e. the number of ones, is calculated. The i-th HW can assume a value between 0 and the amount of samples M. Then, p_{imax} is equal to HW_i/M if $HW_i/M > 0.5$ or to $(1 - HW_i/M)$ otherwise. Once the estimation is done, the average min-entropy per bit is calculated by dividing $(H_{min})_{total}$ by the length N. Involving ten different devices, we evaluated the worst average min-entropy per bit, which turned out to be 0.755. Therefore, in order to extract a key of size L, at least $\lceil L \div 0.728 \rceil$ source bits are needed. With respect to the bit error probability, we used the same samples acquired for the min-entropy calculation. In particular, we estimated the bit error probability over all memory dumps performed on devices. Each comparison led to a different bit error probability, ranging from 0.04 to 0.06. In order to be conservative, we chose the worst value, i.e. 0.06, and we considered an additional value of 0.01, as a secure conservative threshold for the bit error probability. Based on these values, $\lceil 128 \div 0.728 \rceil = 176$ bits are needed to extract a key of 128 bits. For the fuzzy extractor primitives implementation, we consider Reed Muller ECC scheme, Speck cipher [11] for encryption and decryption and a Davis-Mayer scheme on Speck [22] for implementing the privacy amplification. We choose those techniques for reducing the resource demand of the mechanism and limiting the footprint of the extraction algorithm.

Adopting the fuzzy extractor we obtain a stable and reliable high-entropy key from the microcontrollers, eligible to be used for decrypting and validating software components. Indeed, the probability to extract a different key with a Reed Muller scheme of first order and 2^7 as code-length, amounts at $3.278934e^{-09}$.

In order to have the secure boot loader as the root-of-trust, we need a secure perimeter on which it has to be stored. Furthermore, the bootloader has to be the only handler of the SRAM PUF and the only key manager for the node and has to be first one which accesses the memory, in order to have the SRAM in a pristine state. This last point is a fundamental requirement to ensure the correctness of the PUF response.

STM32F7 microcontroller offers a memory protection mechanism to secure the perimeter with different memory read/write restrictions [24]. In particular, we adopt the level 2 read protection, which forbids all external accesses to flash sectors and

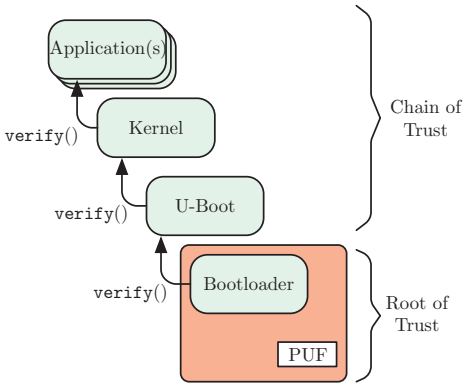


Fig. 3: Secure Booting Sequence

also disables every debug interface, while level 2 write protection prevents flash memory overwrites. A joint use of these two approaches protects the boot loader ensuring its integrity. As the level 2 is a permanent configuration, it follows that the key from the PUF must be extracted before applying the protection mechanism.

The role of the bootloader is to load one or more applications which reside outside the secure perimeter and the PUF mechanism is used to extend trustworthiness to them by means of a key physically dependent to the device. The key is used to decrypt and/or verify a user application image. In case of success, the boot loader prepares to run the user application.

In our case study, we develop a trusted platform which loads a version of uCLinux, a lightweight Linux distribution suitable for ARM Cortex-M. In particular, we use a Linux (uClinux) software of the Emcraft Systems, releases for the STM32F7-Discovery board, associated with a U-Boot configuration under GPL License.

The chain-of-trust which will be ran on the device is depicted in Figure 3. The root-of-trust, namely the bootloader, verifies the U-Boot image by using the key extracted from the PUF. Hence, the U-Boot must be digital signed by using the key from the PUF.

If the verification succeeds, the U-Boot image has to be mapped into microcontroller memory such that it has not be stored into the first flash sectors (which are reserved to the boot loader and cannot be overwritten due to memory protection) or the SRAM area on which PUF responses are extracted. We enrich U-Boot with some primitives that implements what is necessary to verify a uCLinux image. Conversely, in order to have a smaller U-Boot footprint, the verification can be demanded to the bootloader code.

As for the kernel, we store the uCLinux image on the external flash memory of the STM32F7-Discovery. It is signed with the same key previously extracted from the PUF, even though another solution is possible. Indeed, by ciphering the U-Boot with the PUF key, the end-user may define a custom key that can be included into the U-Boot image, as it can be considered a secure storage. Indeed, only entering into

possession of the PUF key enables to decipher the U-Boot image and, hence, obtain the key. This mechanism gives trustworthiness to uCLinux allowing the developing of secure applications.

In order to continue the chain-of-trust, the kernel can use or the SRAM PUF key or, in case of it is ciphered another key defined by the end user and stored into the software image.

3.1 Preliminary Experimental Results

In order to evaluate the overhead introduced with the chain-of-trust, we prepare the previously described set-up and we instrumented each software layer to measure the time needed for completing it.

The information reconciliation algorithm, executed by the bootloader, takes 75 ms to provide the key extracted from the SRAM pattern, while decrypting U-Boot by means of Speck takes about 2 s. The footprint of the bootloader is 16 KB, while U-Boot requires 87 KB, that is the 10% of the whole available flash memory on the device.

4 Conclusion

The needing of security at both device and network level is still an open challenge in the IoT domain. This is due to tight constraints which characterized the design of any security mechanism on IoT devices. In this paper, we showed an architecture for securing a microcontroller based devices without introducing additional costs or hardware functionality, thanks to the adoption of the SRAM PUF. Indeed, we detailed the methodology to extract random, reliable, unclonable, device-dependent keys and, thanks to the PUF nature, tamper evident keys. Then, we proved that a bootloader on a microcontroller is able to be the trust anchor for defining a chain-of-trust, whereby developing trustworthy applications. In particular, with few experimental results, we described a practical implementation of a chain-of-trust on a STM32F7 Discovery board running a uCLinux kernel.

References

1. Amato, F., Barbareschi, M., Casola, V., Mazzeo, A.: An fpga-based smart classifier for decision support systems. In: Intelligent Distributed Computing VII, pp. 289–299. Springer (2014)
2. Amato, F., De Pietro, G., Esposito, M., Mazzocca, N.: An integrated framework for securing semi-structured health records. Knowledge-Based Systems 79, 99–117 (2015)
3. Amato, F., Moscato, F.: A model driven approach to data privacy verification in e-health systems. Transactions on Data Privacy 8(3), 273–296 (2015)

4. Amelino, D., Barbareschi, M., Battista, E., Mazzeo, A.: How to manage keys and reconfiguration in wsns exploiting sram based pufs. In: Intelligent Interactive Multimedia Systems and Services 2016, pp. 109–119. Springer (2016)
5. Bajikar, S.: Trusted platform module (tpm) based security on notebook pcs-white paper. Mobile Platforms Group Intel Corporation pp. 1–20 (2002)
6. Barbareschi, M., Battista, E., Mazzeo, A., Mazzocca, N.: Testing 90nm microcontroller sram puf quality. In: Design & Technology of Integrated Systems In Nanoscale Era (DTIS). pp. 1–6. 2015 10th IEEE International Conference On. IEEE (2015)
7. Barbareschi, M., Battista, E., Mazzeo, A., Venkatesan, S.: Advancing wsn physical security adopting tpm-based architectures. In: Information Reuse and Integration (IRI), 2014 IEEE 15th International Conference on. pp. 394–399. IEEE (2014)
8. Barbareschi, M., Cilardo, A., Mazzeo, A.: Partial fpga bitstream encryption enabling hardware drm in mobile environments. In: Proceedings of the ACM International Conference on Computing Frontiers. pp. 443–448. ACM (2016)
9. Barbareschi, M., Di Natale, G., Torres, L.: Ring oscillators analysis for security purposes in spartan-6 fpgas. Microprocessors and Microsystems (2016)
10. Barker, E., Kelsey, J.: Nist special publication 800-90a: Recommendation for random number generation using deterministic random bit generators (2012)
11. Beaulieu, R., Shors, D., Smith, J., Treatman-Clark, S., Weeks, B., Wingers, L.: The simon and speck lightweight block ciphers. In: Proceedings of the 52nd Annual Design Automation Conference. p. 175. ACM (2015)
12. Cilardo, A.: New techniques and tools for application-dependent testing of FPGA-based components. IEEE Transactions on Industrial Informatics 11(1), 94–103 (2015)
13. Cilardo, A., Fusella, E., Gallo, L., Mazzeo, A.: Exploiting concurrency for the automated synthesis of MPSoC interconnects. ACM Transactions on Embedded Computing Systems 14(3) (2015)
14. Cilardo, A., Mazzeo, A., Romano, L., Saggese, G.: An FPGA-based key-store for improving the dependability of security services. pp. 389–396 (2005)
15. Cilardo, A., Barbareschi, M., Mazzeo, A.: Secure distribution infrastructure for hardware digital contents. IET Computers & Digital Techniques 8(6), 300–310 (2014)
16. Claes, M., van der Leest, V., Braeken, A.: Comparison of sram and ff puf in 65nm technology. In: Nordic Conference on Secure IT Systems. pp. 47–64. Springer (2011)
17. Dodis, Y., Reyzin, L., Smith, A.: Fuzzy extractors: How to generate strong keys from biometrics and other noisy data. In: International Conference on the Theory and Applications of Cryptographic Techniques. pp. 523–540. Springer (2004)
18. Gassend, B., Clarke, D., Van Dijk, M., Devadas, S.: Silicon physical random functions. In: Proceedings of the 9th ACM conference on Computer and communications security. pp. 148–160. ACM (2002)
19. Guajardo, J., Kumar, S.S., Schrijen, G.J., Tuyls, P.: Fpga intrinsic pufs and their use for ip protection. In: International workshop on Cryptographic Hardware and Embedded Systems. pp. 63–80. Springer (2007)
20. Linnartz, J.P., Tuyls, P.: New shielding functions to enhance privacy and prevent misuse of biometric templates. In: Audio-and Video-Based Biometric Person Authentication. pp. 393–402. Springer (2003)
21. Maes, R., Tuyls, P., Verbauwhede, I.: Intrinsic pufs from flip-flops on reconfigurable devices. In: 3rd Benelux workshop on information and system security (WISSec 2008). vol. 17 (2008)
22. Menezes, A.J., Van Oorschot, P.C., Vanstone, S.A.: Handbook of applied cryptography. CRC press (1996)
23. Pappu, R., Recht, B., Taylor, J., Gershenfeld, N.: Physical one-way functions. Science 297(5589), 2026–2030 (2002)
24. STMicroelectronics: RM0090 Reference manual (10 2015)
25. Yan, Z., Zhang, P., Vasilakos, A.V.: A survey on trust management for internet of things. Journal of network and computer applications 42, 120–134 (2014)
26. Zhao, S., Zhang, Q., Hu, G., Qin, Y., Feng, D.: Providing root of trust for arm trustzone using on-chip sram. In: Proceedings of the 4th International Workshop on Trustworthy Embedded Devices. pp. 25–36. ACM (2014)

Part VI
Workshop DEM-2016: 3rd International Workshop on Distributed Embedded Systems

Hybrid Approach on Cache Aware Real-Time Scheduling for Multi-Core Systems

Thomas Huybrechts, Yorick De Bock, Haoxuan Li and Peter Hellinckx

Abstract The Worst-Case Execution Time of a task is important in real-time systems. This metric is used by the scheduler in order to schedule all tasks before their deadlines. However, cache memory has a significant impact on the execution time and thus the WCET. Therefore, different cache analysis methodologies exist to determine the WCET, each with their own advantages and/or disadvantages. In this paper, a new hybrid approach is proposed which combines the strengths of two common analysis techniques. This hybrid methodology tackles the problem that can be described as '*the gap between thinking like a machine and thinking like a human being*'. The two-layer hybrid model splits the code of tasks into so-called *basic blocks*. The WCET can be determined by performing execution time measurements on each block and statically combining those results. The COBRA-HPA framework is specially developed to facilitate the process of generating a hybrid block model and corresponding source files for time measurements. Additionally, the framework is able to generate timed automata models for UPPAAL. In conclusion, the results show that the block size has a great influence on the performance of the hybrid analysis. Thus, future work will focus on improving the hybrid model and determining the optimal size of the blocks.

1 Introduction

Caches are small memory elements which are integrated close to the processor unit. The fast cache memory will fetch data and instructions of the main memory. As a result, it provides a tremendous improvement of the average execution time compared to the slower RAM or storage devices. However, the fast cache memory has a high

Thomas Huybrechts · Yorick De Bock · Haoxuan Li · Peter Hellinckx
MOSAIC, University of Antwerp, Belgium, e-mail:

thomas.huybrechts@uantwerpen.be · yorick.debock@uantwerpen.be
haoxuan.li@uantwerpen.be · peter.hellinckx@uantwerpen.be

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_73

cost per bit and the space close to the processor is limited. Therefore, the maximum cache size is limited to a few MBytes or KBytes for embedded systems to keep the cost affordable. This allows for a limited amount of instructions of a program can be loaded into the cache. Each cache line contains a block of multiple data elements which includes the data or instruction of interest. These cached blocks are very efficient due to the sequential order of instructions in the program flow. This concept is referred to as the locality principle [8].

When the processor requests an instruction, it will first check within the cache if the instruction is present. A cache hit occurs when the instruction is found. In contrast to the slow main memory, the data can now be rapidly returned to the processor in just a few clock cycles. In the opposite case, the cache will miss and thus the instruction needs to be retrieved from a higher level of the memory hierarchy. For instance, L2/L3-cache, RAM or HDD. Each cache miss adds an extra delay to the execution time. This results in a higher variance in the distribution of the program's execution speed (figure 1) [4] [10].

When looking at a multi-core processor, multiple layers of cache memory are introduced. A first level of cache is placed close to each core. This level has a similar functionality as the single-core caches. A second layer of cache is shared between the cores to improve the performance of a multithreaded application working on the same data (cached memory blocks) or the communication between the different cores using these caches [12].

It is important for real-time systems that the different tasks can be scheduled by the scheduler, so that the deadlines of the tasks can be achieved. In order to determine the schedulability of a set of tasks, a timing analysis is performed to calculate the Worst-Case Execution Time (WCET) of each task. In practice, margins are added to the calculated upper and lower bounds of the timing distribution because of the complexity of the cache analysis. As a result, the influences of the cache are taken into account (figure 1) [1]. In this methodology, a cache miss will always be assumed when calculating the upper bound. This pessimistic attitude is not realistic and will lead to over-proportioned systems [1]. In order to approximate the real distribution, one should evaluate the influence of cache memory during timing analysis.

In the current state of the art, there are three main strategies to determine the WCET with the influence of cache. Each of these has their own advantages and disadvantages. A first methodology proposes to lower the non-determinism of the system by adapting the hardware and software components, e.g. pre-allocated caches and deterministic architectures [4] [11] [13]. However, this strategy is not acceptable, since it will eventually have an impact on the performances of the system. A second methodology describes the static cache analysis of the code and architecture to achieve an accurate WCET [7] [8] [10] [14]. However, this method is unusable as the complexity increases. The last methodology is a time measurement based approach of the WCET. This solution is computationally acceptable as the amount of measurements is decided by the researcher. The accuracy on the other hand will drop dramatically when the amount of measurements decreases [8] [10].

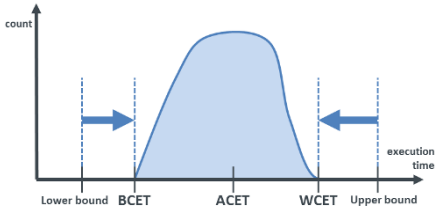


Fig. 1 A distribution of execution times.

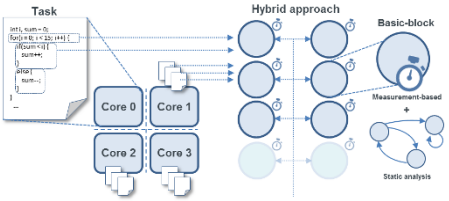


Fig. 2 Hybrid cache analysis methodology.

We believe the solution lies in a hybrid approach. Certainly, when considering caches. In our current research, we examine this new hybrid methodology and start the development of a framework on which we are able to run code analysis to verify our thoughts.

2 Hybrid Methodology

In order to overcome the shortcomings, we suggest a hybrid approach to calculate the WCET of a task. The main problem can be described as '*the gap between thinking like a machine and thinking like a human being*'. We, humans, think in algorithms and want to do estimates based on the structure of the algorithm. Machines think in bits and bytes and in this case more specifically in block sizes handled by the cache replacement algorithms. This hybrid approach will combine a time measurement based approach at machine level with a static analysis approach at algorithm level (figure 2). The static analysis fits in well with the structured algorithmic thinking from the human. Likewise, the time measurements reduce the complexity for the machine. A schematic overview of this methodology is shown in figure 2.

The hybrid model splits the code of a set of tasks into so-called *basic blocks*. Each basic block contains one path trace of instructions with a single set of inputs and a single set of outputs. The size of a basic block varies between the largest possible trace meeting the constraints of a single set of inputs and outputs over a single memory block, i.e. a unit of memory that is replaced by a cache replacement algorithm, to a single program instruction.

The challenge of this two layer hybrid approach is tackling the computational complexity problems within the static analysis layer and the accuracy within the measurements based layer. In other words, a proper balance needs to be found between the two layers of the hybrid model. This goal introduces four main challenges.

1. Finding a right block size in order to keep the complexity of the static analysis contained and to keep the accuracy of the conducted measurements, i.e. more blocks means more complex static analysis and less blocks means less accuracy in the measurement based approach;

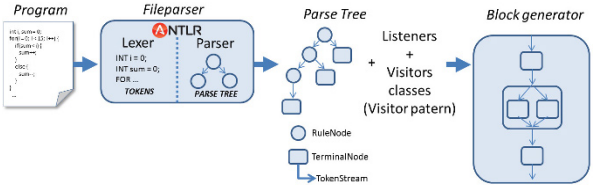


Fig. 3 Schematic overview of the COBRA-HPA framework.

2. How to reduce the required amount of measurements while keeping a high accuracy based on the information from the static analysis layer. This can be achieved by reducing the measurement space using the static analysis information, e.g. excluding non-existing concurrencies;
3. Keeping the resulting WCET low by relying on a correct subset of all WCETs. This can be accomplished by adding additional information to the measurements in order to use only the relevant subset of measurements. Only those results fitting the situation in the current state of the static analysis will be used, e.g. not using correlation with start-up procedures in a shutdown phase;
4. Exploring the influence of the number of measurements conducted and their errors on the static analysis and the total WCET analysis. As a result, we can make the right trade-off between the measurement cost and the required accuracy depending on the specific application domain.

3 COBRA-HPA Framework

In the hybrid methodology discussed in chapter 2, the main idea is splitting the program code into basic blocks on which we can perform time measurements. In order to automate this block creation, we are creating a framework to help us verify our theories and finally allowing us to perform accurate WCET calculations.

The COBRA framework, *COde Behaviour fRAmework*, developed by the MOSAIC research group is a tool to perform various types of analysis to examine the performance and behaviour of code on different architectures. In order to implement the hybrid approach, we created a prototype extension for the COBRA framework, the *Hybrid Program Analyser* or COBRA-HPA. This extension will parse a given C-source file and create a block tree on which we can perform algorithms or apply rules to. In the next step, the framework creates corresponding source files of each block which can be used to run measurement based analysis on. A schematic overview of the COBRA-HPA framework is shown in figure 3. The different modules in the framework are explained in the following sections of this chapter based on an example program.

File Parsing The first stage in creating a block tree is parsing the code file. In this stage we need to interpret the source code to understand the structure of the code. In order to perform this task, we have chosen for the ANTLR v4 framework. ANTLR, *ANother Tool for Language Recognition*, is an open-source tool which can parse a text file according to a given grammar file [9].

The parsing process is done by two components, the *lexer* and *parser* as shown in figure 3. The lexer cuts the character stream into a series of meaningful tokens as defined in the grammar file [9]. These tokens are fed to the parser unit. The parser will then try to identify the parts of speech according to syntactic rules in the grammar file [9]. The final result is a parse tree with the different tokens on its leafs and all rules on the intermediate nodes which where applied for each token [9].

A major advantage of the ANTLR framework is its language independence which allows us to change to another programming language simply by loading the corresponding grammar file. Additionally, the automatic generated listener class enables us to walk through the tree and facilitates the integration of the ANTLR framework with our own.

Block Generator The second step in the generation of the block tree is walking the generated parse tree from the file parser. As previously mentioned, ANTLR v4 implements a visitor pattern with the automatic generated listener class [9]. The pointer object called the walker starts at the root of the tree. It will then travel from a node to the next child node in order, i.e. left to right. This walk pattern is also referred to as a depth-first search [9]. Each intermediate node contains a rule that was applied to a token. When the walker enters or leaves a node, it will notify the framework and trigger a listener event. In order to make the connection between the ANTLR framework and the COBRA-HPA framework, we have subclassed the empty listener class and overwritten the methods of interest. For example, when entering an iteration statement a message will notify the block generator that an iteration statement is detected. All the following statements will be contained inside the body of the created iteration block. The walker will publish an exit message upon leaving the iteration statement to tell the block generator that the block is completed and thus no further actions are required for this particular block.

The block tree used in the COBRA-HPA framework is generated on-the-fly when the parsed tree is traversed by the walker. The result is a tree structure which is built out of blocks. Each block initially contains one instruction or statement of the source code. The blocks are arranged from left to right to their sequential order in the code. Instructions that reside inside a statement body are added as child '*blocks*' to the parent block. These instructions are characterised by their indentation in most programming languages. In this first prototype, we created seven types of blocks in which we categorize each instruction. These blocks are shown in the class diagram in figure 4.

In figure 5, a block tree is rendered from the code of listing 1 to illustrate the structure of a block tree. The attentive reader will notice that the first basic block in figure 5 contains not one but two instructions. This aggregation of two blocks is the result of another implemented feature in the COBRA-HPA framework. This feature is called the *Block Reduction Rule* system.

In chapter 2, we discussed the size of the blocks in our hybrid approach. We need to find a perfect balance to keep the analysis reliable and computable. The default block tree generated by the block generator will always contain one instruction for each block. In order to change the block size, we need to group the smaller blocks

```
int main(int argc, char* argv){
 int var1, var2;
 int array[10];
 if(argc == 1)
  var2 = atoi(argv);
 else
  var2 = 1;
 for(var1=0; var1<10; var1++){
  array[var1] = var1 * var2;
  if(array[var1] > 100)
   break;}
 return 0;}
```

Listing 1 Sample program as an example input file for the COBRA-HPA framework. Source is written in the C-language.

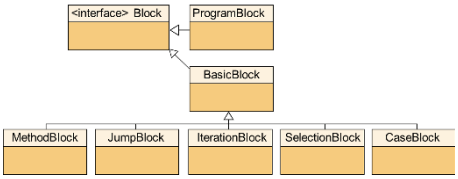


Fig. 4 Class diagram of the COBRA-HPA block types.

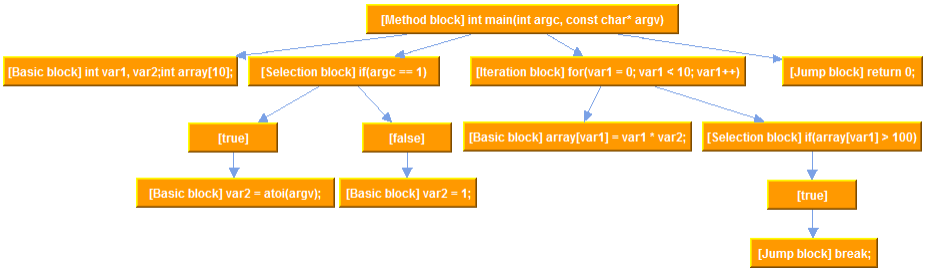


Fig. 5 Block tree with COBRA-HPA of the sample code in listing 1.

into bigger blocks. Therefore, the framework has the Block Reduction Rule system based on rules which allow the user to reduce the number of blocks by creating bigger blocks. When the block generator is finished, the operator has the option to apply a set of rules on the model. The current version of the system has two rules implemented, i.e. basic block reduction and abstraction reduction. The first rule searches for successive basic blocks and replace those by one bigger basic block. The second rule groups all blocks starting from a user-defined depth and thus creates an abstraction from the lower level instructions. In the future, new rules can be added to the system.

Export for Analysis The next step after loading a program file in the system is running time measurements on the created blocks. The COBRA-HPA framework simplifies this process by providing different export methods. The first tool is automatic generated code templates of the blocks for time measurements. The second tool generates a XML project file for the UPPAAL Model Checker [6].

Block Source File for Time Measurement This first tool allows the operator to generate code files according to a default template. Each automatic generated file will include a main method to make the source file executable and a method with the instructions of a block. Before performing the measurements, we need to add code

dependent variable declarations and include header files to make the code runnable. The source file generator is also compatible with timed automaton nodes created with the UPPAAL model exporter.

Timed Automata with UPPAAL In addition to the hybrid methodology discussed in chapter 2, the research group MOSAIC wants to explore other methodologies to determine the WCET. One of these studies is about determining the WCET with probabilistic analysis. For this research, the theory of timed automata is applied. In order to support the generation of models for the experiments, we implemented a feature to create these automata.

A timed automaton is a finite-state machine extended with the notion of time [2]. The model contains clock variables which are in synchronization with each other [2]. Therefore, the transitions in this model can also be guarded by comparing clock variables with numeric values [2].

For modelling the timed automaton, we use the model-checker UPPAAL [6]. This toolbox developed by Uppsala University and Aalborg University is designed to verify systems which can be represented as a network of timed automata [2]. The purpose of using this tool is to model the program flow into a timed automaton. By implementing the UPPAAL model exporter, we are able to create timed automata.

First of all, we start by creating a timed automaton data model. Therefore, we translate our hybrid block model to the syntax of timed automata. Each hybrid block is directly mapped to a node in the automaton. The next step is generating the necessary links between the nodes. These links are the transitions which represent the program traces that can be traversed during program execution. In order to create those links, we need to perform a program flow analysis to determine the different links or traces between the nodes. As previously discussed in subchapter 3, we define seven types of blocks in our hybrid model. Each of these types refers to a specific structure inside the '*flow*' of the program such as iterating and selecting a program trace. To determine which links need to be created, we have to interpret the statements inside the blocks with the exception of generic basic blocks and case blocks. Each specific statement will eventually result in another program trace. For example, a while-loop will only iterate over the code as long the condition is true. Whereas a do ... while-loop will always execute the code at least once regardless of the result of the condition.

Second, the generated automaton model is exported into a XML project file for UPPAAL. The challenge in creating this file is to comply to the correct syntax defined by the UPPAAL framework and also to deliver a proper and clear layout of the model in the GUI based tool [5] [6].

In figure 6, an automaton is generated in UPPAAL for the example code given in listing 1. Each node in the model has a name corresponding to the line numbers in the original source file of the code included in that node. The associated code is included in the comments of each node. For a Case Block, i.e. true/false or cases in a switch instruction, the statement is added to the guard of the link that refers to the matching program trace.

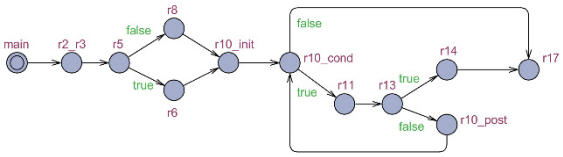


Fig. 6 Timed Automaton in UPPAAL generated from the sample code in listing 1.

4 Results

During development of the COBRA-HPA framework, we are using sample code files to test the functionality of the framework. These reference files are used from the TACLeBench. The TACLeBench is a benchmark project of the TACLe community to evaluate timing analysis techniques and comparing their performances [3].

Apart from testing the COBRA-HPA framework creating blocks, we need to verify what the performances are of our hybrid methodology. For these tests, we have selected two testbenches from the TACLeBench to perform timing measurements on their generated hybrid model, i.e. qsort and h264dec_ldecode_macroblock. The timing measurements are executed on an ATMEL AT90CAN128 and a single core of the ARM Cortex M7 processor. The M7 processor architecture allows us to enable or disable the built-in cache memory on the fly. Each testbench is tested on three different abstraction depths of the hybrid model. The resulting blocks are then executed on the ATMEL processor and twice on the ARM processor, once with cache enabled and again with cache disabled.

Figure 7 is a sample from the numerous measurement results of the performed tests on the function *quicksort_init*. When we look at the right side of figure 7, we first notice the node *r68_init* with a execution time of less than one clock tick on the ARM. This result is due to the size of the block that is to small for the measurement. The instructions inside the block, in this case a variable declaration, are executed too fast to obtain an accurate result with the current time resolution.

The second node in the row contains an iteration statement. This is indicated by the red arrow. The number next to this arrow specifies the range in which the loop will be repeated. The ranges for all iteration statements in the TACLeBench testbenches are annotated inside the source code. These annotations specify the minimum and maximum loop boundaries. In our case, we are interested in the maximum number of iterations to determine the WCET.

Since we are interested in the WCET of a task, we assume that the total WCET of a block equals to the sum of the WCETs of its child blocks. As we apply this theorem to the results in figure 7, we find a total WCET of 670/8347 clock ticks for the individual blocks. If we measure the WCET of total block at once however, we obtain a WCET of 394/4695 clock ticks. This is a difference of almost 56-59%.

The major difference in WCETs between the abstract block and its composed form is caused by several factors. The foremost reason, that has the largest impact on the results, is the size of the blocks. Each block contains a part of code from a larger whole. These parts follow each other in a sequential order and can share data with each other using variables. For example, the first block calculates the result of an algorithm. This result is than passed to the second block to perform a comparison

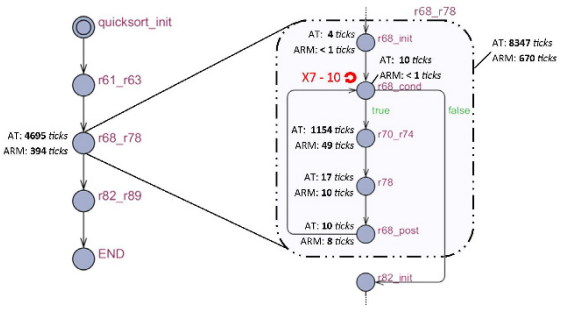


Fig. 7 Measurement results of the quicksort testbench with disabled cache in clock cycles. *Left*: abstraction depth = 1, *Right*: abstraction depth = 2 (AT: ATMEL AT90CAN128, ARM: ARM Cortex-M7)

to decide which trace to continue. When we evaluate each block separately to determine the WCET, we have to generate a set of all possible inputs in order to find the worst case possible. However, this set of inputs can be significantly larger than the actual input set during normal program execution. Therefore, the obtained WCET will be larger than the real WCET and thus raising the upper bound of the execution time distribution as shown in figure 1.

A second reason, that lies at the root of the time difference, are the built-in performance enhancement techniques on the processor which improve the overall execution speed. For instance, the ARM Cortex-M7 processor used in this research has a six-stage pipeline. The pipelining of instructions results in faster program execution by performing multiple instructions in parallel, e.g. resolving instruction *A* while fetching instruction *B*. As we look back to figure 7, we see that most blocks on the right side consist of few instructions. The abstracted 'bigger' block, on the other hand, can execute all instructions consecutively and thus benefit from the performance boost.

In conclusion, the two main factors which have a major impact on the WCET analysis are the size of the blocks and the non-deterministic/complex features of the Cortex-M7 processor architecture. When performing a WCET analysis with the hybrid methodology, we must keep in mind that our blocks are not too small to maintain accurate values. In the conducted test setup, we have executed small pieces of code on a very powerful processor with lots of enhancement features, which influenced the results even more. A better approach for our test setup would be using a smaller processor (e.g. ARM Cortex-M0) with all 'extra' improvement features disabled or creating larger tasks with larger blocks for the Cortex-M7.

5 Conclusion

In this paper, we suggested a new hybrid approach to determine the WCET by combining two existing methodologies. This technique creates an opportunity to merge the benefits of algorithm based thinking with a measurement based approach.

In order to verify the new mindset, we created the COBRA-HPA framework which allows us to split code into blocks according to the hybrid methodology. The resulting prototype is successfully capable of parsing C-source files and generating

a corresponding hybrid block scheme. Additionally, an extension on the framework makes it possible to translate a existing block model into a timed automaton for the model checker tool UPPAAL. With the generated source files of the COBRA-HPA framework, we could easily run time measurements for hybrid blocks.

Finally, we performed measurements with the generated files on an AT90CAN128 and an ARM Cortex-M7 processor. We can conclude from these results that the size of the blocks have a tremendous impact on the obtained WCET.

In future work, we can repeat the measurements creating larger tasks with bigger blocks to examine the influence of the size on the results of a block and its composed form. In addition, improvements can be applied to the COBRA-HPA to increase the user experience and lower the manual labour.

References

1. Axer, P., et al.: Building timing predictable embedded systems. ACM Transactions on Embedded Computing Systems **13**(4), 1–37 (2014)
2. Behrmann, G., David, A., Larsen, K.G.: Formal Methods for the Design of Real-Time Systems, *Lecture Notes in Computer Science*, vol. 3185, chap. A Tutorial on UPPAAL, pp. 200–236. Springer Berlin Heidelberg (2004)
3. Falk, H., et al.: Taclebench: a benchmark collection to support worst-case execution time research. Proceedings of the 16th International Workshop on Worst-Case Execution Time Analysis (WCET'16) (2016)
4. Hahn, S., Reineke, J., Whilhem, R.: Towards compositionality in execution time analysis - definition and challenges. In: The 6th International Workshop on Compositional Theory and Technology for Real-Time Embedded Systems (CRTS 2013) (2013)
5. Koolmees, B.: Validation of modeled behavior using uppaal. Final bachelor project, University of Technology Eindhoven (2011)
6. Larsen, K.G., et al.: Uppaal. http://www.uppaal.org/ (2015). URL http://www.uppaal.org/
7. Liang, Y., et al.: Timing analysis of concurrent programs running on shared cache multi-cores. Real-Time Systems **48**(6), 638–680 (2012)
8. Lv, M., et al.: A survey on cache analysis for real-time systems. ACM Computing Surveys p. 45 (2015)
9. Parr, T.: The Definitive ANTLR 4 Reference. The Pragmatic Bookshelf (2013)
10. Reineke, J.: Caches in wcet analysis. Ph.D. thesis, University of Saarlandes (2008)
11. Schoeberl, M., et al.: T-crest: Time-predictable multi-core architecture for embedded systems. Journal of System Architecture (2015)
12. Tian, T.: Software techniques for shared-cache multi-core systems (2012). URL https://software.intel.com/en-us/articles/software-techniques-for-shared-cache-multi-core-systems
13. Ungerer, T., et al.: parmerasa - multi-core execution of parallelised hard real-time applications supporting analysability. In: Euromicro Conference on Digital System Design, vol. 16 (2013)
14. Yan, J., Zhang, W.: Wcet analysis for multi-core processors with shared l2 instruction caches. In: Real-Time and Embedded Technology and Applications Symposium. IEEE, IEEE (2008)

Towards Recognising Activities of Daily Living through Indoor Localisation Techniques

Andriy Zubaliy, Anna Hristoskova, Nicolás González-Deleito, Elena Tsiporkova

Abstract Recognition of Activities of Daily Living is a problem that is still not yet solved. In addition, it is important to realise that a neat solution can support health-care systems to meet increasing population and ageing. Near existing fall detection solutions, a monitoring system of person's daily activities will provide a lot of opportunities. For instance, doctors and caretakers will spend more time for patients who truly need them and the patient, in its turn, will receive recommendations in order to prevent disease development and progression. This paper presents a novel user-centric approach proposed and researched by the first author in the context of his Master thesis. The initial idea was to push off from the indoor localisation technique as it might improve recognition of person's activities significantly. The experiments using Bluetooth beacons and a smartwatch demonstrated good results and a lot of potential.

1 Introduction

In this paper, we propose a novel approach and present the results of our research of technologies using which it is possible to recognise daily activities from *waking up* up to *going to sleep* including *washing hands, tooth brushing, taking a shower or a*

Andriy Zubaliy
University of Antwerp, Belgium, e-mail: zubaliy@gmail.com

Anna Hristoskova
University of Antwerp, Belgium e-mail: anna.hristoskova@uantwerpen.be

Nicolás González-Deleito
Data Innovation Group, Sirris, Belgium e-mail: nicolas.gonzalez@sirris.be

Elena Tsiporkova
Data Innovation Group, Sirris, Belgium e-mail: elena.tsiporkova@sirris.be

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_74

bath, preparing breakfast, taking breakfast, watching TV, working at the computer and other simple activities as well, such as sitting, walking, jumping etc. To the best of our knowledge, this is the first work that uses the presented technologies in order to solve this problem.

The rest of the paper is organised as follows: Section II briefly presents related work and Section III defines the term activity in our context. In Section IV we propose the configuration of a monitoring system and explore the selected technologies. We continue in Section V with analysis of the first measurements. Section VI describes our research methodology in order to recognise activities. We conclude with the challenges we encountered and the future work in Section VII.

2 Related Work

Due to the evolution in low power technologies there have been many researches done in the past years using wearable sensors [1, 2, 3]. However, none of them tried to build a comfortable and non-obtrusive system for daily usage. Their systems required to install many sensors in the environment and the user had to wear several custom-made devices. Their approach can only be applied in single-user environment. In order to recognize activities from multiple users, a user-centric approach is preferable [2]. Currently, Android smartphones are increasingly researched due to the built-in sensors and easy access to the development tools [4, 5, 6]. Since the arrival of modern smartwatches in 2013, there have been only a few studies done whether it is possible to use them for accurate activity recognition such as sitting, walking, running etc. [7, 8, 9].

3 Activity Recognition

In general, an activity (e.g. tooth brushing) in our context can be defined as a combination of:

- a location or whether the person was in the vicinity of an object during a certain amount of time (e.g. in the bathroom near the sink for 3 minutes).
- performed actions (e.g. turning the light on, opening the tap, opening the door)
- a set of used objects (e.g. light switch, toothbrush)
- specific movements (e.g. arm movement while brushing teeth)

Furthermore, we can assume the person performed tooth brushing while s/he was near the sink for a certain amount of time after s/he woke up. The previous activity determines the following and vice versa, taking into account other parameters as well.

More specifically, an activity can be simple (or atomic, e.g. sitting, walking, running, opening the door, etc.) or complex that consists of many sub-activities (e.g. cooking pasta) [1]. Additionally, the activities can be sequential (SA), interleaved (IA) or concurrent (CA) [1, 2]. This leads to a complexity in distinguishing them. Almost all researches in this field mainly consider sequential activities, because of the simplicity [1].

4 System Configuration Proposal

Based on our previous research [10] on state-of-the-art of localisation technologies and on current research on ubiquitous devices (the results are not shown to conserve space), we concluded that Bluetooth is the best suitable technology for our purposes. This technology can solve two problems: it can serve for determining location of the person using beacons (described in Section 4.1) and determining which object is used by this person. The concept of the monitoring system we propose, is shown on Figure 1. In this paper, we do not focus on the doctor's side, but concentrate more on the patient side.

In order to reach as much as possible target groups, already adopted devices by the public should be used, such as wristbands, smartwatches and smartphones. The on-board sensors such as skin temperature, galvanic skin response, motion sensors, heart rate, UV light and other allow us to measure user's health parameters and to track a part of their daily activities. These devices allow us to interact with the user through the screen, coloured LED lights, vibration and sound. Smartwatches and wristbands are made as an extension of the smartphone and should be coupled with a smartphone in order to transfer collected data for further analysis and to show results. Smartphones, in turn, can serve as an intermediate device to communicate data further to the cloud. Some smartwatches, on the other side, can also work autonomously, as they are equipped with Wi-Fi or even with cellular support and thus can replace both wristbands and smartphones.

Figure 1: Concept of the Monitoring System

For our research, we chose the combination of an Android smartwatch and Bluetooth beacons.

4.1 Analysis of Selected Technologies

Bluetooth Low Energy (BLE) was introduced in 2010. It provides a possibility to advertise packets without establishing a connection as was the case in the classic version of Bluetooth. Based on this technology, Apple introduced iBeacons in 2013, followed by Google with their Eddystone protocol in 2015.

A beacon is a small BLE device that is able to advertise packets. These packets are meant to be collected by other devices (e.g. smartphones) without user interaction. Such packets contain different information, such as the id of the beacon, temperature, light intensity and other data depending on the available sensors. Once a packet is received, an action can be triggered automatically (e.g. an advertisement is shown on the user's screen).

4.1.1 Estimote Stickers

We choose Estimote Stickers because of their small size, robustness, design, availability of accelerometer and temperature sensors, and plug-and-play capability. A Sticker has a maximum broadcasting range of about 12 meters, 8 levels to adjust broadcasting power and can last up to 1 year on a single coin battery. In addition, when a Sticker is in motion, it starts to broadcast twice as frequently in order to track a moving object.

One disadvantage is proximity-based nature, the signal fluctuates a lot and therefore the stickers are not supposed to estimate distance by design. The other disadvantage is the proprietary protocol, called Nearable, and thus we have to use Estimote SDK. Currently, it is only possible to configure Stickers using an Estimote application on an iOS device.

The radiation pattern of the Sticker is depicted on Figure 2. This is important to know in order to have a better understanding of how and where to attach beacons.

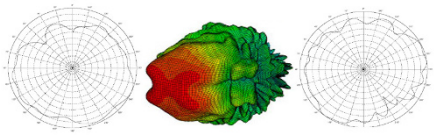


Figure 2: Radiation pattern of the Stickers [12]

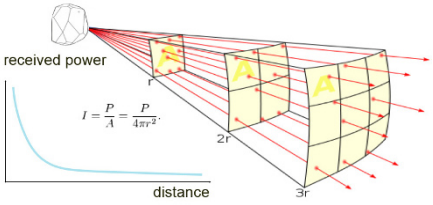


Figure 3: RSSI [12]

The broadcasted signal attenuates by distance (Figure 3) and is very susceptible to the environment. Any object between a beacon and a receiver results in a fluctuation of the signal due to reflection and absorption. When a packet from a beacon is received, the Received Signal Strength Indicator (RSSI) value is measured.

4.1.2 Android Smartwatch

The main advantage of a smartwatch is that the user can wear it almost continuously, except the time when the device needs to be charged. We decided to do our research on an Android based smartwatch above other platforms because it uses the same Android Stack as in Android smartphones. These devices are widely spread and have a huge development community because the platform provides an easy access

to the development tools. The other known platforms such as Tizen from Samsung, WatchOS from Apple and Microsoft Band do not yet offer these advantages.

After the analysis of modern wearable devices, we chose the Urbane W150 waterproof smartwatch from LG above all other devices because of direct availability on the market and because of various built-in technologies such as Bluetooth v4.1, WiFi b/g/n, haptics engine, accelerometer, gyroscope, compass, heart rate sensor and a battery of 410mAh. The other advantage against other smartwatches is the sensor batch queue [13] for 10000 events. This will allow us to do battery usage optimisations in the future.

4.1.3 Scan interval

The beacons advertise packets with fixed interval. The scan interval on Android is more a hint for the system. The chip decides when to return scanning results to the OS. Smart chips will return results less often if the results do not change too much. They can also return results immediately, rather than wait for the scan to end if they discover a new beacon. Therefore, it is both much more energy and CPU efficient, and can actually be more responsive than fully software defined scanning.

4.2 Sensors in the smartwatch

The most valuable physical sensors that can be used in our context are: accelerometer and heart rate sensor. The composite sensors, such as significant motion, step detector and step counter do not show good accuracy in indoor environment due to their low power nature, i.e. high latency.

5 First Measurements

In this section, we explain how data is collected. For each beacon following parameters are set: room, object and a list of user's activities associated with this object. E.g. Watching TV, Reading a Book and Relaxing can be added to the Living room-TV Sticker. This list of user's activities can be used later to label recognised patterns of activities defined in recognition algorithms.

Currently, the following data from the beacons is logged for further processing and analysis: timestamp, beaconId, rssi, temperature, moving, orientation, accelerometer data (x, y, z). From the smartwatch: battery level once started monitoring and once finished, timestamps when a step was detected and how many steps where done since the start of the monitoring, raw data from accelerometer.

Together with radiation pattern of a beacon shown on Figure 2 it is also important to know how different settings of broadcasting power affects the range. We noted the distance to the Sticker when RSSI was about -100 dB for each broadcasting level. The results of this test are shown on Table 1.

5.1 One beacon for one activity

Each room has one or more specific functions and thus is meant to do one or more activities, i.e. a room is linked to certain activities. It is also logical to attach beacons directly to the objects that "describe" an activity, e.g. on a desktop or a monitor in the office to describe the user is working. It is better to place beacons as close as possible to the user in order to have a better accuracy in activity recognition. Based

Table 1: Sticker's maximum broadcasting distance

Setting in dBm	Distance in m
-30	0.2
-20	1.3
-16	2.7
-12	3.8
-8	5
-4	7
0	9
+4	12

Table 2: One beacon for one activity

Activity	Possible locations of the beacons	RSSI	Moving temp.
Sleeping	On the wall above the bed On the nightstand	x	
Working at the computer	On the desktop On the monitor	x	(x)
Watching TV	On top of the TV, place that gets warm when TV is on	x	x
Washing hands Tooth brushing	On the wall above the sink On the sink	x	
Taking a shower or a bath	On the wall near a shower or a bath	x	x
Cleaning up	On the vacuum cleaner	x	x x

on the RSSI it is possible to distinguish whether the person is near or far, i.e. s/he is working at the computer or just walking in the room. Consequently, by *used object*, we mean an object near which a person spent a certain amount of time. In addition, an object is also *used* when the temperature has risen or a movement is detected.

Another option is to shorten the range of the beacons by reducing the broadcasting power in order to monitor user's presence near a specific object instead of the presence in the whole room. This way we can determine only one specific activity or a usage of the object and we should not consider the significance of RSSI. As additional benefit, the battery consumption of the beacon will be reduced as well.

In Table 2, we show a few examples of how to place a beacon on an object and which data can be used in order to recognise an activity with only one beacon.

5.2 Kitchen

On Figure 4, a *cooking activity* is shown. In comparison to the activities from Table 2, more beacons are needed in order to detect activities such as: cooking, heating food, just consuming food or water.

We defined following objects: (1) A sticker on the door edge of a *fridge*. The fridge is used when the door opening is detected. (2) A sticker on the door of a *microwave*. If movement is detected, then it means the microwave was used to heat up the food. If temperature rising is detected after a while, then it means the person used the grill option. (3) A sticker on the door of an *oven*. If movement is detected, then it means that the person wanted to use the oven. If temperature begins to rise, then it means the oven is being used and the user is cooking. (4) A sticker on a *cabinet's* or a *shelf's* door. If cabinet door is opened, then it means the person took some tableware. (5) On the side of the *kettle* or on the lid. If kettle is moved, we can suppose it was used. If temperature raising was detected, then we can be sure the person boiled water e.g. to prepare some tea. (6) In order to detect the *hob* usage, the sticker should be placed above on the wall or on the edge of a furniture. The person

is cooking when temperature is raising. (7) A sticker on the side of a *ater pitcher or a bottle*. The person consumed water when movement is detected.

In general, activities follow each other. If the user spent some amount of time in the kitchen, used some objects and then moved and spent some amount of time at the dining table, then we can be sure that at first the user performed a *cooking* or *preparing a meal* activity and then was *eating*.

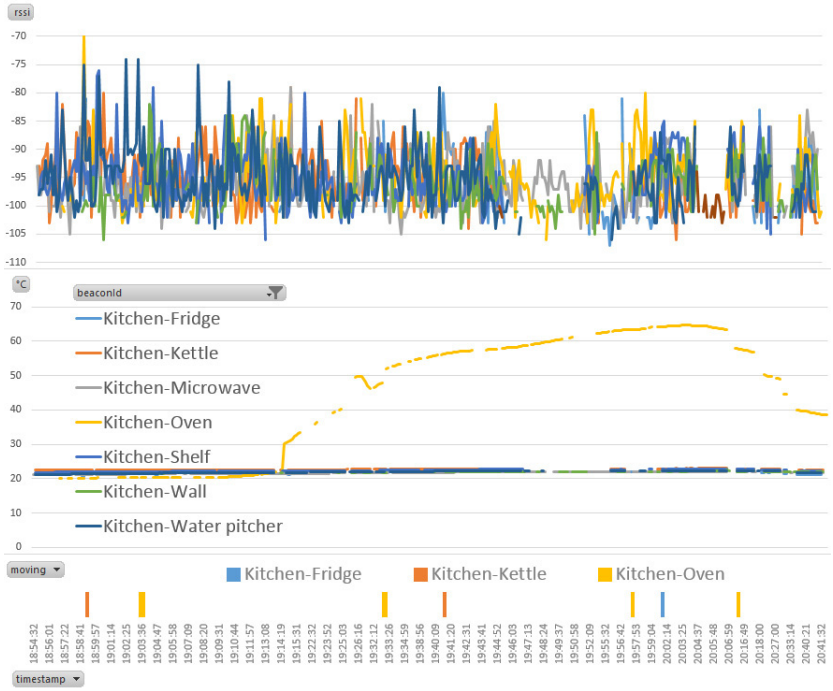


Figure 4: Cooking activity: RSSI

6 Research Methodology

In this section, we describe which steps we made in our research in order to recognize activities in a home environment.

The person can move relatively fast from one place to another and for instance opening the fridge happens fast. Therefore, appropriate broadcasting and scan settings are important. We set the broadcasting interval of each beacon to a bit less than 1000ms. The scan interval on the smartwatch is set to 1000ms. Because the scan interval is higher than the broadcasting interval, we should never loose packets.

In the following tests we attached beacons as described in Table 2. The test person went from one room to another, spent some time near each beacon or walked through the whole room to collect as much as possible representative data.

6.1 Measurement 1

8 beacons are set to the highest broadcasting power level of +4dBm in order to see what can in general be derived from the measurements. The results are shown on Figure 5. On the bottom of the figure the manually annotated corresponding room-level locations are shown. Note that there are no beacons placed in the bathroom.

In order to write only a room level based localisation system some complex rules and filters have to be written, but they will only work for one specific environment. If the environment changes, the algorithms should be adjusted as well.

One of the most important conclusions we can derive from this test is that high broadcasting power should doubtless be set to the beacons to continuously monitor objects which temperature is a determinative parameter in order to prevent accidents, e.g. the oven in the kitchen.

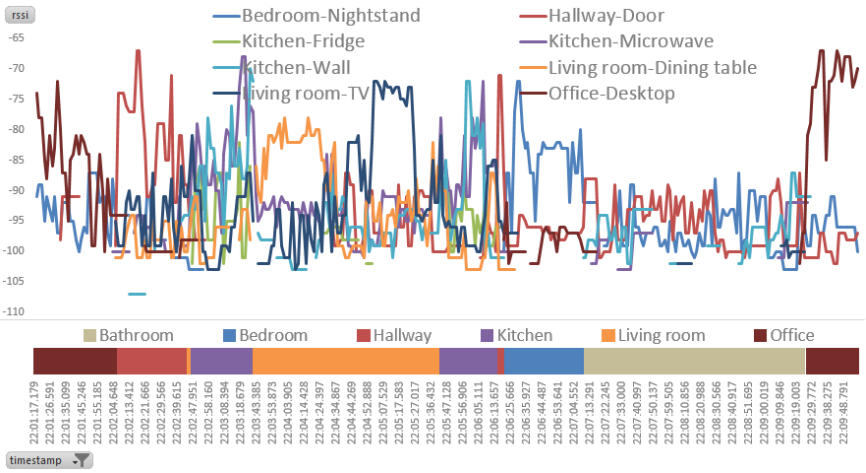


Figure 5: +4dBm, Walked through all rooms

6.2 Measurement 2

In this test we could use only 6 beacons with the same broadcasting power of -12dBm. One reason was the manufacturing fault that resulted in a damage of some beacons. The other reason was an update of the Estimote application that contained a bug that did not allow us to change settings of the Stickers for a long period.

The results are shown on Figure 6. We clearly see separated groups of signals. On the bottom of the figure the manually annotated corresponding room-level locations are shown. As a result of a lower broadcasting power, there are places where packets do not reach the smartwatch. Therefore, an *EMPTY* state (i.e. empty log) is introduced in order to know whether the smartwatch works properly and keeps scanning the environment. *EMPTY* log is translated into an *Unknown* location.

In comparison to the Measurement 1, it is now possible to write simple algorithms in order to retrieve meaningful information from the data such as: the amount of time the person spent near each beacon. However, the algorithms to determine the nearest beacon can become very complex when a beacon is in the intersection of several beacons. Fortunately, there exist algorithms, such as classification that are able to solve such problems.

6.3 Neural Network as Classifier

Neural networks is one of the areas of research in the field of artificial intelligence based on trying to reproduce the ability of the brain to learn and correct mistakes. To do our analysis, we used an open source, web-based framework called Jupyter

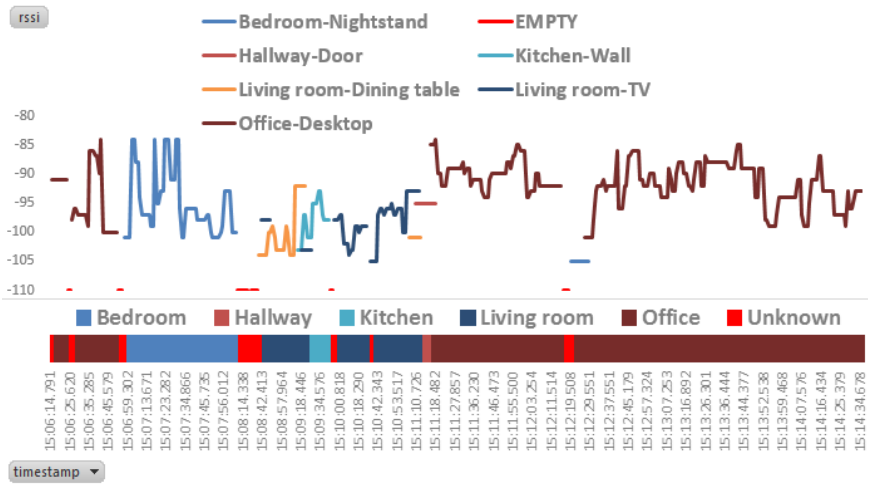


Figure 6: -12dBm, Walked through all rooms

Notebook, loved by many researchers. We used Multilayer Perceptron Classifier (MLPC) from scikit-learn machine learning library written in Python [14] in order to determine the nearest beacon. At the time of writing this paper, the MLPC existed only in the development version *(0.18.dev0)* of scikit-learn.

We trained the MPLC to recognize the nearest beacon on the RSSI values only. 4 test files were used, containing 8, 10, 18 and 30 minutes of data. The first file was used as training set. The data was collected from 8 beacons with broadcasting power set to -12dBm. We achieved a very high recognition accuracy of 99.29%.

In order to determine the nearest beacon or a room-level localisation of a person, definitely the classifications algorithms should be used. Finally, the combination of the temperature rising, door movements, other user's actions and location should give us appropriate results in determining the activity.

7 Challenges and Future Work

Working with very new technologies expects a lot of patience as they might be buggy as it was in our case.

It is already possible to recognise many of the main daily activities using only 10 beacons, however it would be interesting to try to make as much as possible objects smart in order to investigate what is the limit of this approach. The effectiveness of a humidity sensor in the kitchen should be investigated. In addition, the Nearable protocol used in Stickers can be reverse-engineered and Android native BLE API for beacon discovery should be used in order to eliminate smoothing algorithms used in Estimote SDK that hold already invisible beacon for a few scan periods more.

It should be investigated how locomotions such as walking, stand up, sit down and other could help to recognise activities. However, this can lead to a more complex system.

When working with data collection, it is important to have an efficient procedure to label your data. Therefore, an application for the watch be developed that will simplify data annotation during measurements. In addition, more test data should be

collected in order to test all kind of conditions; other classification methods, such as Decision Trees, SVM, kNN and others should be tested and compared; and instead of writing own code, the existing frameworks with GUI should be used, e.g. Weka. Own recognition algorithm can provide good results in recognition of activities' patterns, but this should also be compared with effectiveness of the classifiers.

References

1. A. Benmansour, A. Bouchachia, and M. Feham, "Human activity recognition in pervasive single resident smart homes: State of art," in *Programming and Systems (ISPS), 2015 12th International Symposium on*, April 2015, pp. 1–9.
2. T. Gu, Z. Wu, X. Tao, H. Pung, and J. Lu, "An Emerging Patterns based approach to sequential, interleaved and Concurrent Activity Recognition," in *Pervasive Computing and Communications, 2009. PerCom 2009. IEEE International Conference on*, March 2009, pp. 1–9.
3. O. Lara and M. Labrador, "A Survey on Human Activity Recognition using Wearable Sensors," *Communications Surveys Tutorials, IEEE*, vol. 15, no. 3, pp. 1192–1209, Third 2013.
4. M. Mitchell, F. Sposaro, A.-I. Wang, and G. Tyson, "BEAT: Bio-Environmental Android Tracking," in *Radio and Wireless Symposium (RWS), 2011 IEEE*, Jan 2011, pp. 402–405.
5. Y.-S. Lee and S.-B. Cho, "Activity recognition with android phone using mixture-of-experts co-trained with labeled and unlabeled data," *Neurocomputing*, vol. 126, pp. 106 – 115, 2014, recent trends in Intelligent Data AnalysisOnline Data ProcessingSelected papers of the The 6th International Conference on Hybrid Artificial Intelligence Systems (HAIS 2011)Including a selection of papers from the International Conference on Adaptive and Intelligent Systems 2011 (ICAIS 2011). [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231213006966
6. D. Lou, X. Chen, Z. Zhao, Y. Xuan, Z. Xu, H. Jin, X. Guo, and Z. Fang, "A Wireless Health Monitoring System based on Android Operating System," {*IERI*} *Procedia*, vol. 4, pp. 208 – 215, 2013, 2013 International Conference on Electronic Engineering and Computer Science (EECS 2013). [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2212667813000336
7. C. Boletsis, S. McCallum, and B. Landmark, "The Use of Smartwatches for Health Monitoring in Home-Based Dementia Care," in *Human Aspects of IT for the Aged Population. Design for Everyday Life*, ser. Lecture Notes in Computer Science, J. Zhou and G. Salvendy, Eds. Springer International Publishing, 2015, vol. 9194, pp. 15–26. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-20913-5_2
8. J. Trimpop, M. Haescher, G. Bieber, D. J. Matthies, F. Lämmel, and P. Burggraf, "The Digital Health Companion: Personalized Health Support on Smartwatches via Recognition of Activity-and Vital-Data," 2015.
9. G. Bieber, M. Haescher, and M. Vahl, "Sensor Requirements for Activity Recognition on Smart Watches," in *Proceedings of the 6th International Conference on PErvasive Technologies Related to Assistive Environments*, ser. PETRA '13. New York, NY, USA: ACM, 2013, pp. 67:1–67:6. [Online]. Available: http://doi.acm.org/10.1145/2504335.2504407
10. A. Zubaliy, A. Hristoskova, N. González-Deleito, and E. Tsiporkova, "A survey on indoor localisation and related sensor technologies for enabling quality of life applications," goo.gl/SdzDMI, (unpublished manuscript).
11. "Estimote beacon stickers track movement and temperature," http://thenextweb.com/gadgets/2014/08/21/estimotes-beacon-stickers-can-track-movementtemperature-communicate-youre-nearby/, (Accessed on 05/15/2016).
12. "Estimote // hardwired nyc // september 2014," http://www.slideshare.net/firstmarkcap/estimote-hardwired-nyc-september-2014, (Accessed on 05/13/2016).
13. "Batching — android open source project," https://source.android.com/devices/sensors/batching.html, (Accessed on 05/13/2016).
14. "sklearn.neural_network.mlpclassifier scikit-learn 0.18.dev0 documentation," http://scikit-learn.org/dev/modules/generated/sklearn.neural_network.MLPClassifier.html, (Accessed on 05/15/2016).

Powerwindow: a Multi-component TACLeBench Benchmark for Timing Analysis

Haoxuan Li, Paul De Meulenaere, Peter Hellinckx

Abstract Timing analysis is used to extract the timing properties of a system. Different timing analysis techniques and tools have been developed over the past decades. Well-rounded benchmark suites are in a high demand for the evaluation of those techniques and tools. TACLeBench is aiming to provide a benchmark suite to fulfil the requirement. It is not only limited to single-core systems but also will be replenished with complex benchmarks for multi-core systems.

In this paper, we propose a multi-component benchmark which can be easily converted to benchmarks for single-core, parallel, distributed or even multi-core systems without fundamental changes.

1 Introduction

Real-time embedded systems are widely adopted in applications such as automotive, avionics and medical care. Many of these systems are required to meet strict deadlines. Understanding the timing behaviour of such systems at the early design stage can reduce the number of design iterations, which ultimately reduces the design cost and time-to-market. Timing analysis is the process of deriving or estimating the timing properties of a system.

Different timing analysis techniques and tools have been developed over the past decades [7]. Well-rounded benchmark suites are in a high demand for the evaluation of those techniques and tools. The current existing benchmark suites for code level

Haoxuan Li
CoSys-Lab, MOSAIC, University of Antwerp e-mail: Haoxuan.Li@uantwerpen.be

Paul De Meulenaere
CoSys-Lab, Flanders Make, University of Antwerp e-mail: Paul.Demeulenaere@uantwerpen.be

Peter Hellinckx
MOSAIC, iMinds, University of Antwerp e-mail: Peter.Hellinckx@uantwerpen.be

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_75

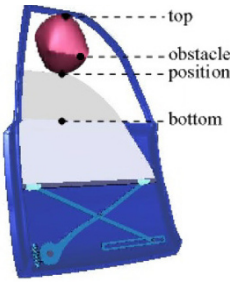


Fig. 1 Power window [4].

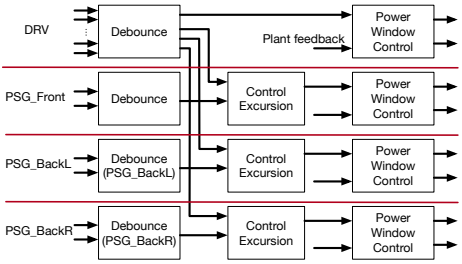


Fig. 2 The complete scope of the power window Simulink® model.

timing analysis, however, do not cover multi-core softwares. TACLeBench [1] is aiming to provide an available and comprehensive benchmark suite to meet the requirement. It is not only limited to single-core systems but also will be replenished with complex benchmarks for multi-core systems in the near future.

Powerwindow is a multi-component benchmark included in the TACLeBench. It contains four main components including one driver-side window and three passenger-side windows. These components allow the benchmark to be easily converted to benchmarks for single-core, parallel, distributed or even multi-core systems without fundamental changes. The goal of this paper is to prove that the powerwindow benchmark is suitable to be used by different timing analysis techniques according to the requirements. In this paper, we performed measurement-based timing analysis with three case studies covering the benchmark in isolation, the benchmark with interactions and the benchmark with interactions in semi-parallel.

The rest of this paper is structured as follows. Section 2 gives a detailed description about the powerwindow benchmark. Section 3 provides an overview of the case studies. The results are elaborated and compared in Section 4 followed by discussion in Section 5. The conclusion and the future work are covered in Section 6 and Section 7 respectively.

2 Powerwindow Benchmark

2.1 Power Window Description

A power window is an automatic window which can be raised and lowered by pressing a button or a switch instead of a hand-turned crank handle (Fig. 1). The complete application of a power window in a car normally consists of four windows which can be operated individually. The three passenger-side windows can be operated by the driver with higher priority. In addition to the basic function that controls the position of the window with push-buttons, the power window model we used in this paper contains the following functions [6]:

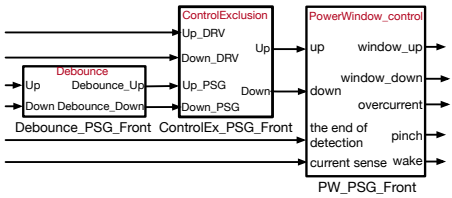


Fig. 3 The front-passenger-side power window Simulink® mode.

Autoup & **Autodown:** The autoup and autodown functions can be triggered by pressing the push-down button then quickly release it within 0.5 second. Once initiated, the window will raise up or lower down automatically until it reaches the end of the window frame, or until an object is detected between the rising glass and the window frame.

Object (Pinch Force) Detection: If an obstacle is present between the top of the window frame and the glass when the window is closing (Fig. 1), the motor will automatically revert. The window glass will lower down by a certain distance to prevent potential injuries to hands, fingers or even accidents to children.

End of the Range Detection: The end of the range detection mechanism guarantees that when the window is fully closed, no force higher than a certain value (e.g. 100 N) is present between the glass and the window frame.

2.2 Power Window Model

The complete power window model (Fig. 2) is elaborated in Simulink® and consists of the driver-side window (*DRV*), front-passenger-side window (*PSG_Front*), back-left-passenger-side window (*PSG_BackL*) and back-right-passenger-side (*PSG_BackR*). The outputs of the driver-side debounce module are used as inputs by the control exclusion modules of the passenger-side windows.

Since the three passenger-side power windows are identical, we will use the front-passenger-side window in this paper (Fig. 3). Compared with the passenger-side power window, the driver-side power window excludes the control exclusion module and contains six more debounce circuits in the debounce module.

Debounce: When a push-button is pressed, it rebounds a bit before settling. This mechanical reaction often generates spurious inputs, for example, it may be read as multiple quick presses in a short period by the program. The debounce module (*Debounce_PSG_Front*) is implemented to filter out these faulty inputs. The debouncing function is triggered on the rising edge of a pulse signal with a period of 10 ms. Any signal that is shorter than 30 ms will not be forwarded to the next module. The debounce module is trivial because it guarantees the input only changes when push-button is definitely pressed or released.

Control Exclusion: The control exclusion module (*ControlEx_PSG_Front*) is only included in the passenger-side windows. By using basic logic gates, the control exclusion module assigns higher priority to the driver when control inputs come from driver and passenger simultaneously.

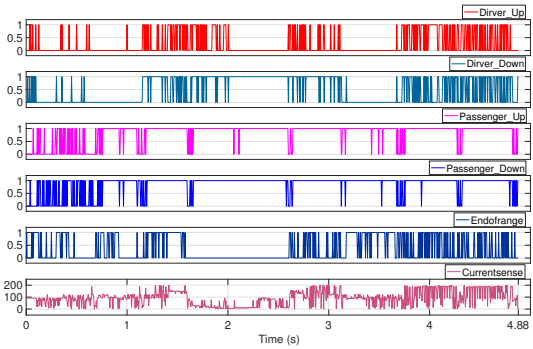


Fig. 4 The front-passenger-side power window inputs. The cases covered by the input set include window position control from the passenger, simultaneous controls from both the driver and the passenger, pinch detection, and end of the range detection.

Power Window Control: The power window control module (*PW_PSG_Front*) uses the control signals (*up*, *down*) and the environment signals (*the end of detection*, *current sense*) to control the motor and the safety mechanism. Similar to the debounce module, the power window control function is connected to a pulse signal with a period of 50 ms and triggered on both the rising and falling edges. The input signals *the end of detection* and *current sense* are the feedback from the plant, which is a part of the environment.

2.3 Benchmark Description

The benchmark was initially generated in C code by Matlab® Embedded Coder® targeting Atmel AVR (8-bit) micro processor with a discrete fixed-step solver at a fixed-step size of 5 ms from the Simulink® model. Then modifications were made to convert the generated code to TACLeBench standards [1]. Table 1 is a brief elaboration of the source files included in the powerwindow benchmark. The input set included in *powerwindow_inputs* can be replaced by user-defined inputs.

The input set was generated exclusively from the model. One of the reasons is to guarantee the power window model can be deployed independently from the inputs. This feature provides flexibility to the benchmark. For example, when a microprocessors has limited data memory, the input set can be used as external inputs from communication interface such as serial port. Additionally, it is almost effortless to connect the model with new input set or real-time inputs, in such way the users can create new benchmarks adjusted to the needs. By using real-time inputs, the benchmark accomplishes the functionality of a real-world power window application.

Fig. 4 illustrates the input set of the front-passenger-side window. The input set was created using Simulink® Design Verifier^{TM} with MC/DC (modified condition/decision coverage) coverage objectives. MC/DC is a code coverage criterion to assess whether adequate testings have been performed for the most critical (Level A) software [2]. It is used in avionics software development guidance DO-178B [3] and DO-178C [5] and highly recommended for automotive. The generated input set achieved the highest possible MC/DC coverage rate for the power window model. Thus, it may have a higher possibility to hit the WCET in the timing analysis.

The inputs were sampled every 5 ms, which means there are 977 groups of inputs in the input set. Note that this input set is used for testing the front-passenger-side window as an individual benchmark. When using the complete power window benchmark, the *Driver_Up* and *Driver_Down* will be connected to the driver-side power window.

Table 1 The name and function of the source files included in powerwindow benchmark.

Name	Function
debounce	the Debounce module
controlexclusion	the Control Exclusion module
powerwindow_control	the Power Window Control module
PW_Control_DRV	the driver-side power window
PW_Control_PSG_Front	the front-passenger-side power window
PW_Control_PSG_BackL	the back-left-passenger-side power window
PW_Control_PSG_BackR	the back-right-passenger-side power window
const_params	the lookup table for stateflow chart in Power Window Control module
powerwindow	the final implementation of the complete power window
inputs	the input set of the power window, can be replaced by user-defined inputs

[a] Note: apart from powerwindow, the complete name of all the other files include prefix *powerwindow_*

3 Case Study

In the case studies, the powerwindow benchmark was deployed on a DVK90CAN1 development board with an 8-bit Atmel AT90CAN128 single-core microprocessor. Technically speaking, every component of the benchmark is a periodic task. Every time a task arrives, it takes one group of inputs from the input set. Ideally, the benchmark should be connected to a 16-bit timer which triggers one interrupt every 5 ms. However, in our case, due to the limited data memory, the input set was sent through a serial port. This communication was not included in the main function, so that the measured execution time would not be contaminated. The input set included in *powerwindow_inputs* set was used as the test case. One complete execution of the test case is considered as one test. The n^{th} execution is represented by *test–n*. In total, the test case was executed 50 times consecutively in every case study.

In the case study we performed timing analysis on the benchmark in three cases:

1. case study 1: the driver-side and passenger-side windows execute in isolation as individual benchmarks.
2. case study 2: four windows execute on one single-core processor in sequential order with interactions.
3. case study 3: four windows execute in semi-parallel with interactions under an operating system.

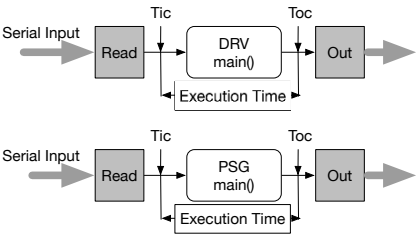


Fig. 5 The implementation of case study 1. The driver-side window and passenger-side windows are used as individual benchmarks.

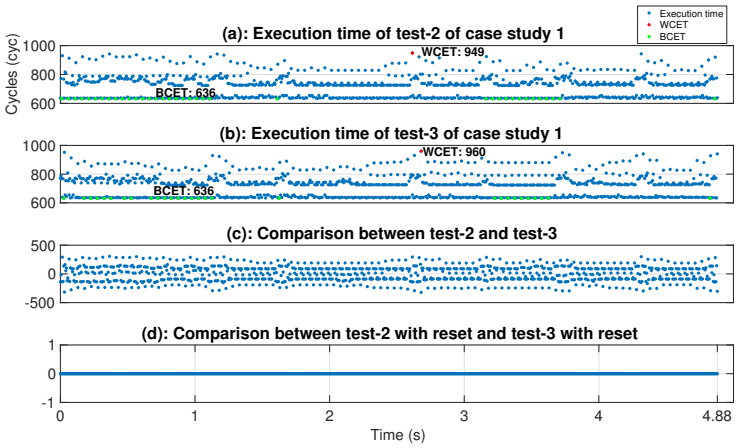


Fig. 6 The results (partially) of execution time of case study 1 of PSG_Front.

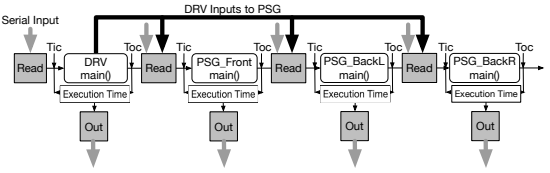


Fig. 7 The implementation of case study 2. Four power windows are chained in sequential order. The outputs of driver-side window are used as inputs for passenger-side windows.

4 Results

Case Study 1 In case study 1 (Fig. 5) the test case was executed consecutively without resetting the board or reinitializing the function. The timing analysis results are demonstrated in Table 2. The BCET (Best-Case Execution Time) of the driver-side window was reached at 1189 *cyc* (148.625 μs), the WCET (Worst-Case Execution Time) was at 1986 *cyc* (248.25 μs). The timing behaviours of the three passenger-side windows were identical. The BCET and the WCET were reached at 632 *cyc* (79 μs) and 977 *cyc* (122.125 μs) respectively.

As can be seen in Fig. 6:(a) and (b), the BCETs for both tests were at 632 *cyc*, whilst the WCETs were at 949 in *test*–2 and 960 in *test*–3. The BCET occurred much more frequently than the WCET. In Fig. 6:(c), although the case study was carried on a very simple single-core processor without any non-deterministic components, *test*–2 and *test*–3 showed different execution patterns. After reinitialization of the power window before the execution of the complete test case, the benchmark showed exactly same timing behaviours across the different tests (Fig. 6).

Table 2 The results of case 1.

Task	BCET (*cyc*)	BCET (μs)	WCET (*cyc*)	WCET (μs)
DRV	1189	148.625	1986	248.25
PSG_Front	632	79	977	122.125
PSG_BackL	632	79	977	122.125
PSG_BackR	632	79	977	122.125

[a] DRV: driver-side window; PSG: passenger-side window; BackL: back-left; BackR: back-right; BCET: best-case execution time; WCET: worst-case execution time; *cyc*: system clock cycle.

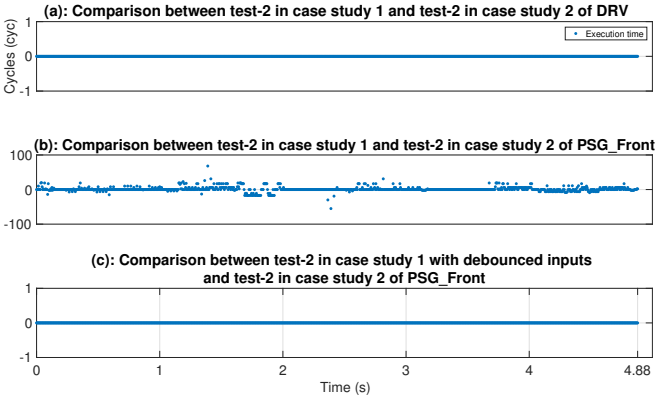


Fig. 8 The results (partially) of case study 2.

Case Study 2 In case study 2, four power windows were chained in sequential order to test whether the powerwindow benchmark is applicable when there are interactions between different components (Fig. 7). The results are presented in Table 3. The BCET and the WCET of the driver-side window had same values as in case study 1. The WCETs of the passenger-side windows were at 991 *cyc* (123.875 μs), which were slightly higher compared with the 977 *cyc* observed in case study 1.

Table 3 The results of case 2.

Task	BCET (*cyc*)	BCET (μs)	WCET (*cyc*)	WCET (μs)
DRV	1189	148.625	1986	248.25
PSG_Front	632	79	991	123.875
PSG_BackL	632	79	991	123.875
PSG_BackR	632	79	991	123.875

[a] DRV: driver-side window; PSG: passenger-side window; BackL: back-left; BackR: back-right; BCET: best-case execution time; WCET: worst-case execution time; *cyc*: system clock cycle.

As can be seen in Fig. 8:(a) *test*–2 in case study 2 and *test*–3 in case study 1 had identical timing behaviours at the driver-side window. However this observation did not hold for the passenger-side windows (Fig. 8:(b)).

In order to inspect the reasons behind the observed changes in passenger-side windows in case study 2, we created a new input set, in which the input signals from the driver were debounced before being used by the passenger-side power

Table 4 The results of additional experiment in case study 2.

Task	BCET (*cyc*)	BCET (μs)	WCET (*cyc*)	WCET (μs)
PSG_Front	632	79	991	123.875
PSG_BackL	632	79	991	123.875
PSG_BackR	632	79	991	123.875

[a] PSG: passenger-side window; BackL: back-left; BackR: back-right; BCET: best-case execution time; WCET: worst-case execution time; *cyc*: system clock cycle.

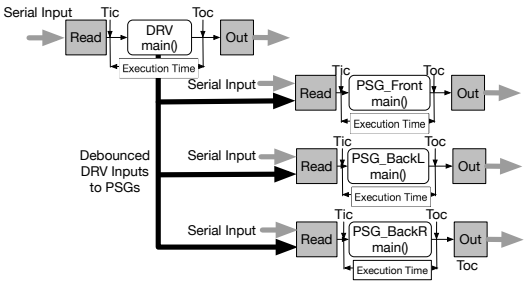


Fig. 9 The implementation of case study 3. The OSEK [8] operating system is used to schedule the four power windows. The tasks are defined as non-preemptive tasks with a period of 5 ms.

window in isolation as in case study 1. The results of this experiment were identical compared with the results of passenger-side windows in case study 2 (Fig. 8:(c)).

Case Study 3 Case study 3 targets the possiblity of using the powerwindow benchmark to perform timing analysis on parallel tasks under an operating system (Fig. 9). Due to the delay introduced by serial communication, the tasks will pile up in the task queue. Tasks with lower priority will be constantly pre-empted. Therefore, all tasks needed to have the same priority to guarantee the correctness of the inputs received from the serial port. This is the reason why case study 3 is "semi-parallel".

Table 5 The results of case 3.

Task	BCET (*cyc*)	BCET (μs)	WCET (*cyc*)	WCET (μs)
DRV	1246	155.75	3523	440.375
PSG_Front	743	92.875	2762	345.25
PSG_BackL	743	92.875	2754	344.25
PSG_BackR	743	92.875	2754	344.25

[a] PSG: passenger-side window; BackL: back-left; BackR: back-right; BCET: best-case execution time; WCET: worst-case execution time; *cyc*: system clock cycle.

As can be seen in Table 5, the BCETs increased by 111 *cyc* in case study 3. The WCET of driver-side window almost doubled compared with the other case studies. The increases of the WCETs of the passenger-side windows were even more notable. The results of the 3 passenger-side windows were no longer identical. It can be observed in Fig. 10 that, most of the execution times in case study 1 and 2 were around 650 *cyc* and 750 *cyc*. The WCETs occurred rarely but still close to the average execution time. However, in case study 3, the distribution changed drastically. Most of the execution times were lower than 1000 *cyc*. Only several execution times were located around 2500 *cyc*, which was much higher than the average execution time.

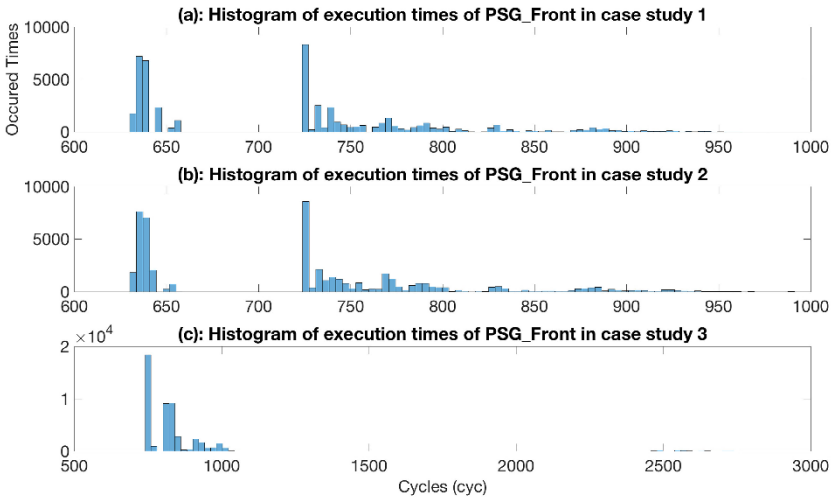


Fig. 10 The histogram of execution times (from *test*–1 to *test*–50) of PSG_Front in 3 case studies.

5 Discussion

Case study 1 has shown that the variations between different executions without reset were due to the fact that the benchmark is sensitive to the initial state. The execution time is not only dependent on the inputs but also on the state (the variable values) of the power window function. The changes in the execution pattern is because the final state of one test becomes the initial state of the following test. This can be further proved by the experiment with reinitialisation, in which the power window showed exactly the same behaviour across different tests.

In case study 2, after connecting the passenger-side windows with the driver-side window, different timing behaviours occurred in the passenger-side windows. The additional experiment indicated that the differences were due to the debounce function in the driver-side window, from which we concluded that the test case generated by Simulink® Design Verifier^{TM} was not sufficient to cover the worse case scenario. Special cautions should be given when using Simulink® to generate test case for critical systems. After further investigation we noticed the test case generated by Simulink® is based on the step-size of the solver (5 ms in our case). However, when the trigger of the functional component is not synchronized with the solver, the scenario covered by the test case may be missed. Due to this, even though the test case covers all possible inputs and correlated outputs under MC/DC coverage criterion, it is still not sufficient for the WCET detection.

The increase of the execution time in case study 3 may relate to the OSEK operating system. The increases in the WCETs were considerably higher compared with the increases in BCETs. To fully understand the reasons behind the dramatic increases, a further investigation towards the OSEK operating system is needed.

Although the test case was not complete for detecting the absolute WCET for the passenger-side window in case study 1, the powerwindow benchmark still is sufficient to evaluate timing analysis techniques with different requirements. Because of the flexibility provided by the benchmark, more complicated and new benchmarks could be created with limited efforts.

6 Conclusion

In this paper we presented a multi-component TacleBench benchmark. By adjusting the four components, the benchmark can be used for evaluating timing analysis techniques and tools for single-core, parallel, distributed or even multi-core systems. This flexibility provided by the components is the main advantage of this benchmark. In addition, the powerwindow benchmark showed the possibility to generate benchmarks using Matlab® Embedded Coder®, with which only minimal efforts are needed to create new benchmarks, especially for specific hardware platforms.

7 Future Work

In the next step, we will use input pins to replace the serial port. In such way we shall be able to perform case study 3 using preemptable tasks with different priorities. Timing analysis on processors with cache and multi-core will be studied thereafter. For the future timing analysis, the impact of the initial state will be considered. Anther topic which needs attention is how to improve the test case generated by Simulink® Design Verifier^{TM}.

Acknowledgement

This research was partially supported by Flanders Make vzw, the strategic research centre for the manufacturing industry.

References

1. Falk, H., et al.: Taclebench: A benchmark collection to support worst-case execution time research. In: 16th International Workshop on Worst-Case Execution Time Analysis (2016)
2. Hayhurst, K.J., et al.: A practical tutorial on modified condition/decision coverage (2001)
3. Johnson, L.A., et al.: Do-178b, software considerations in airborne systems and equipment certification. Crosstalk, October (1998)
4. Prabhu, S.M., et al.: Model-based design of a power window system: Modeling, simulation and validation. In: Proceedings of IMAC-XXII: A Conference on Structural Dynamics, Society for Experimental Mechanics, Inc., Dearborn, MI (2004)
5. Rierson, L., et al.: Developing safety-critical software: a practical guide for aviation software and DO-178C compliance. CRC Press (2013)
6. Vanherpen, K., et al.: Model transformations for round-trip engineering in control deployment co-design. In: Proceedings of the Symposium on Theory of Modeling & Simulation: DEVS Integrative M&S Symposium, pp. 55–62. Society for Computer Simulation International (2015)
7. Wilhelm, R., Altmeyer, S., Burguière, C., Grund, D., Herter, J., Reineke, J., Wachter, B., Wilhelm, S.: Static timing analysis for hard real-time systems. In: Verification, Model Checking, and Abstract Interpretation, pp. 3–22. Springer (2010)
8. Zahir, A., et al.: Osek/vdx-operating systems for automotive applications. In: OSEK/VDX Open Systems in Automotive Networks (Ref. No. 1998/523), IEE Seminar, pp. 4–1. IET (1998)

Predictive Algorithms for Scheduling Parameter Sweep Calculations in a Cloud Environment

Stig Bosmans, Glenn Maricaux and Filip Van der Schueren

Abstract This paper investigates various techniques for scheduling parameter sweep calculations cost efficiently in a hybrid cloud environment. The combination of both a private and public cloud environment integrates the advantages of being cost effective and having virtually unlimited scaling capabilities at the same time. To make an accurate estimate for the required resources, multiple prediction techniques are discussed. The estimation can be used to create an efficient scheduler which respects both deadline and cost. These findings have been implemented and tested in a Java based cloud framework which operates on Amazon EC2 and OpenNebula.

1 Introduction

The increasing demand for processing power over the past years forced technology to evolve from supercomputers to other solutions. Grid systems have been a viable solution for a long time, however their resources aren't unlimited. For additional computing power, techniques were developed to address idle resources on desktop computers. [5] [11] More recently, cloud computing is changing information technology. Various companies offer public cloud services, providing virtually unlimited resources on demand, but with a cost. These resources can be complemented with self owned resources such as a private cloud. To keep the cost at a

S. Bosmans
University of Antwerp, Paardenmarkt 92,2000 Antwerp,Belgium, e-mail: stig-bosmans@gmail.com

G. Maricaux
University of Antwerp, Paardenmarkt 92,2000 Antwerp,Belgium, e-mail: glenn.maricaux@student.uantwerpen.be

F. Van der Schueren
University of Antwerp, Paardenmarkt 92,2000 Antwerp,Belgium, e-mail: filip.vanderschueren@uantwerpen.be

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_76

minimum, a scheduler will distribute a minimum of workload to the public cloud, while the workload will still be processed before a given deadline. Related work shows how a cost optimization scheduler can reduce the cost to process a workload [10] [1]. However, this research uses heuristics to estimate the workload. In this paper, various (non-heuristic) techniques will be discussed to create a cost efficient scaler which will allocate resources in the public and private cloud.
In our research the workload is defined as a parameter sweep which is an application that executes the same piece of code multiple times with a set of input parameters. For example the surface of an airplane wing or the weight of an airplane can be parameters to determine the fuel consumption of the plane. Each combination of different input parameters is referred to as a job. Our research focuses on how we can cost efficiently schedule a parameter sweep application in a hybrid cloud environment while respecting the deadline. In order to do that prediction techniques are used to estimate the total execution time to construct a reliable model as fast as possible. Different sample selectors schedule the workload of the parameter sweep by increasing priorities, the higher the priority the sooner the job has to be executed. A scaler is used to distribute the workload on both the private and the public cloud. Also, the scaler is responsible to start extra instances when the predicted runtime exceeds the deadline. To respect the cost it might be possible that the scaler will stop public instances in order to save money.

2 Sample selector

The sample selector assigns priorities to each individual job of a parameter sweep. The higher the priority of the job, the sooner that the job will be executed. Once a job or parameter combination is executed the runtime of that job is logged into a database. These runtime measurements are then used as scattered points in the prediction model (section 3). The order of the execution of the parameter combinations is very important in regard to get a reliable prediction within an acceptable timeframe. In order to have a reliable prediction, testing reveals that the whole range of parameters should be covered.

Three different sample selectors were used in our framework:

1. **Sequantial:** When using a sequential sample selector, parameter values will be fetched from the database one after another. This will make it very difficult to estimate the runtime at an early stage because only one part of the parameter range is taken into account.
2. **Random:** A random sample selector will select random parameters within the parameter range. The random sample selector is the most effective because each part of the parameter range is taken into account. Experimental results show that a wide area of the parameter sweep is covered at an early stage.
3. **Spread Algorithm:** The goal of this custom algorithm is to maximize the spread between the selected samples. The algorithm will select unprocessed samples so that the distance between processed samples is the greatest possible. Currently this algorithm is still outperformed by the random sample selector.

3 Prediction Techniques

To execute a parameter sweep application as cost efficient as possible and to respect a deadline, it is necessary to estimate how much resources will be required to execute a complete parameter sweep. Regression algorithms are used to create an estimate of the runtime and of the corresponding resources required to construct a reliable model as fast as possible. A significant dataset needs to be fetched to the algorithms in order to obtain a reliable estimation. Depending on this estimation the framework will automatically start or stop cloud instances to optimise the cost. In our research we implemented three different mathematical approaches to predict the duration: Polynomial, Radial basis functions and Kriging [3].

3.1 Techniques used

All three techniques are able to interpolate discrete scattered data [4]. In our case the scattered data represents the durations of already executed jobs.

1. **Polynomial:** When using the polynomial regression technique, a polynomial will be constructed with runtimes of jobs that are already executed. The constructed polynomial model approximates the known points as good as possible using the least squares method. To limit the processing power required and prevent the polynomial from overfitting, the degree of the polynomial will be set independent of the number of points provided to construct the polynomial. This results in a polynomial which will not intersect with the provided points, but they will be approximated as good as possible using the least squares method.
2. **Radial Basis Functions:** The RBF model constructs a function around all the known runtimes. The weights of these functions will be adapted with the least squares method to approximate the model consisting of the existing durations as good as possible. RBF has a faster convergence and is more accurate in many situations. Different kernel functions [6, 7] can be implemented in the radial basis functions.
3. **Kriging:** Kriging is very similar to radial basis functions, but Kriging evaluates the kernels in every dimension of a sample [7] this results in better results when multiple dimensions are used. The dimensions of the samples are defined by the number of parameters. Just as with radial basis functions the three described kernel functions are implemented in the framework.
4. **Other prediction techniques:** In related work [7], the use of neural network were investigated, but the results were very poor for our use case. Our developed framework is modularly constructed so that other prediction techniques can be easily incorporated.

3.2 Overhead

Each job has to be fetched from the database before it can be executed and after the execution the result has to be written to the database. These operations take about 1,5s on average in the test environment described in section 6. When a single parameter sweep consists of 1000 jobs we will have to take 1500s of overhead into account. The overhead is monitored and logged to the database so that the real total duration of the parameter sweep can be estimated.

4 Scaling

Scaling keeps the balance between the provided deadline and the necessary cost optimization. The scaler will evaluate the predicted remaining durations and consider whether the current deployed cloud resources are feasible to reach the deadline. If the predicted duration exceeds the deadline, extra resources will be allocated. The deadline is evaluated multiple times before actually scaling. If there is no hardware available to host a private cloud, it might be beneficial to use a public cloud provider. To reduce costs, the scaler will always opt for private instances unless these are already exhausted. The scaler will also monitor if the allocated resources are redundant. In the framework multiple scaling techniques can be used:

1. **Private, public or hybrid scaling mode:** Scaling modes provide more control over different clouds. When using private or public cloud, only the selected one will be used. In a hybrid environment, both clouds will be used, with the private cloud as the preferred cloud provider. If the private cloud provider is not able to provide the requested resources, public instances will be used instead.
2. **Strict versus Tolerant scaling:** When using Amazon EC2 as a public cloud provider usage per hour will be charged. If the instance is only used for 5 minutes and terminated afterwards, a full hour will still be charged. To make a cost effective scaler, it is necessary to keep the remaining time of the instances in mind. When set to tolerant, scaling will not occur if the deadline will be missed, but there is enough time left in running instances to finish the parameter sweep. The maximum time passed a deadline with tolerant scaling is 1 hour. If the deadline is very important, the strict scaling method will deploy instances, without considering remaining time for the running instances. Strict and Tolerant scaling has no effect in a private cloud.
3. **Private Greedy:** When a private greedy scaler is used, all resources available to the private cloud will be allocated. Using This overallocation of resources will cause a parameter sweep to be processed as fast as possible. If a deadline

for a specific parameter sweep requires more resources than the private cloud can provide, the scheduler would use all private resources, whether the private greedy method was used or not. Private greedy is only available when using a private cloud (private/hybrid mode).

5 Cloud Framework

5.1 Overview

The architecture of the framework consists of three major components:

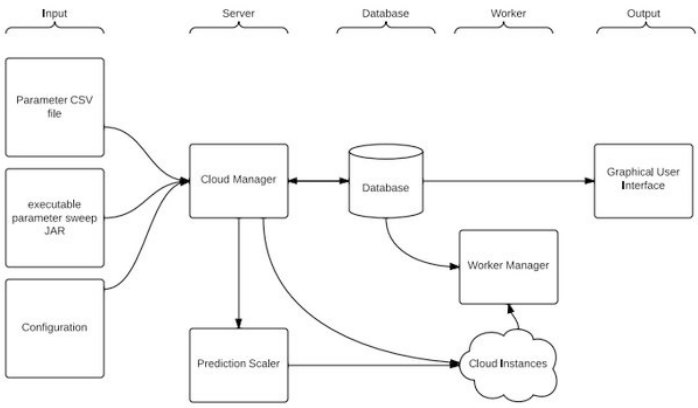


Fig. 1 Overview architecture of the cloud framework

1. **Cloud Manager:** The cloud manager is the heart of the framework. It will generate all possible parameter combinations of a parameter sweep and it will write these parameter combinations to the database. At the same time the server installs, configures and launches the workers. The workers are public or private cloud instances operating on a Linux Ubuntu operating system. Once the workers are launched, they will get and execute the next job with the highest priority. The worker will then write the result and the execution time to the database. The server continuously reads the measured runtime of the jobs that are already executed from the database. Using these runtimes it can build a prediction model which estimates the total execution time of the entire parameter sweep. This process will repeat until all the jobs are executed.
2. **Database:** A MySQL database with an InnoDB engine was used for the framework because of its row-locking features. Without these it might be possible for

jobs to be fetched and processed multiple times. A MongoDB implementation for example with a shadow table which overcomes this shortcoming was implemented as well. Nevertheless MySQL was better suited to deal with the different synchronisation issues. Also, it must be noted that the MySQL database has a negative impact on the overhead and thus the overall performance of the framework

3. **Worker:** The worker will execute the actual calculations, it can be deployed on either a private or public cloud instance. In our research OpenNebula was used as a private cloud provider and Amazon EC2 as the public cloud provider. The greatest advantage to use a private cloud provider is that it runs in a controlled environment, however resources are limited to the available hardware. To overcome this, a public cloud provider is also incorporated in the framework. The great advantage of using a public cloud is that there is a virtually unlimited amount of resources available on demand, but (in most cases) at a greater cost than a private cloud provider.

6 Test Results

6.1 Test Setup

In this test setup two simulations of a parameter sweep application were build. Each simulation uses only one input parameter. We decided to use simulations instead of real parameter sweep applications because this gives us the possibility to fully control the runtime of each job. The input parameter immediately impacts the runtime because this parameter is passed into a combination of a polynomial and kriging model, the output of these two models results into the final duration.

Each simulation uses a different polynomial and Kriging model and in one of the two simulations pseudo random noise was added. The use of only one parameter enables us to easily plot the data into a two-dimensional graph, which is more convenient to evaluate.

6.2 Results

In this section we will evaluate all implemented prediction techniques, scheduling techniques and sample selectors. All tests were executed ten times to be statistically significant.

6.2.1 Sample Selectors

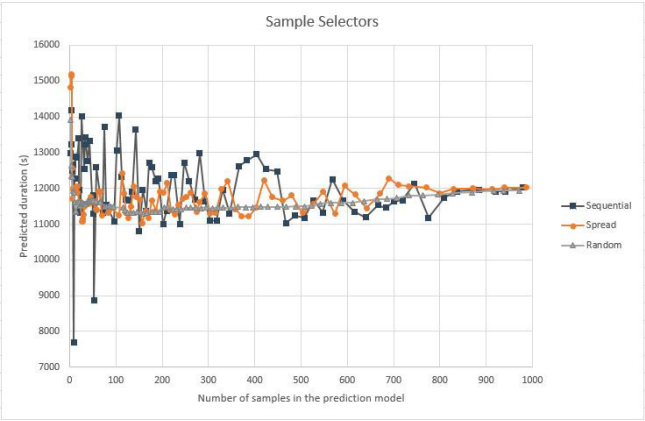


Fig. 2 Three sample selectors combined - Real runtime: 12003s

Figure 2 shows the predicted total duration in function of the number of samples added into the prediction model. To come to these results the parameter sweep application with noise described in section 3.1 was used together with a radial basis function predictor and a multiquadric kernel. The real total duration was on average 1203 seconds.

The figure clearly illustrates that the sequential sample selector is of no use because it fluctuates too much. The spread sample selector performs better but still isn't perfect because it stabilizes only after 800 input samples. Optimization of the spread algorithm to decrease the gaps between the selected samples might improve this. The results of the random sample selector are very stable. Tests with other configurations and parameter sweep applications confirm these results.

6.2.2 Predictions

All three prediction techniques and all kernels discussed in section three were tested multiple times with different applications. The three best techniques are covered: Polynomial, Kriging with an inverse multiquadric kernel and RBF with a multiquadric kernel. The predicted duration in function of the number of samples added into the prediction model are plotted in Figure 3 and 4. Both tests were performed using a random sample selector. Figure 3 uses the parameter sweep application with noise and Figure 4 uses the parameter sweep application without noise.

Both graphs show that it takes a to long until the Kriging technique reaches a trustworthy estimate. Other techniques perform better, the RBF technique and the

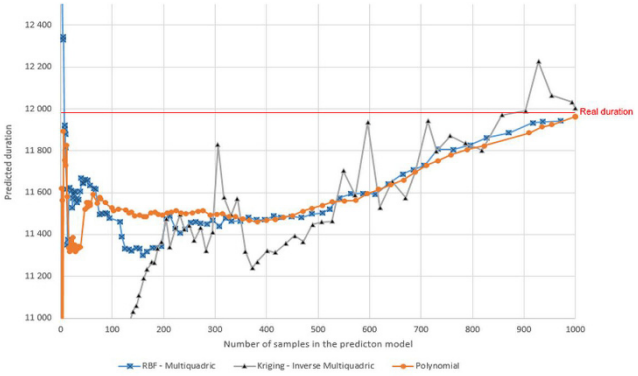


Fig. 3 Prediction techniques compared - Noise added

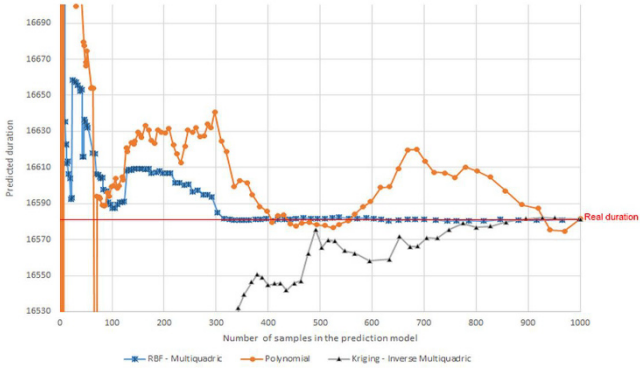


Fig. 4 Prediction techniques compared - Without noise

polynomial technique are well matched. However most of the tests with different parameter sweep applications pointed out that the RBF technique performs slightly better.

6.2.3 Scaling

- Test setup 1: No scaling, 2 instances started
- Test setup 2: Scaling, started with 2 instances (tolerant)
- Test setup 3: Scaling, started with 2 instances (strict)
- Test setup 4: No scaling, 12 instances started

Tests were performed in Public mode, the deadline was set one hour later, using radial basis functions with a multiquadric kernel and a random sample selector.

As can be seen in the results (Figure 5), scaling has a significant impact, even with

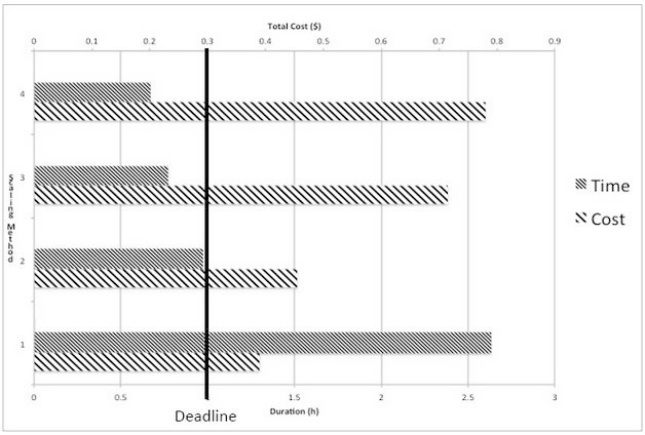


Fig. 5 Cost and duration of a parameter sweep with scaling

a very short deadline and relatively simple parameter sweep. We can see that test setup one (without scaling) processed the full parameter sweep the fastest, but it is also the most expensive option.
The other test setup without scaling, setup four is the cheapest setup, but the deadline was abundantly exceeded. In both setups where scaling was used, the deadline was met, but when using the tolerant method, parameter sweep was executed as cost effective as possible.
The reason why the strict method is almost as expensive as the setup without scaling is because of its behaviour to strictly reach the deadline. When the predictor's estimation went up at the end, the scaler deployed extra instances just a few minutes before the deadline, while the tolerant method would consider the remaining time of the running instances.

7 Conclusion

Different techniques are discussed on how a parameter sweep application can be cost efficiently scheduled in a hybrid cloud environment while respecting the deadline. Prediction techniques are used to determine the remaining execution time. The radial basis function with a multiquadric kernel together with the polynomial regression technique showed the best results depending on the application. In order to schedule the jobs of the parameter sweep a sample selector was used. The sample selector assigns priorities to each job, the higher the priority the sooner it will be executed. The execution order has a great impact on the predicted execution times. Results showed the random sample selector as the better one. The spread algorithm is also very promising but it needs further optimization. The implementation of various scaling methods makes it possible to launch extra instances in the private or

public cloud when the estimated execution time exceeds the deadline. On the other hand the scaler will also downscale to not over-allocate resources and thus reduce the cost. Globally we can conclude extensive cost optimisation in deadline dependent parameter sweeps can be reached by introducing runtime prediction in scaling and load balancing algorithms on a public, private or hybrid cloud environment.

8 Future Work

The techniques used and implemented in the framework can still be greatly expanded or combined with other techniques for greater control and options. The ability to define multiple parameter sweeps and scale on different levels would be a great addition. If there are multiple parameter sweeps, the discussed techniques can be used to implement further prioritization and scaling can be implemented. It's also a challenge to further reduce the overhead of the framework.

References

1. H. Casanova, A. Legrand, D. Zagorodnov and F. Berman: Heuristics for scheduling parameter sweep applications in grid environments. In: Heterogeneous Computing Workshop, 2000. (HCW 2000) Proceedings. 9th
2. P. Hellinckx, S. Verboven, F. Arickx and J. Broeckhove: Predicting Parameter Sweep Jobs: From Simulation to Grid Implementation. In: 2010 International Conference on Complex, Intelligent and Software Intensive Systems
3. P. Hellinckx, S. Verboven, F. Arickx and J. Broeckhove: Runtime Prediction in Desktop Grid Scheduling. In: Parallel Programming, Models and Applications in Grid and P2P Systems
4. Y. Beeckmans: Predictive Scheduling of Parameter Sweep Jobs Predictive Scheduling of Parameter Sweep Jobs in the Cloud.
5. Open-source software for volunteer computing and grid computing http://boinc.berkeley.edu/
6. Grady B, Wright: Radial Basis Function Interpolation: Numerical and Analytical Developments
7. P. Hellinckx, S. Verboven, F. Arickx and J. Broeckhove: Runtime Prediction Based Grid Scheduling of Parameter Sweep Jobs
8. Armbrust, Michael and Fox, Armando and Griffith, Rean and Joseph, Anthony D. and Katz, Randy H. and Konwinski, Andrew and Lee, Gunho and Patterson, David A. and Rabkin, Ariel and Stoica, Ion and Zaharia, Matei: Above the Clouds: A Berkeley View of Cloud Computing http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-28.html
9. P. Hellinckx, S. Verboven, F. Arickx and J. Broeckhove: Dynamic Grid Scheduling Using Job Runtime Requirements and Variable Resource Availability. In: Euro-Par - Lecture Notes in Computer Science
10. Bittencourt, LuizFernando and Madeira, Edmundo Roberto Mauro: HCOC: a cost optimization algorithm for workflow scheduling in hybrid clouds
11. P. Hellinckx, F. Arickx, G. Stuer and J. Broeckhove: The CoBRA Grid: a Highly Configurable Lightweight Grid http://dx.doi.org/10.1504/IJWGS.2007.014954

Part VII
Workshop BIDS-2016: International Workshop on Business Intelligence and Distributed Systems

A Study of the Internet Financial Interest Rate Risk Evaluation Index System in Context of Cloud Computing

——Rough Set-Particle Swarm-based Support Vector Machine (SVM) Technology

Mu Sheng-dong[1], Tian Yi-xiang[1], Lili[1], Xu An Wang[2]
1 School of Management and Economics,
University of Electronic Science and Technology of China, Chengdu 614100
2 Engineering University of CAPF, Xi'an 710086

Abstract: Cloud computing is a product of computer technologies combined with network technologies and it has been widely applied in China. Experts and scholars in all fields begin to make many studies of cloud computing infrastructure construction and effective resource utilization. With the improvement of cloud computing technology (especially security technology), Internet finance will be deployed widely and will develop rapidly. ITFIN (Internet finance) is the results of finance comprehensively combined with network technology and also a new ecological finance fermenting in this Internet era. In ITFIN, it integrates online transaction data generating in various social network, studies and judges the credit standing of customers, and completes credit consumption, loan and other borrowing behavior by e-payment, so that people can enjoy financial services in dealing with various problems. However, one person can play different identities in network, which has posed a severe challenge to ITFIN network security and has largely intensified risk including operational risk, market selection risk and other business risks as well as network and information security risk. Therefore, it is an effective technological approach and operable measure to prevent or eliminate ITFIN risks by establishing a reliable, reasonable and effective ITFIN risk assessment model. This paper conducted theoretical and empirical analysis, then constructed an assessment model against China's ITFIN risk that integrates rough set and PSO-SVM (particle swarm optimization support vector machine, and finally used this model in China's ITFIN risk assessment. The empirical research results indicate that, through introducing rough set theory, it could effectively reduce redundant data information, guarantee a reliable, reasonable and scientific model, and enhance classification effect of model. The parameters of SVM model obtained by optimizing with PSO could effectively avoid local optimum, it objectively can improve the effect of the classification model, this model has good generalization ability and learning ability.

Keywords: Cloud Computing ;ITFIN; Risk assessment; Rough set; PSO; SVM

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_77

Introductions

In 2013, the upsurge of ITFIN took on an unstoppable trend. Ali finance increasingly expanded its business scope, Tencent launched its fund strategy, and JD.com started to exert into creating supply chain finance. In the meantime, some traditional financial tycoons took actions in succession: the China Construction Bank brought its "Shan-Rong Business" online; Industrial and Commercial Bank of China created a "large e-commence platform"; Agricultural Bank of China initiated and established an "ITFIN lab" etc [1-3].

Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources which includes network, server, storage, application and service that may be provided and released quickly through minimized management or interaction with service providers. Called "cloud", such a pool of resources is made up of some self-maintained and managed virtual computing resources which may be expanded infinitely, acquired at any place at any time, used on demand, and paid as used. Networks providing resources are divided into separate pools of resources, which may be provided and released quickly and may be monitored and controlled. The basic principle of cloud computing is to distribute computing to a number of distributed computers instead of local computers or remote servers. A cloud computing system comprises of two parts, i.e. front end and rear end, which are connected via a network. A front end refers to an extended terminal, including various cloud terminals and wireless access devices. A rear end refers to a system's computer cluster made up of control nodes, databases, application servers, and computer networks, which are integrated in a trans-boundary way. The application programs needed by a user does not run on a terminal, such as a PC or a mobile phone, at the front end but run in a large-scale server cluster on the Internet. The data processed by a user is not stored locally but stored in a data center on the Internet. Cloud computing service providers can switch resources over to the application needed and access a computer and a storage system on demand.

Currently, regarding Internet financial risk research in countries out of China, the majority used empirical method, which employed a viewpoint of customer perception risk to study ITFIN risks. They thought a root reason that restricted financial business to expand in Internet platform lied on that customers had lost trust on ITFIN. While, in China, many scholars utilized analytic hierarchy process, fuzzy mathematics and unascertained models or methods to evaluate risk management in ITFIN business; in existing researches, few analysis and conclusion were made on key factors that influenced ITFIN risks from an empirical viewpoint. With prosperous development of ITFIN, financial risks aggravates day by day. Thus it becomes particularly important and urgent to implement domestic ITFIN risk assessment and relevant empirical research, which also involves whether domestic financial industry could stably and healthily develop in future [2-5].

1.1 Technical frameworks and key technologies of could computing

The key concept of cloud computing is "everything is service" and it is a collection of services. The technical frameworks of cloud computing include three forms, i.e. IaaS (Infrastructure as a Service), PaaP (PowerShell as a Platform), and SasS (Software as a Service). Cloud computing provides IT services to users with a unprecedented scale based on such key technologies as virtualization technology, parallel programming model, massive data distribution & storage technology, massive data management technology, and cloud computing platform management technology.

The IaaS provides processing and storage services to users, and its common services include communication services provided by networks and communication systems, computing services provided by server devices, and storage services & operating system services provided by data storage spaces.

As a complete virtual platform, the PaaS aims at program developers. It provides such services as end-to-end distributed software development, tests, deployment, operating environments, and complicate program hosting to simplify program development and deployment by making use of programming languages and tools provided cloud service providers.

The SaaS provides software services via the Internet. Service providers are responsible for maintaining and managing software and providing hardware facilities for software running. A user may use software at any place at any time as long as he/she gets an access to a SaaS network terminal.

Through virtualization technology, computing resources are abstracted and specific technical characteristics are encapsulated and hidden while uniform logic interfaces are provided so as to avoid variance caused due to diversified physical devices. By means of virtual machine encapsulation technology, resources are shared, private servers are configured to realize customized resources, and physical servers are divided into several virtual machines to realize fine-grained resource management.

Parallel programming model: cloud computing is a kind of simplified distributed programming model and highly-effective task scheduling model. A simple program is complied in the context of cloud computing to realize specific purposes.

Massive data distribution & storage technology: cloud computing stores data in the form of distributed storage. Redundant storage is used to guarantee the reliability of data stored. Highly-reliable software is used to rule out unreliability of hardware.

Massive data management technology: through cloud computing platform management technology, a number of servers may work collaboratively to ensure that business is deployed and opened easily, system failure is found and restored quickly, and information technology is used to make a large-scale system work reliably.

Through cloud computing platform management technology, a number of servers may work collaboratively to ensure that businesses are deployed and opened easily, system failure is found and restored quickly, and information technology is used to make a large-scale system work reliably.

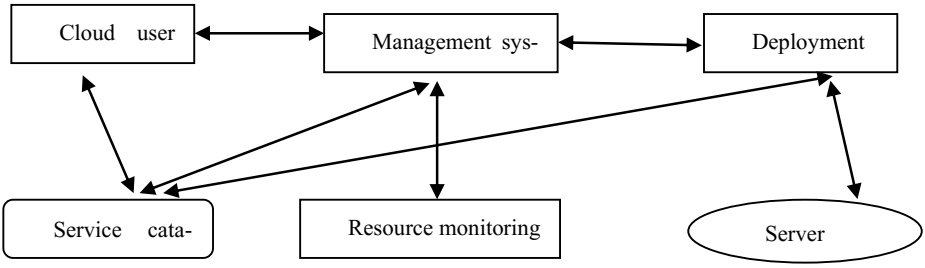


Fig. 1: Cloud computing system structure

1.2 Basic operation of rough set

(1) Decision information system and decision sheet reductions

$S=(U, A, V, f)$ is simply denoted as $S=(U, A)$, named as information system. The universe of discourse U is a non void and limited set, A is a non void and limited attribute set, any attribute $a \in A$, V_a refers to its range, i.e. $\forall x \in U, a(x) \in V_a, V=\cup V_a$, which means V is the range of A.

(2) Data analysis method of information sheet deduction

Regarding given information system $S=\{U,R\}$, when we remove an attribute a from R one by one and immediately examine whether the classification ability of R after such remove, if classification ability remains unchanged, i.e. $U/\{R-a\}=U/R$, then attribute a could be reduced; instead, if classification ability changes, i.e. $U/\{R-a\} \neq U/R$, then a is coral attribute of R. The specific method is: as to one known information sheet, it is possible to remove the row where attribute a is located, and in case no new and same line appears, a could be reduced, which is to say, at this moment, a is not a coral attribute; otherwise, it is impossible to reduce a, which is to say, a is a coral attribute [7-9].

Reduction process as follows [7-9]:

Step 1: Find the core. After examining p, q and r in R, only p could not be reduced. So only p is a coral attribute in P.

Step 2: Examine whether it construct a reduction. If yes, it is a sole and smallest reduction. Otherwise, take next step. Examination method: examine and check if the classification equivalence class of U is the same with the classification equivalence class of U divided by R. Specific method: simply see if new same line appears after removing q and r in sheet. As line 4 & 5 are new same lines, so the core cannot construct reduction.

Step 3: Each time, randomly take an attribute out of coral attribute and incorporate in coral attribute, then see if it could be further reduced. As coral attribute P cannot construct reduction (for that its classification ability differs from the classification ability of R), therefore, the coral set is not yet reduction. If the coral set is reduction, then it is the sole and smallest reduction that people expect. But the real application is normally not like this, as it needs to add partial off-core attributes to construct reduction. Method: each time, randomly take one attribute from off-core attributes and incorporate into the core, and see whether it could gain same classification ability with R.

(3) Decision rule reduction

Attribute reduction can be divided into information sheet reduction and decision sheet reduction. We assume to conduct reduction to information system $S=\{U,A\}$, in other words, to conduct rough set reduction to entire attribute set A, i.e. it requires us to use least attribute set to contain equal amount of feature information with original attribute set A. If $a \in A$ is redundant attribute, we need to remove a to obtain $A-\{a\}$, and this $A-\{a\}$ possesses same distinguish and recognition abilities with A, i.e. it has equivalence class family below [7-9]:

$$U/A-\{a\}=U/A \tag{1}$$

If $A-\{a\}$ is relatively dependant, removing of any one attribute in $A-\{a\}$ would change its distinguish and recognition abilities, then we can consider $A-\{a\}$ is reduced A. This reduction of A is relatively independent certain subset P in A, and this subset P possesses same distinguish and recognition abilities with A, which could also explain that, all attribute have no impact on P and belong to a basic feature of information sheet formed by redundant attributes. At this time, attribute set A cannot be divided into condition attribute set and decision attribute set.

However, as to decision sheet $T=(U, A=C \cup D)$, the reduction is not completely the same, because its reduction is not only conducted on entire attribute set A, but also on condition attribute. Thereby the reduction of decision sheet can be divided into two parts:

Step 1: Attribute set reduction

In case $P \subseteq C$ meets that P is independent from decision attribute D, and $Pos_P(D)=Pos_C(D)$, then P is D reduction of C. Intersection $\cap RED_C(D)$ of all D reduction in C is called the core of C, denoted by $Core_D(C)$.

Step 2: Attribute value reduction

Suppose $T=(U, C \cup D)$ is an uniformity decision sheet, then we consider $P \subseteq C$ is reduction of C to D. With regard to reduction value reduction P, it could also say that reduction of attribute value is expressed under each decision principle of above decision sheet. In case one certain rule of decision sheet can be reduced, when and only when this attribute value is reduced, such rule still possesses original uniformity principle, i.e. no situation shows to not uniform with such rule [7-9].

2 SVM Algorithms

The SVM method, as structural risk prevention, is previously used to solve classification problem and now possesses perfect theory. *SVR* learning algorithm as below [10-14]:

Suppose the training data $\{(x_i, y_i), i=1,2,\cdots,l\}$ is given, where, $x_i \in R^d$ is an input value of learning sample of number i and a one-dimensional vector $x_i=(x_i^1, x_i^2, \cdots, x_i^d)$, $y_i \in R$ is the corresponding target output value. At first, define ε insensitive loss function as

$$|y-f(x)|_\varepsilon = \begin{cases} 0 & |y-f(x)| \le \varepsilon \\ |y-f(x)|-\varepsilon & |y-f(x)| > \varepsilon \end{cases} \tag{2}$$

Through appropriately defined coral function $k(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$ ($\phi(x)$ as one certain nonlinear function, we nonlinear transform input sample space into another feature space for linear regression estimation, thus we suppose nonlinear classification estimation function to be [10-14]

$$f(x) = \omega \cdot \phi(x) + b \tag{3}$$

In above equation, $\omega \in R^d$ refers to weight value vector, $b \in R$ refers to threshold value, and $(\cdot)$ represents inner product operation. Introducing slack variable ξ_i, ξ_i^* , so optimization problem changes into

$$\begin{aligned} &\min_{w,b,\xi_i,\xi_i^*} \|w\|^2 / 2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) \\ &s.t. y_i - (w \cdot \phi(x_i) + b) \le \varepsilon + \xi_i \\ &(w \cdot \phi(x_i) + b) - y_i \le \varepsilon + \xi_i^* \\ &\xi_i, \xi_i^* \ge 0, i = 1,2,\cdots,l \end{aligned} \tag{4}$$

Implementing Lagrange conversion, its pairing optimization problem is as below:

$$\begin{aligned} &Max[-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(a_i - a_i^*)(a_j - a_j^*)k(x_i, x_j) - \varepsilon\sum_{i=1}^{l}(a_i + a_i^*) + \sum_{i=1}^{l} yi(a_i - a_i^*)] \\ &s.t.\sum_{i=1}^{l}(a_i - a_i^*) = 0 \\ &0 \le a_i, a_i^* \le C \end{aligned} \tag{5}$$

Where, C is a positive constant, called penalty parameter. We seek to solve above optimization problem, and the training sample corresponding to $(a_i - a_i^*) \ne 0$ is support vector, thus we obtain $w = \sum_{i=1}^{l}(a_i - a_i^*)\phi(x_i)$. The formula as below:

$$b = \frac{1}{N_{Nsv}}\left\{\sum_{0 \le a_i \le C}[y_i - \sum_{x_j \subset sv}(a_j - a_j^*)k(x_j, x_i) - \varepsilon] \sum_{0 \le a^* \le C}[y_i - \sum_{x_j \subset sv}(a_j - a_j^*)k(x_j, x_i) + \varepsilon]\right\} \tag{6}$$

In above equation, N_{Nsv} is the quantity of standard support vector and the learnable and obtained classification function is:

$$f(x) = \sum_{x_i \in sv}(a_i - a_i^*)k(x_i, x) + b \tag{7}$$

3 ITFIN Risk Assessment Model Based on Rough Set and PSO-SVM

Use of the advantages of the two theories respectively, and can make use of rough set to the original data preprocess, then uses the support vector machine (SVM) to black box processing, the processed data in this way can better avoid inherent in these two methods of imperfections, and play their respective advantages to the greatest extent. Through rough set theory, we do not need to depend on any prior knowledge or information to extract correspondent features of data, reduce redundant information, acquire coral indicator system, further give full play to information process

merit of rough set theory, and thus provide objective, reliable and accurate data source for subsequent SVM processing. In late period of precise leaning training and assessment, the complexity of SVM original data input and the efficiency of training model will be largely decreased, but classification accuracy and scientific nature. On basis of above statement, this paper highly integrated rough set with SVM algorithm to build ITFIN risk model, see basic procedure as below and detailed flow chart in Fig.2.

(1) In order to effectively evaluate ITFIN risk, it is a must to guarantee and build a complete and comprehensive indicator system. Firstly, establish a widely generalized indicator system, then collect, clear up and classify according to factors that influence enterprise financial risk, and find out the relation between each influence factor with ITFIN risk, i.e. build one set of systematic, scientific and comprehensive integral assessment indicator system.

(2) Rough set theory is used to reduce the comprehensive indicator system established in step 1. Specific procedure: conduct reduction of decision sheet formed from these indicators, then find out from these data which are coral attributes and which are redundant attributes; due to complexity of attribute reduction algorithm, this paper used MATLAB programming to implement attribute reduction. This indicator system reduced redundant attribute and extracted coral attributes so that laid a solid foundation for SVM classification recognition.

(3) SVM trains the data in these indicator systems till they meet the requirement of accuracy, and related parameters shall be optimized via PSO algorithm during training.

(4) Apply well-trained model to make early warning.

After above four steps, we can obtain ITFIN risk assessment results.

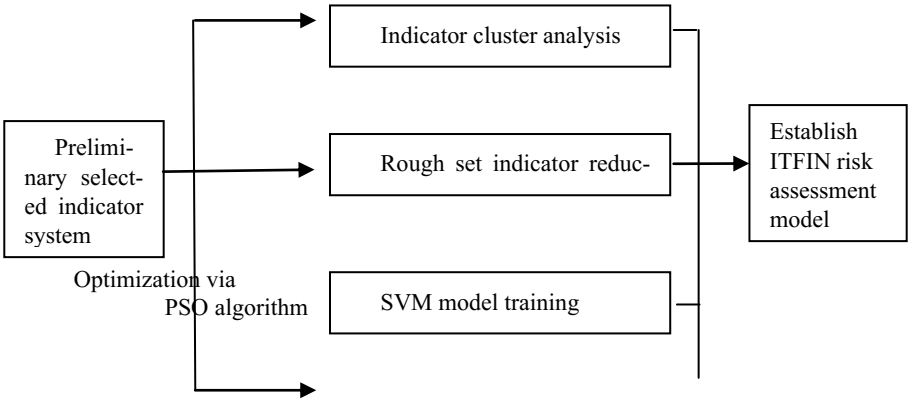


Fig.2 Basic flow chart of combined rough set and SVM

4.1 Selection principle of ITFIN risk assessment indicators

when establishing ITFIN risk assessments indicator system, we have to follow the principles below:

(1) Scientific principle. The comprehensive assessments indicator system should be constructed on a reasonable, standard and scientific premise, so as to accurately, objectively and faithfully reflect ITFIN risk.

(2) Comprehensive and information abundant principle. We have to choose comprehensive and abundant information as possible for comprehensive assessments indicator system, so that it could perfectly reflect every aspects of ITFIN risk.

(3) Quantitative and qualitative combined principle. In building comprehensive assessments indicator system, it is better to choose indicators being liable to quantize or already quantized to describe; regarding qualitative indicators hard to quantitatively describe, we have to try every efforts to quantize them.

(4) Acceptable and comparable principle. When choosing each kind of indicators for comprehensive assessments indicator system, we have to meet comparability of space time as possible, adopt one standard to analyze, calculate and summarize each indicators, use indicators with better comparability and evident common features as possible, and simultaneously clarify physical implication of each indicator and reasonably determine its content and range.

(5) Dominant principle. In selection and establishment course of comprehensive assessments indicator system, we should select indicators with high comprehensiveness or representativeness as possible.

(6) Operable principle. The selected indicators for assessment indicator system should be convenient for follow-up operation, such as explanation, quantification and analysis. Besides, these indicators should be more reliable to obtain basic data.

4.2 Establishment of ITFIN risk assessments indicator system

This paper utilized the selection principle of ITFIN risk assessments indicator system stated in Section 4.1 and the domestic and foreign research results, and consulted many ITFIN risk assessment experts for their opinions to finally determine to select indicators, such as network security risk, operation risk, credit risk, financial business risk, legal and reputational risk, to establish appropriate ITFIN risk assessments indicator system and compose of original assessment set, details as shown in Table 1. In this paper, the author set ITFIN risk assessment grading as five grades, such as “Very High”, “High”, “Medium”, “Low” and “Very Low”. In which, according to the basis of grading as shown in Table 2, we invited 50 ITFIN risk assessment experts to use the proposed model to make assessment. Those invited experts shall consider both the probability of possible risks and the resulted loss once risk takes place when they mark to grade of risk.

Table 1. INFIN risk assessments indicator system

S/N	Name of indicator	Mark
1	Risk of tampered website	X_1
2	Risk of webpage Trojan horse	X_2
3	Risk of network phishing	X_3
4	Risk of website back door	X_4
5	Risk of network delay	X_5
6	Risk of payment mode innovation	X_6
7	Risk of payment mode	X_7
8	Inter-industrial correlation risk	X_8
9	Mal-operational risk	X_9
10	Consumer operational risk	X_{10}
11	Internal fraud risk	X_{11}
12	External fraud risk	X_{12}
13	Personal credit risk	X_{13}
14	Economic environmental risk	X_{14}
15	Policy risk	X_{15}
16	Liquidity risk	X_{16}
17	Market risk	X_{17}
18	Industry risk	X_{18}
19	Business risk	X_{19}
20	Legal leg risk	X_{20}
21	Subject qualification risk	X_{21}
22	Ideal money risk	X_{22}
23	Risk of network money laundering	X_{23}
24	Reputational risk	X_{24}

Table 2. Risk Grading Standards

Grade	Standard	Description
1	Very Low	Once risk occurs, there will be scarcely no influence that could be remedied with simple measures.
2	Low	Once risk occurs, there will be low influence that could be solved via ordinary internal treatment.
3	Medium	Once risk occurs, there will be certain economic or production influence in controllable range.
4	High	Once risk occurs, there will be large economic or production influence that can result in organizational damage within certain scope.
5	Very High	Once risk occurs, there will be severe economic or production influence that can seriously affect organizational reputation.

4.2 Reduction of rough set attributes

Regarding the existing corresponding supervision value of risk assessment indicators, we can obtain the original data. Regarding the indicator with none qualitative data, we invited 100 professionals who have job categories similar to ITFIN and many year of working experience to evaluate each indicator data. After standard processing, the indicator data are all distributed between 0~1. Based on these, we utilized ROSETTA software to conduct discretization process and added risk condition decision attribute formed by cluster to form a decision table. In which, company is represented by code,

24 indicator data forms the condition attribute set, and risk condition belongs to decision attribute, see Table 3 for process results.

Table 3. Decision Table after discretization process

Code	X_1	X_2	X_3	X_4	X_5	…	X_{24}	D
1	0	1	2	1	2	…	2	…
2	0	2	1	2	1	…	1	…
3	0	2	2	2	2	…	2	…
4	0	1	1	2	2	…	0	…
5	0	1	1	2	2	…	1	…
…	…	…	…	…	…	…	…	…

Significance of each attribute in new indicator system that calculated based on attribute significance formula and software computational results is shown in Table 4.

Table 4 Reduced Indicator System

Indicator name	Denoted by	Significance of attribute	Indicator name	Denoted by	Significance of attribute
Risk of tampered website	C_1	0.0935	External fraud risk	C_{10}	0.1476
Risk of webpage Trojan horse	C_2	0.1206	liquidity risk	C_{11}	0.2375
Risk of network phishing	C_3	0.4872	Market risk	C_{12}	0.1231
Risk of website back door	C_4	0.2365	Industrial risk	C_{13}	0.1532
Risk of payment mode	C_5	0.1823	Business risk	C_{14}	0.1832
Inter-industrial correlation risk	C_6	0.2364	Legal lag risk	C_{15}	0.0556
Mal-operational risk	C_7	0.0265	Subject qualification risk	C_{16}	0.2512
Consumer operational risk	C_8	0.2689	Ideal money risk	C_{17}	0.2981
Internal fraud risk	C_9	0.2321	Risk of network money laundering	C_{18}	0.1965

The reduce indicator system contains 18 indicators, which fully and systematically cover network security risk, operational risk, credit risk, financial business risk and legal and reputational risk, thus they could be input data of PSO-SVM model. This enables us to establish corresponding assessment model more efficiently, reasonably and scientifically, so as to avoid excessive dimensionality of input data, bring down training effect of model and further influence the effect of ITFIN risk assessment.

5 Conclusions

This paper will put forward a series of ITFIN risk prevention advices on basis of above assessment results. According to risk assessment indicator system in this paper, relevant measures can be divided into four levels as below:
From the technical prospective, the following aspects should be considered in an Internet financial security solution: (1) High-performance security control functions;

security gateway devices are deployed at key positions, with high requirements for the transmission performance and the throughout rate. (2) Highly-effective protection functions against internal & external network attacks, e.g. firewall deployment, IDS and defense systems. (3) Highly-effective anti-virus functions; (4) High-performance application-layer control capacity to detect the statuses of data packets, restrict Session data based on application characteristics and IP addresses, and optimize the network efficiency; (5) Flexible and high-efficiency bandwidth management; The Quality of Service (QoS) should set a rate for each IP address. Application programs should be blocked and restricted strictly. Elastic QoS functions should be supported. There are powerful flow monitoring, statement and session control functions. (6) Effective authentication management mechanisms and flexible, safe, and reliable VPN applications, which can control specific network resource visitors; security gateway supports such authentication ways as local authentication, RADIUS and CA defined redundant servers. Through deployment of encrypted and certified VPNs, the data transmitted will not be stolen and falsified. (7) Network architectures with redundant backup functions, which can improve the stability and reliability of software and hardware systems; (8) Safe and reliable centralized application management; Internet financial networks are tremendous, trans-platform, and trans-regional networks. If centralized equipment deployment, monitoring, maintaining, auditing, and upgrading platforms are used, management workloads will be reduced greatly. Centralized end-to-end life-circle management can realize elaborate equipment configuration, network setting, VPNs configuration, and security strategy control.

1 References

[1] Xie Ping and Yin Long. Financial Theory and Governance under Network Economics [J]. Economic Research, 2001, 4: 12-16.

[2] Zhang Huolin. China's Security E-Commerce Development Problems and Countermeasures [J]. Finance and Computer in Southern China, 2003, 11:22-28.

[3] Shi Fangli & Wu Shimin. Research of Correlation between Financial Informationalization and Network Finance [J]. China's Management Informationalization, 2005, 12:39-44.

[4] Cai Xuefei. Brief Discussion on Internet Finance and Bank [J]. Technological Economical Market, 2007, 1:79-86.

[5] Liang Xun. Network Financial System Design and Realization Case Collection [M]. Beijing University Press, 2009s

[6] Cui Zeyuan. Network Finance [M]. Press of Dongbei University of Finance and Economics, 2011.

[7] Zhang Wenxiu, Wu Weizhi, Liang Jiye, etc. Rough Set Theory and Methods [M]. Beijing: Science Press, 2003.

[8] Pawlak Z.Rough Sets[J].Inter J of Computer and Information Sciences,1982,11(2):341-356.

[9] Zeng Huanglin. Rough Set Theory and Applications [M]. Chongqing: Chongqing University Press,1996.

[10] Vapnik V N. The nature of statistical learning theory [M]. New York: Springer-Verlag, 1995.

[11] Wen Wenbo & Du Wei. Overview of Ant Colony Algorithm [J]. Control System, 2002, (1) : 19-22.

[12] Van den Bergh F. An Analysis of Particle Swarm Optimizers [D]. University of Pretoria, South Africa, 2002.

[13] Jae H M, Y C L. Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters [J]. Expert Systems with Application, 2005, 28: 603~614.

[14] C W HSU, C J LIN. A Comparison of Methods for Multi-class Support Vector Machines [J]. IEEE Transactions on Neural Networks, 2002, 13(2): 415-425.

[15] R. Madala, A.G. Ivakhnenko. Inductive Learning Algorithms for Complex Systems Modeling [M]. Florida: CRC Press Inc, 1994: 69-86.

[16] Mueller A, Lemke F. Self-organizing data mining [M]. Herstellung Berlin: Libri Books on Demand, 2000: 63-83.

[17] He Changzheng. Slef-organizing Data Mining and Economic Predictions [M]. Beijing: Science Press, 2005. (In Chinese). 42-43.

[18] M.S. Sohail, B. Shanmugham. Information Sciences, 2003, 150(3-4): 207-217.

[19] B. Jaruwachirathanakul, D. Fink. Internet Research, 2005, 15(3): 295 – 311.

[20]International Journal of Space-Based and Situated Computing,2011 Vol. 1 No. 1 .3-17

[21]Evaluation of MANET protocols for different indoor environments: results from a real MANET testbed Masahiro Hiyama; Elis Kulla; Makoto Ikeda; Leonard Barolli .DOI: 10.1504/IJSSC.2012.047465.83-103

[22]A multi-objective approach for high quality tree-based backbones in mobile ad hoc networksApivadee Piyatumrong; Pascal Bouvry; Frédéric Guinand; Kittichai Lavangnananda.DOI: 10.1504/IJSSC.2012.047466.104-111

[23]Provisioning vehicular ad hoc networks with quality of serviceGongjun Yan; Danda B. Rawat; Bhed BahadurBista.DOI:10.1504/IJSSC.2012.047467.112-122

[24]Data transfer exploiting multiple heterogeneous challenged networks - implementation and applicationAkira Nagata; Shinya Yamamura; Masato Tsuru DOI: 10.1504/IJSSC.2012.047468.123-138

A Dynamic System development Method for Startups Migrate to C loud

Asmaa Abdelrehim Selim Ibrahim
Master Student, Nile University
0201022670050
a.elshopki@nu.edu.eg
Mohameed Mamdouh Awny
Professor, MOT at Nile University
021222106397
mmawny@nu.edu.eg

ABSTRACT

Cloud computing has become the most convenient environment for startups to run, build and deploy their products. Most startups work on availing platforms as a solution for problems related to education, health, traffic and others. Many of these platforms are mobile applications. With platforms as a service (PaaS), startups can provision their applications and gain access to a suite of IT infrastructure as their business needs. But, startups face many business and technical challenges to adapt rapidly to cloud computing. This paper helps startups to build a migration strategy. It discusses steps, techniques and approaches for moving from enterprise applications to cloud based applications. There are many several strategies for cloud migration but most of them are for SMEs. This research work, an approach for cloud migration to be followed by startups. In addition, cloud migration challenges facing startups are discussed. One to one meetings with 18 different startups who are moving to cloud computing approach have been conducted. In addition, a study about four successful startups which have already moved to cloud has been conducted, and their approach to cloud migration strategy is presented.

Keywords – Cloud migration; Startups; Agile Methodology; Migration Challenges; DSDM;

1. INTRODUCTION

Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.

The developing world's cloud computing sector has drawn considerable attention from global and local IT players, national governments and international

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_78

agencies. In theory, it is possible for developing economies to catch up with developed countries as the cloud allows them to have access to the same IT infrastructure, data centers and applications [2]. Clouds provide a powerful—and often unattainable—IT infrastructure at a modest cost. In addition, they free individuals and small businesses from worries about quick obsolescence and a lack of flexibility [3]. According to Blank (2012), a startup is "a temporary organization in search of a scalable, repeatable, profitable business model". By finding a scalable business model, entrepreneurs also need to scale their IT resources. However, with cloud technologies, cloud consumers do not have to think about overprovision in case the demand of computing resources increases. Cloud computing gives the ability to manage scaling instantly by adjusting the storage, server and network needs according to the consumers' demand. It gives the opportunity of accessing economies of scale that only the larger firms were able to take advantage of [4]. Startups have always been very good at bootstrapping, getting as far as possible on no or very little funding. But now, with cloud computing and high quality open source software, better and less expensive network, a more open mobile market, and more evolved customer base, the bootstrapping can take business very far forward possibly all the way to a self-supported profitable business [5].

Public clouds like Amazon AWS, Microsoft Azure, and Google AppEngine offer infrastructure and platforms as services over the internet. In public clouds, resources and costs are shared by users who use them over the internet on pay per use model. This model appeals especially for startups and small organizations that cannot invested in hardware resources and are looking for ways to avoid the large Capital expenditure (CAPEX) involved in procuring infrastructure upfront [6].

2. THE CLOUD MODEL

The cloud model is composed of five essential characteristics, three service models, and four deployment models [1].

2.1 Essential Characteristics:

- *On-demand self-service*. A consumer can unilaterally avail provision computing capabilities, such as server time and network storage, as needed automatically without requiring human interaction with each service provider.
- *Broad network access*. Capabilities are available over the network and accessed through standard mechanisms that promote the use by heterogeneous thin or thick clients' platforms (e.g., mobile phones, tablets, laptops, and workstations).
- *Resource pooling*. The provider's computing resources are pooled to serve multiple consumers using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumers' demand. There is a sense of location independence in that the customer generally has no control or knowledge over the exact location of the provided resources but may be able to specify location at a higher level of abstrac-

tion (e.g., country, state, or datacenter). Examples of resources include storage, processing, memory, and network bandwidth.

- *Rapid elasticity.* Capabilities can be elastically provisioned and released, in some cases automatically, to scale rapidly outward and inward commensurate with demand. To the consumer, the capabilities available for provisioning through the cloud system often appear to be unlimited and can be appropriated in any quantity at any time.
- *Measured service.* Cloud systems automatically control and optimize resource usage by leveraging a metering capability at some level of abstraction appropriate to the type of service (e.g., storage, processing, bandwidth, and active user accounts). Resource usage can be monitored, controlled, and reported, providing transparency for both the provider and the consumer of the utilized service.

2.2 Service Models:

- *Software as a Service (SaaS).* The capability provided to the consumer is to use the provider's applications running on a cloud infrastructure. The applications are accessible from various client devices through either a thin client interface, such as a web browser (e.g., web-based email), or a program interface. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, storage, or even individual application capabilities, with the possible exception of limited user specific application configuration settings.
- *Platform as a Service (PaaS).* The capability provided to the consumer is to deploy onto the cloud infrastructure consumer-created or acquired applications created using programming languages, libraries, services, and tools supported by the provider. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, or storage, but has control over the deployed applications and possibly configuration settings for the application-hosting environment.
- *Infrastructure as a Service (IaaS).* The capability provided to the consumer is the provision of processing, storage, networks, and other fundamental computing resources where the consumer is able to deploy and run arbitrary software that can include operating systems and applications. The consumer does not manage or control the underlying cloud infrastructure but has control over operating systems, storage, and deployed applications; and possibly limited control of select networking components (e.g., host firewalls).

2.3 Deployment Models:

- *Private cloud.* The cloud infrastructure is provisioned for exclusive use by a single organization comprising multiple consumers (e.g., business units). It may be owned, managed, and operated by the organization, a third party, or some combination of them, and it may exist on or off premises.

- *Community cloud.* The cloud infrastructure is provisioned for exclusive use by a specific community of consumers from organizations that have shared concerns (e.g., mission, security requirements, policy, and compliance considerations). It may be owned, managed, and operated by one or more of the organizations in the community, a third party, or some combination of them. It may exist on or off premises.
- *Public cloud.* The cloud infrastructure is provisioned for open use by the general public. It may be owned, managed, and operated by a business, academic, government organization, or some combination of them. It exists in the premises of the cloud provider.
- *Hybrid cloud.* The cloud infrastructure is a composition of two or more distinct cloud infrastructures (private, community, or public) that remain as unique entities, but are bound together by standardized or proprietary technology that enables data and application portability (e.g., cloud bursting for load balancing between clouds).

3. THE SELCTION OF CLOUD PROVIDER

In the following we discuss, how the startups could select the most suitable cloud system provider for the business.

It is worth noticed that most of cloud migrations strategies are exist or SMEs are customized based on the vendor tools. There are few migration challenges need to be tackled.

3.1 Migrations Challenges

Business owners especially startups planning to cloud migration face many challenges. First should do a complete study before deciding about migration process. There are many aspects of challenges for the selection. Some aspects are described in figure (1).

3.1.1 Business Aspect:

Interoperability and Lock-In

As cloud offerings proliferate, there will be ongoing challenges with interoperability, portability, and migration. To be sure, interoperability is also an issue for on-premises applications, but this challenge is magnified in the cloud system. In an on-premises model, enterprises control their infrastructure and platforms at any time. In the cloud, they're locked in to a provider and no longer control their own IT [7]. So, startups might be locked in a specific vendor.

Adoption to cloud computing

An organization adopting cloud computing will encounter changes to the mission, authority, funding and staffing of various departments within the organization [8]. So, it needs a good plan to minimize the risk of adoption

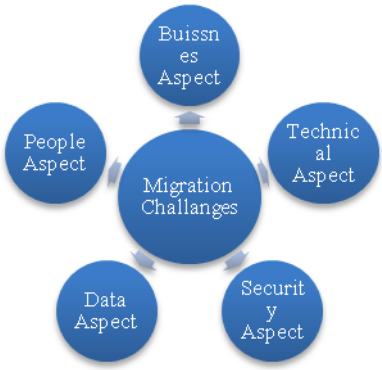


Figure (1): Proposed aspects of cloud migration challenges

Risk Analysis

Startups should do a risk analysis study regarding to the cloud-shift. If startups are moving to the public cloud, their cloud service will be shared among subscribers and the provider will not be able to offer a service customized to suit startups needs and requirements. Confirm with the provider or find the one who can customize the services for you.

Cost of migration

Most of cloud providers who avail a specific program for startups have pricing calculators. These calculators help startups to estimate the real cost of migration in comparison to traditional IT. Doing a full cost compassion will save time and cost for startups.

Provider support

It is useful for startups to find a variety of free credit cloud programs. Startups should find how much support they could gain from the selected cloud provider. So, the provider support will represent an important parameter of selection.

Lack of control

PaaS as a delivery model in cloud system gives startups a ready environment for development. This will make the development faster. However, startups should realize that their control will be loose. So, they should select the cloud provider who will be capable of giving a complete support during extending or shrinking their resources.

Support for Service Level Agreement negotiation

Cloud computing platforms should support standardized service expressions for quality of service (QoS) and service level agreement (SLA) negotiation mechanisms between consumers and providers to state the terms under which the service need to be provided. Most current IaaS platforms provide simple SLAs based on the availability of the resources, but they lack advanced SLA mechanisms based on service-oriented QoS metric. Examples are: service performance, service availability, service response time, or service cost, etc. Moreover, as different cloud providers can offer different

types of SLAs, defined with a variety of service level objectives (SLOs) and QoS metrics that makes the negotiation of similar SLAs with different cloud providers difficult. This may become an obstacle for the expansion of aggregated cloud infrastructures [9].

3.1.2 Technical Aspect:

Proof of concept (POC)

It's always a good idea to build a small POC before you actually migrate your workload to the cloud. It will give startups greater clarity and understanding about their cloud migration and the challenges they can face [10]. Startups could start with using the free model availing by the cloud provider. Where, they could develop one feature of the project testing the current environment before complete migration to the selected cloud provider.

Integration

The technical issues are complex. Most firms that migrate to the cloud do so in a hybrid model, keeping certain key elements of their infrastructure under their direct control, while outsourcing less sensitive or core components. Integrating internal and external infrastructures can be a technical quagmire.

Internet connectivity

Internet connectivity is a challenge facing startups and business owners in developing countries Like Egypt. Egypt was ranked among the 15 slowest countries for Internet speed according to the NET INDEX interactive map, ranking 184 among the 200 countries included on the list [11]. So, an existing good internet connectivity is very important for startups working in cloud platforms and are willing for cloud migration.

3.1.3 Data Aspect:

If the application stores and retrieves very sensitive data, you might not be able to maintain it in the cloud. This will limit the choice towards cloud migration. In addition, startups will ask about the data center location. Startups should investigate what will happen if the data has been lost. They should find out whether there will be backup and restore plan or not.

3.1.4 Security Aspect:

It is clear that the security issue has played the most important role in hindering coud computing acceptance. Without doubt, putting your data, running your software on someone else's hard disk using someone else's CPU appears daunting to many [12]. While there are many security concerns in the cloud, [13] published a report which focuses on specifically related to the shared, on-demand nature of cloud computing.

3.1.5 People Aspect:

Expertise

The organization needs to possess, or to be able to acquire, the skills necessary to prepare and migrate the application components. The preparation of virtual machine images and their deployment to a cloud service may involve skills new to the organization, and this must be considered in the project plan [14].

3.2 Related Work

There are many approaches that can now serve an enterprise to develop migration strategy to cloud. IBM presented an approach to capture the potential of cloud [15]. IBM approach needs more analysis and strategic thinking, and this approach is suitable for Small & Medium-sized Businesses (SMBs). Amazon also proposed an approach for cloud migration strategy. The model name is AWS cloud adoption framework (CAF). CAF is a guide to help with adopting AWS cloud-based services [16]. Not only IBM and Amazon put an approach for cloud migration. [17] Presents an approach built by Amazon for cloud migration. But, this model is a waterfall model, which is a one way model. It is a phase driven model approach to cloud migration. There are many approached for other vendors. Every vendor launched an approach which mapped with its platform features. In addition, there are free cloud programs for startups. That is why the startup faces a dilemma in selecting the best provider.

4. THE APPROACH

As startups look for a cost control option. They have always looked forward for paying less. There are many cloud programs for startups. From which they could select among them. As we described before the problem is answering this question what is the best cloud provider that suits individual startups demand. The proposed model for cloud migration is based on agile method. Agile method deals with unstable and volatile requirements by using a number of techniques. The most notable are: 1) simple planning, 2) short iteration, 3) earlier release, and 4) frequent customer feedback. The agile approach turns the traditional software process sideways. Based on short releases, agile methods go through all development stages a little at a time, throughout their software development life cycle. In one agile release, the steps may not be clearly separated, as they are in a waterfall development model. But the requirements recognition, planning, implementation and integration sequences are the same as in waterfall model [18]. There are many agile methods as Extreme Programming (Kent Beck, Ward Cunningham, and Ron Jeffries), Scrum (Jeff Sutherland and Ken Schwaber), Crystal Methods (Alistair Cockburn), Feature Driven Development (Jeff DeLuca) and Dynamic Systems Development Method (DSDM Consortium) [19]. Dynamic System Development Method is an agile software development approach that provide a framework for building and maintaining a system that meets tight time constrain through the use of incremental prototyping in a controlled project management. This method is particularly useful for the systems to be developed in short time span and where the requirements cannot be frozen at the start of the application building. Whatever requirements are known at a time, design for them is prepared and design is developed and incorporated into system [20].

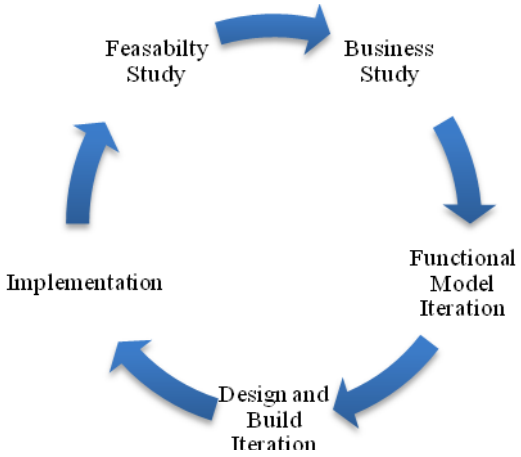


Figure (2): DSDM approach lifecycle.

Dynamic Systems Development Methods is an iterative and incremental process. Urgent business needs can be addressed early while less immediately important functionality to be addressed later. The iterative nature of DSDM enables users to see work under construction, comment on it and request changes during its further the development [21]. Startups face challenges as requirements changes, customer turnover and technology innovation. So, DSDM agile model will help startups to overcome these challenges and select the best cloud provider. As explained below in the figure (2).

DSDM includes 3 main processes. These processes are included in five phases:

- *Pre-project process*, which provides the necessary resources for starting the project, along with preliminary planning.
- *Project- proper*, during which the five main phases of the DSDM are applied; the first two sequentially at the start of the project, and the remaining three as interwoven cycles: as interwoven cycles:
    - *Sequential Phases*: studying the business domain and performing a preliminary analysis of the system: 1) Feasibility Study, 2) Business Study,
    - *Iterative Phases* (The Development Cycle): iterative and incremental analysis, design, coding and deployment of the system through evolution prototyping: 1) Functional Model Iteration, 2) Design-and-Build Iteration. 3) Implementation.
- *Post-project, where* system maintenance through further iterations of the main phases.

The DSDM five phases provides an adaptive approach which is able to respond to the changing requirements of the clients. These phases are described below in figure (2):

4.1 Feasibility study Phase

In this phase, startups should start to do a study about the cloud program availed by cloud providers and comparing between providers guided by aspects of migration challenges as described before. In this phase, the startup would select the cloud provider based on the study results. Startups could do this phase again according to the technology developments. Because, this phase is an iterative one. The deliverables from this stage are the Feasibility Report and the Feasibility Prototype that address the feasibility of the selected cloud provider. In addition, startups should do a simple cost and risk analysis. This phase can be repeated again, as the requirements of the business are not completely clear.

4.2 Business study Phase

In business study phase, startups identify the processes and information entities in their business domain, in addition to determining the main users in the systems. One of the main inputs of this phase is the selected cloud program as an early specification of the physical architecture. In addition, startups should do risk assessment for the selected vendor, provide a firm basis for technical development to proceed.[22] described the key products as follows: 1) Business Area Definition, 2) Prioritized Requirements List, 3) Development Plan, 4) System Architecture Definition, 5) Updated Risk Log.

4.3 Functional Model Iteration Phase

This phase is the first iterative phase, where it may be repeated more than one time. Startups could do a complete risk analysis of the selected cloud program. This analysis will be refined during the iterations based on the prototype. This will help startups to follow the right directions to either continue with the selected provider or reselect another cloud provider.

4.4 Design-and-Build Iteration Phase

Design and build iteration phase includes the refinining prototypes of the product. Testing process is part of this phase, where startups could test the cloud program from all prospectives. where, sartups could validate the design.

4.5 Implementation Phase

In this phase, the customer's role appears. Or could be renamed by time to the product launch phase. Customers could access the final product and they put their modification, if exists. The products of this phase include the Delivered System that contains all the agreed points.

Post-project process:

The Post-Project phase contains the activities that occur once the project team has disbanded. These include support and maintenance activities and (optionally) a Post-Implementation Review to assess the system in use. The objectives are to keep the Delivered System operational; assess whether or not the proposed benefits of the project as stated during its initial phases have been achieved; enable development processes to improve; and review the Delivered System in use [23]. It includes the steps in case of the startup decided to execute the exit plan.

5. RESEARCH METHODOLOGY

A qualitative research using case studies is used to identify how bootstrapped startups could select the best cloud program using one of agile methodologies. Qualitative method has been adopted, as this research is exploratory in nature. The case study method provides the opportunity to ask penetrating questions and to capture the participant's rich experiences and thoughts. A total of 18 cases studies reflecting 18 different startups were carefully selected and semi-structured interviews with startups owners were carried out.

6. ANALYSIS

According to theoretical analysis of the 18 startups going to cloud, there have been serious indications that cloud computing has high positive impact on entrepreneurs and startups in Egypt. Due to the current situation of Egypt's economy, startups could play an excellent role in economic prosperity. Using cloud computing as enabling technology startups would focus on business more than technical. Besides that, startups gain benefits from cloud, in addition to cost reduction, flexibility, etc. It minimizes the risk on startups, where startups is able to decide the best cloud provider with investing less time and money. They could decide that after the feasibility phase or after the implementation the selected cloud provider. Because, the startups know end result of the entire project. So, startups have a better opportunity with cloud computing. Using the standard DSDM approach, startups will avoid the pain of selecting the best-of-breed cloud program for their business case. Although there are many cloud providers which give startups a special free credit cloud program, but what matters is how startups could avoid the risk of discontinuity with this vendor from all migration challenges aspects. DSDM approach is an iterative approach which could be applied easily on different free credit cloud programs. And, it will give startups an opportunity to generate a pilot free cloud program platform easily and quickly with low risk in time and cost. In addition, it will give startups a practical result of the best vendor.

7. CONCLUSIONS AND FUTURE WORK

A DSDM as an agile approach to Cloud migration is required to move applications to the Cloud while navigating through various minefields. This would enable organizations to ensure that the costs of migration to the free cloud programs in the long term after finalizing the free duration do not exceed its benefits.

In the future work, DSDM approach with selected startups planning to cloud migration will be applied. This plan will convert DSDM approach from theoretical to practical area for cloud migration. Also, there are limitations in applying this approach as there are just a limited number of participating startups. In addition, one of the main disadvantages of DSDM approach for startups is multiple team roles. Customizing the team structure to a limited number will helps startups to find capable team members. Because, it requires users and developers both to be trained to employ it effectively. It is a relatively a new model. So, it is not very common and an easy to understand. To be able to approve this approach, a tracing plan is needed to see the startup progress and how this approach would effect on the startups performance.

REFERENCES

[1] Peter Mell Timothy Grance. : The NIST Definition of Cloud Computing, (September 2011), P3.

[2] Edlund, Åke, Livenson, Ilja. : Cloud Computing in Startups, (2011).

[3] Kshetri, Nir. : Cloud Computing in Developing Economies, IEEE Computer, (2010), pp. 47-55.

[4] Dimitrios Gkikas. : The Impact of Cloud Computing on Entrepreneurship and Start-ups: Case of Greece, (2014).

[5] Samuel Greengard. : Cloud Computing and Developing Nations, ACM 0001-0782/10/0500, (2010).

[6] Gaurav Rastogi, George Eby Mathew, Kochikar V P, Raj Joshi, Rajiv Narvekar, Ranganath, Subu Goparaju. : Cloud Computing: Pinnacle of IT infrastructure Democratization, (2009).

[7] Paul Hofmann, Dan Woods. : Cloud Computing: The Limits of Public Clouds for Business Applications.

[8] Shrikant D. Bhopale. : Cloud Migration Benefits and Its Challenges Issue, IOSR Journal of Computer Engineering (IOSR-JCE) ISSN: 2278-0661, ISBN: 2278-8727, PP: 40-45.

[9] Rafael Moreno-Vozmediano, Rubén S. Montero, and Ignacio M. Llorente. : Key Challenges in Cloud Computing to Enable the Future Internet of Services, (2011), IEEE.

[10] Vineet Badola. : Cloud migration: benefits and risks of migrating to the Cloud, (October 2015).

[11] NETINDEX. : Working for a better internet, annual report, (2014).

[12] Kuyoro S. O, Ibikunle F and Awodele O.: Cloud Computing Security Issues and Challenges, International Journal of Computer Networks (IJCN), Volume (3): Issue (5), (2011).

[13] Cloud Security Alliance. : Cloud Computing Top Threats in 2016, (February 2016).

[14]Cloud Standards Customer Council. : Migrating Applications to Public Cloud Services: Roadmap for Success, (December, 2013).

[15] IBM Global business services. : Capturing the potential of cloud: https://www.ibm.com/ibm/files/K640311W72867H78/12Capturing_the_Potential_of_Cloud_1_5MB.pdf.

[16] Miha Kralj, Pervez Kazmi and Andy Ruth. : An Overview of the AWS Cloud Adoption Framework, (February 2015), P25.

[17] Jinesh Varia. : Migrating your Existing Applications to the AWS Cloud - A Phase-driven Approach to Cloud Migration, (October 2010), P4.

[18] Ming Huo, June Verner, Liming Zhu, Muhammad Ali Babar. : Software Quality and Agile Methods, Proceedings of the 28th Annual International Computer Software and Applications Conference, (2004), IEEE.

[19] Sanjiv Augustine. : Introduction to Agile Methods: http://www.slideshare.net/Samuel90/introduction-to-agile-project-management, (July, 2006).

[20] DSDM Consortium. : DSDM: Enabling Business Agility" http://intra.iam.hva.nl/content/0708/propedeuse/ontwikkelmethoden_en_techniken/intro-en-materiaal/DSDM.pdf, (2002-2008).

[21] Umer Siddique. : Dynamic System Development Method: http://dsdm-model.blogspot.com.eg/2012/05/dynamic-system-development-method.html, (May 2012).

[22] J. Stapleton. : DSDM Consortium. DSDM: Business Focused DSDM: Business Focused Development, 2nd Ed., AddisonWesley, (2003)

[23] Open Group. Available:https://www.opengroup.org/architecture/wp/dsdm/DSDM_Framework_and_TOGAF.pdf.

Integrated Algorithms for Network Optimization of Dangerous Goods Transportation

Haixing Wang[1], Guiping Xiao[1], Tao Hai[1]
[1] Beijing Jiaotong University,Shangyuancun 3, Xizhimenwai, Haidian district, Beijing, China, 100044
Haixing21@163.com

Abstract. Network Optimization of Dangerous Goods Transportation (NODGT) belongs to a class of problems referred to as NP-Hard, and strict constraint of it and then make it harder to solve. Considering the capacity of the road network and the maximum expected risk limits, a network optimization model to minimize the total cost was established based on the network flow theory. Improvement on flow distribution model was adopted on the basis of former algorithm. In order to dealing with NODGT, improved algorithms were devised. Furthermore, the problem had been discussed with integrated use of immune cloning algorithm and LINGO. An example was analyzed to demonstrate the correctness of the application. It was proved that improved algorithms were efficient and feasible in solving NODGT.

Keywords: LINGO Algorithm, Simulated annealing Algorithm, Network Optimization of Dangerous Goods Transportation

1 Introduction

Dangerous goods, which have different physical and chemical properties, have high risk during transportation, but a series of problems may arise in this process network optimization is a complex combinatorial optimization problem, which is a typical NP-complete problem and difficult to come up with a direct answer. It is a practical problem in urgent need of solution that how we can find the optimal plan under the restrictions quickly, accurately, safely and economically.

Network Optimization of Dangerous Goods Transportation (NODGT) is to determine plans which can make a minimum of risk and safety under the premise of restrictions of the placements. Both the risk model and the flow model should be considered. That is to seek a safe routing plan which could make the cost least with the conditions have been known.

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_79

2. Algorithm based on LINGO

The above model considers the safety factor of the arc, which means each path should meet the limitation safety factor. So the alternative path set was determined first, and then the flow distribution was carried out.

The k shortest path between OD was selected as the alternative path set. The most commonly used algorithm for solving K shortest path problem should be Yen algorithm, Deviation algorithm and Double-sweep algorithm. Double-sweep algorithm was selected after comparison. The safety factor represents weight in each arc.

The above model belongs to a kind of multi-commodity network flow problem with a particular constraint, so you can use classical precise algorithms such as column generation, Danzig-Wolfe decomposition method to solve it However, the precise algorithm in solving large-scale problems often consume a long time, and it is not practical. So the algorithms based on simulated annealing and LINGO was given to solve the model.

LINGO, namely "linear interactive and general optimizer", is an effective tool for solving linear, nonlinear, integer programming, two programming and linear, nonlinear equations and other optimization problems. LINGO language model expresses simple. LINGO can express linear, nonlinear and integer programming problem by understandable mathematical formulas, require no special programming language, and the program is easy to read and modify. The input and output of date select easily. To the solving model LINGO can directly read the relevant data from the other database or spreadsheet, save a large amount of data input time and save manpower. At the same time, the data initialization part and the other part is separated, for the same model, when calculate with different data, only need to change the objective function and constraint conditions.

Steps of the algorithm:

Step 1 get the k shortest path set based on Double-sweep algorithm;

Step 2 delete some paths from k shortest path set according to the limit of safety factor to each arc;

Step 3 input the alternative path set to LINGO and get the result of the flow distribution. Associated matrix was adopted to represent the relationship between flow, path and arc, and all paths have uniform numbers.

3. Simulated annealing algorithm based on cloud computing

An advantage of algorithm based on LINGO is that the optimal solution could be obtained in theory, however, the computation time is too long when the cost function is a nonlinear function, or in a large-scale network, and sometimes even could not get the results. In view of this, the following simulated annealing (SA) algorithm based on cloud computing was also given to improve the algorithm.

Simulated annealing is a probabilistic technique for approximating the global optimum of a given function. Specifically, it is a metaheuristic to approximate global

optimization in a large search space. It is often used when the search space is discrete (e.g., all tours that visit a given set of cities). For problems where finding the precise global optimum is less important than finding an acceptable local optimum in a fixed amount of time, simulated annealing may be preferable to alternatives such as brute-force search or gradient descent.

Simulated annealing interprets slow cooling as a slow decrease in the probability of accepting worse solutions as it explores the solution space. Accepting worse solutions is a fundamental property of metaheuristics because it allows for a more extensive search for the optimal solution.

Cloud computing is an emerging field of computing where a set of resources (i.e. hardware and software resources) are accessible as a service to the user but not as a product. Cloud computing is the result of the evolution and adoption of existing technologies and paradigms. The goal of cloud computing is to allow users to take benefit from all of these technologies, without the need for deep knowledge about or expertise with each one of them. The cloud aims to cut costs, and helps the users focus on their core business instead of being impeded by IT obstacles. Cloud Computing refers to the use of computing, platform, software, as a service. It's a form of utility computing where the customer need not own the necessary infrastructure and pay for only what they use. Computing resources are delivered as virtual machines.

The best part about this computing paradigm is, user need not to be worried about the physical resource locations and number of available instances of any resources. Cloud computing is efficient because it provide multilevel abstraction and a series of virtualization layers by which it becomes a more resourceful network-based resource computing. Broadly, we can say that there are three types of services of cloud computing:

(1) platform as a service ;

(2) software as a service;

(3) infrastructure as a service.

There are some good cloud offerings for such type of services i.e. Google, Amazon web services, GoGrid, etc

In such condition, task scheduling algorithms play an important role where the aim is to schedule the tasks effectively so as to reduce the turnaround time and improve resource utilization. Task scheduling has become one of the major key areas for research. There are a number of soft computing techniques available for solving complex problems. After dealing with soft computing techniques to solve various problems in task scheduling in cloud computing environment, simulated annealing algorithm for scheduling tasks taking into consideration their computational complexity and computing capacity of processing elements. Cloud tool is used for experimentation. Experimental results show that proposed algorithms exhibit good performance under heavy loads. Cloud computing resource scheduling management model was shown in figure 3.

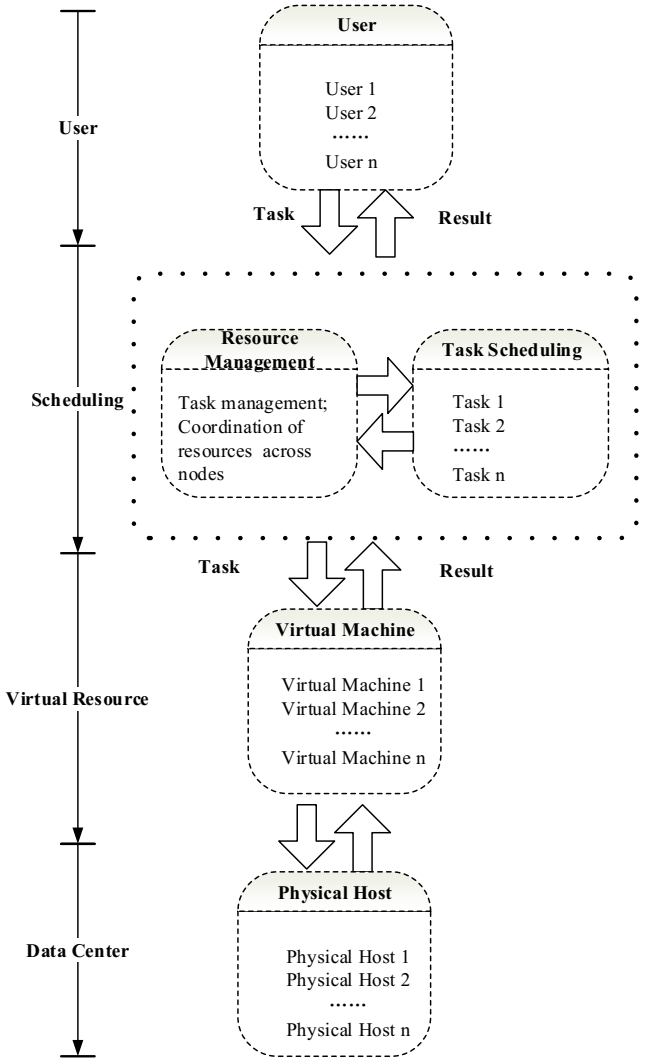


Figure 1 Cloud computing resource scheduling management model

Steps of the algorithm:

Step 1 select a path randomly to generate an initial feasible solution x and calculate the objective function value $\overline{f}_x$, set the initial temperature as t_0, set the temperature fall times as $l = 0$;

Step 2 Repeat these steps W times

Search for feasible solutions $\tilde{x}$ within $N(x)$(the new neighborhood of x),set $\Delta f = \overline{f}_{\tilde{x}} - \overline{f}_x$,if $\Delta f \le 0$,then $x = \tilde{x}$;if $\Delta f > 0$,when $\exp(-\Delta f / t_l) > random(0,1)$ (It represents a uniform random number between 0-1), $x = \tilde{x}$;

Step 3 If the algorithm satisfies the termination rule, quit; otherwise, go back to proceed the temperature operation $t_{l+1} = \alpha \square t_l$, $l = l + 1$,back to Step 2.

4. Example analysis

C ++ language was adopted to achieve the above two algorithms based on the example in the following figure.

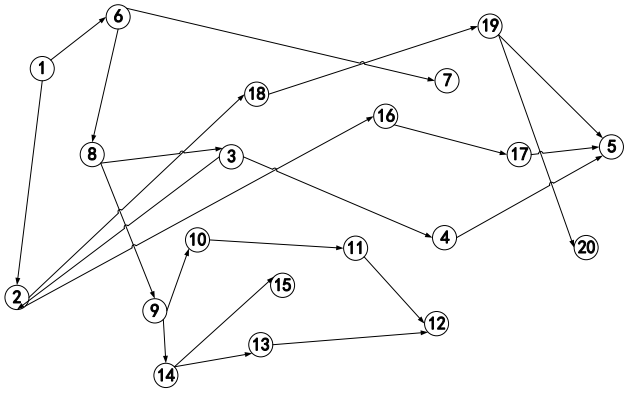


Fig. 2. Example for the algorithms.

The relevant parameters of each arc was shown in the following table:

Table 1. Relevant parameters of each arc.

No.	O	D	Unit transportation costs	capacity	Safety factor
1	1	6	6	28	0.5
2	6	7	3	25	0.6
3	1	2	4	27	0.7
4	8	3	3	25	0.4
5	8	9	7	20	0.4
6	9	10	6	25	0.4
7	10	11	5	25	0.5
8	11	12	6	25	0.7
9	13	12	4	25	0.8
10	9	14	4	25	0.6
11	14	13	4	20	0.4
12	14	18	5	25	0.2
13	2	18	4	25	0.4

14	3	2	5	25	0.3
15	3	4	3	25	0.6
16	16	17	4	20	0.7
17	2	16	5	30	0.6
18	18	19	4	25	0.6
19	14	15	5	25	0.7
20	19	20	4	25	0.4
21	19	5	5	30	0.4
22	17	5	5	25	0.7
23	4	5	3	25	0.7
24	6	8	5	30	0.7
25	12	19	4	30	0.7

Flow between nodes and their associated parameters were shown in the following table:

Table 2. Flow between nodes and their associated parameters.

No.	O	D	Flow	Safety factor Limitation
1	1	7	10	0.5
2	1	2	20	0.3
3	8	10	10	0.4
4	8	12	10	0.3
5	1	3	10	0.4
6	11	20	10	0.3
7	9	19	12	0.4
8	2	5	10	0.3
9	8	5	10	0.2

Alternative path set could be get use the algorithm based on LINGO;

Table 3. Alternative path set based on LINGO.

No.	O	D	Alternative path
1	1	7	1→6→7
2	1	2	1→2; 1→6→8→3→2
3	8	10	8→9→10
4	8	12	8→9→10→11→12; 8→9→14→13→12
5	1	3	1→6→8→3
6	11	20	11→12→19→20
7	9	19	9→14→18→19
8	2	5	2→18→19→5; 2→16→17→5
9	8	5	8→3→4→5; 8→3→2→18→19→5; 8→9→14→18→19→5

Alternative path set could be get use the algorithm based on Simulated annealing;

Table 4. Alternative path set based on LINGO.

No.	O	D	Flow	Alternative path
1	1	7	10	1→6→7
2	1	2	20	1→2
3	8	10	10	8→9→10
4	8	12	10	8→9→14→13→12
5	1	3	10	1→6→8→3
6	11	20	10	11→12→19→20
7	9	19	12	9→14→18→19
8	2	5	10	2→18→19→5
9	8	5	10	8→3→4→5

In addition to the above example, a number of large-scale examples are also tested, the information of these examples is shown in the following table:

Table 5 Information of large-scale examples.

No.	Notes	Arcs	Flows
1	20	30	15
2	25	40	15
3	25	50	10
4	30	50	20
5	50	70	25
6	60	100	30

In the above example, all the computing time of simulated annealing algorithm to get alternative path set was within 2 minutes, the computing time of LINGO algorithm was less than 15 minutes, which show algorithm 2 fit the demands of the large-scale examples. The algorithm Comparison result was shown in the following table.

Table 6. Algorithm Comparison.

No.	Optimal objective value		deviation
	Algorithm 1	Algorithm 2	
1	15673	16272	3.8%
2	15948	15948	0
3	16229	17245	6.3%
4	16448	17140	4.2%
5	18392	18864	2.6%
6	18994	19594	3.2%

We can see that the optimal objective function value gap between two algorithms was within a tolerable range, thus satisfying the needs of practical large scale problems.

5. Conclusion

Flow distribution model was presented to solve NODGT at first, then the integrated simulated annealing algorithm and LINGO algorithm were adopted, And also the two algorithms were tested and compared combined with examples. Through a series of comparison and analysis, it shows that the optimal objective function value gap between two algorithms was within a tolerable range, thus satisfying the needs of practical large scale problems.

Acknowledgment

The authors thank all the reviewers and editors for their valuable comments and works. This paper is supported by the Supported by the Fundamental Research Funds for the Central Universities (No. 2016JBM027) and Natural Science Foundation of China (NSFC) (No. 51408237).

Reference

1. Lucio Bianco, Massimiliano Caramia, Stefano Giordani. A bilevel flow model for hazmat transportation network design [J]. Transportation Research: Part C, 2009, 12(2): 175-196
2. Erhan Erkut, Osman Alp. Designing a road network for dangerous goods shipments [J]. Computers & Operations Research, 2007, 34 (5): 1389-1405
3. Liu Yu. The model of urban transport path selection for dangerous goods based on ArcGIS. The thesis of Jilin University master's degree thesis. 2007
4. Wang Yin Gao. Risk assessment of road dangerous goods transport. Thesis of Chang'an University master's degree thesis. 2006
5. Rink K A, Rodin E Y, Pandian V Sundara. A Simplification of the Double-sweep Algorithm to Solve the K-shortest Path Problem [J], Applied Math. Lett.,2000,13(8): 77-83
6. Sayarshad H R, Ghoseiri K, A Simulated Annealing Approach for the Multi-peoriodic Rail-car. Fleet Sizing Problem [J], Computer & Operation Research, 1983, 220: 671--80.
7. Qingni Shen; Yahui Yang; Zhonghai Wu; Dandan Wang; Min Long. Securing data services: a security architecture design for private storage cloud based on HDFS, International Journal of Grid and Utility Computing, 2013 Vol. 4 No. 4 P 242-254.
8. Shu Guo; Haixia Xu. A secure delegation scheme of large polynomial computation in multi-party cloud, International Journal of Grid and Utility Computing, 2015 Vol. 6 No. 1 P 1-7.
9. Shuaishuai Zhu; Xiaoyuan Yang. Protecting data in cloud environment with attribute-based encryption, International Journal of Grid and Utility Computing, 2015 Vol. 6 No. 2 P 91-97.
10. Madoka Yuriyama, Takayuki Kushida. Integrated cloud computing environment with IT resources and sensor devices, International Journal of Space-Based and Situated Computing, 2011 Vol. 1 No. 2/3 P 163-173.

11.Cristina Dutu; Elena Apostol; Catalin Leordeanu; Valentin Cristea. A solution for the management of multimedia sessions in hybrid clouds, International Journal of Space-Based and Situated Computing, 2014 Vol. 4 No. 2 P 77-87.
12.Tao Hai, The Theory and Key Technology of Safety Control for Hazardous Materials' Tanker Transportation in Railway, Doctoral thesis of Beijing Jiaotong University , 2010.
13.Qingni Shen; Yahui Yang; Zhonghai Wu; Dandan Wang; Min Long. Securing data services: a security architecture design for private storage cloud based on HDFS, International Journal of Grid and Utility Computing, 2013 Vol. 4 No. 4 P 242-254.
14.Shu Guo; Haixia Xu. A secure delegation scheme of large polynomial computation in multi-party cloud, International Journal of Grid and Utility Computing, 2015 Vol. 6 No. 1 P 1-7.
15.Shuaishuai Zhu; Xiaoyuan Yang. Protecting data in cloud environment with attribute-based encryption, International Journal of Grid and Utility Computing, 2015 Vol. 6 No. 2 P 91-97.
16.Madoka Yuriyama, Takayuki Kushida. Integrated cloud computing environment with IT resources and sensor devices, International Journal of Space-Based and Situated Computing, 2011 Vol. 1 No. 2/3 P 163-173.
17.Cristina Dutu; Elena Apostol; Catalin Leordeanu; Valentin Cristea. A solution for the management of multimedia sessions in hybrid clouds, International Journal of Space-Based and Situated Computing, 2014 Vol. 4 No. 2 P 77-87.

Study on Dynamic Monitoring System for Railway Gaseous Tank Cars Transportation

Haixing Wang[1], Jujian Wang[1], Chao Li[1]

[1] Beijing Jiaotong University,Shangyuancun 3, Xizhimenwai, Haidian district, Beijing, China, 100044

Haixing21@163.com

Abstract. Transport of dangerous goods accidents are caused mainly by changes in the container environment during transportation (such as temperature, humidity, pressure, etc.) or a mixture of goods caused by a chemical reaction and lead to combustion, explosion, toxic gas leaks and so on. Therefore, it has great significance of improving the security of maritime container transport of dangerous goods if the real-time state can be monitored during whole the transportation of dangerous goods. This paper presents dangerous goods tanker monitoring systems in railway and discusses the system's architecture and key technologies. In addition, real-time information for the state of dangerous goods will be communicated to the system center, thus optimally combining safety with transportation efficiency enhancement.

Key words：Dangerous goods Transportation, Dynamic Monitoring System, Railway Gaseous Tank Cars

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_80

1 Overview

Special protection shall be given to dangerous goods as they are combustible, explosive, poisonous, corrosive and radiative that they are liable to result in the personal injury and property damage. Moreover, the accident of hazardous material may lead to very serious consequence if not controlled or rescued in time.

Railway tank car is the main vehicle for the railway transportation of dangerous goods, transporting over 85% of dangerous goods. Besides, tank car usually transports liquid or gaseous dangerous goods, which are especially hazardous and are quite combustible, explosive and poisonous that it is necessary to focus on the monitoring to control the safety.

Dynamic monitoring technology is the core technology for the safety control of railway transportation of dangerous goods. As dangerous goods transported by the tank car of gaseous dangerous goods are major hazardous sources, it is quite necessary to handle the operation status and operation position in the railway transportation to monitor the mobile hazardous source effectively to prevent any accident. This paper studies the dynamic monitoring system of railway tank cars of gaseous dangerous goods and utilizes the modern sensing technology, monitoring technology and information technology to realize the dynamic management of transportation of dangerous goods, and researches the overall frame, core modules and key technologies of the system.

2 System Structure

2.1 Physical Structure

Physical structure of the system is demonstrated in Figure 1:

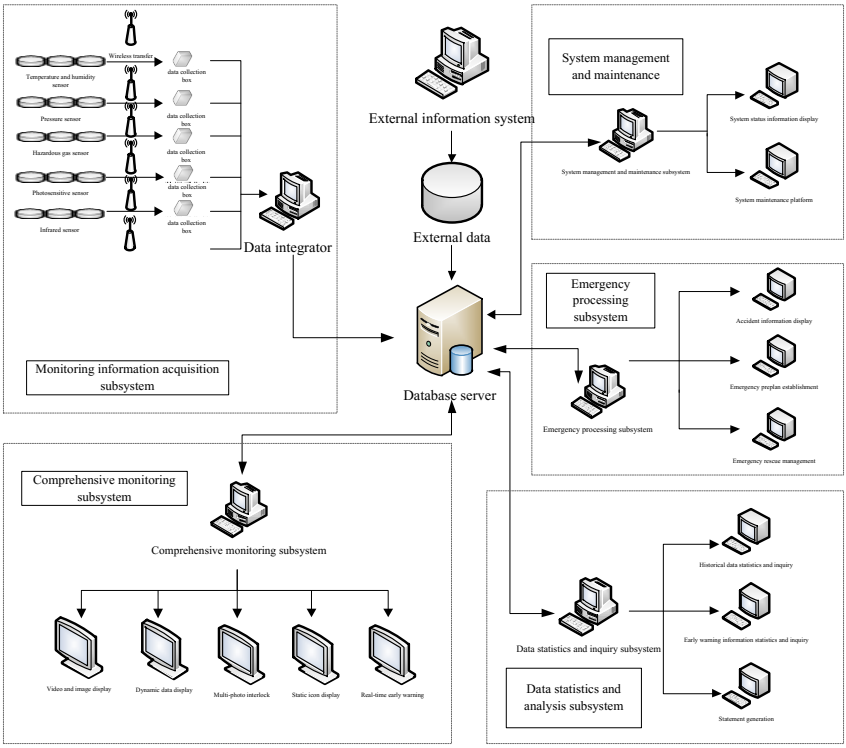


Fig. 1. Physical Structure of System

2.2 Operation Flow

Operation flow of the system is demonstrated in Figure 2:

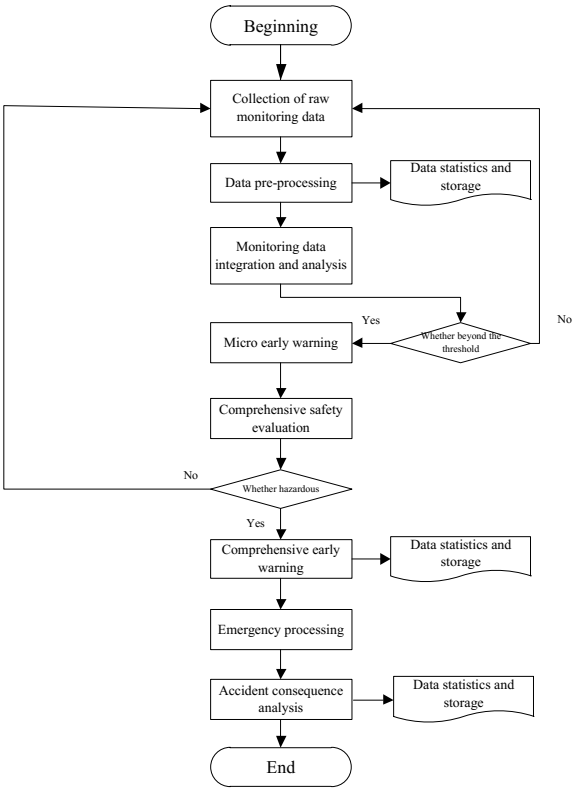


Fig. 2. Operation Flow of System

3 Core Functions

Core subsystems of the system are demonstrated in Figure 3:

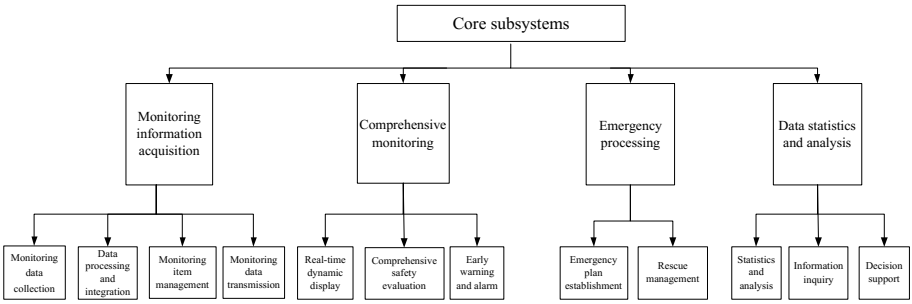


Fig. 3. Functional Structure of System

3.1 Monitoring Information Acquisition Subsystem

(1) Monitoring data collection module

Monitoring data collection module mainly collects the raw data of the monitored object by the monitoring device. Monitoring devices of the tank car of gaseous dangerous goods mainly include the temperature and humidity sensor, pressure (hydraulic pressure) sensor, hazardous gas sensor, infrared sensor, photosensitive sensor and smoke sensor, etc.

(2) Monitoring data processing module

Monitoring data collected in a real time manner includes temperature, humidity, pressure and density of hazardous gas. Such data is transmitted to the data processing and integration module in a real time manner for the module to conduct the integration and analysis of the monitoring data collected by the sensors.

(3) Monitoring item management module

Monitoring item management module manages different monitoring items according to different types of materials to extract the monitoring information selectively. Different items correspond to different sensors. Each sensor depends on an algorithm, which is under the administration by the comprehensive monitoring subsystem. Whenever the monitoring item changes, the sensor will be added, removed or replaced correspondingly. The monitor may carry out the matching, adding, deletion or revision of the corresponding algorithm of sensor through the module, and may enable the inquiry of the algorithm.

(4) Monitoring data transmission module

Monitoring data transmission module mainly transmits various kinds of monitoring data, including temperature, humidity, pressure and density of hazardous gas. Relevant data may be transmitted by the wireless sensor network.

3.2 Comprehensive Monitoring Subsystem

(1) Real-time dynamic display

Real-time dynamic display module mainly displays the dynamic change of temperature, humidity, pressure (hydraulic pressure) and density of hazardous gas in a

real time manner. Real-time monitoring data from the database may be processed to form the graph of monitoring information, which may display dynamically each monitoring data in the form of curve diagram to enable the monitor to judge whether the monitoring data is beyond the safe scope or safe threshold, or the extent of deviation.

Real-time dynamic display flow is demonstrated in Figure 4:

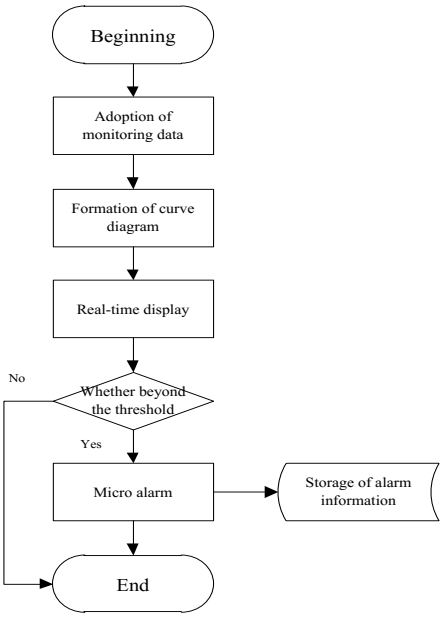


Fig. 4. Flow of Real-time Dynamic Display Module

(2) Comprehensive safety evaluation module

Comprehensive safety evaluation module mainly utilizes the operation safety evaluation system to comprehensively analyze the monitoring information, material status information, comprehensive management and regulation information, and technical condition information of vehicles and equipment, and adjusts the threshold of indexes according to the actual situation and relevant information, so as to comprehensively evaluate the safety status and determine the accident level correctly.

Flow of comprehensive safety evaluation module is demonstrated in Figure 5:

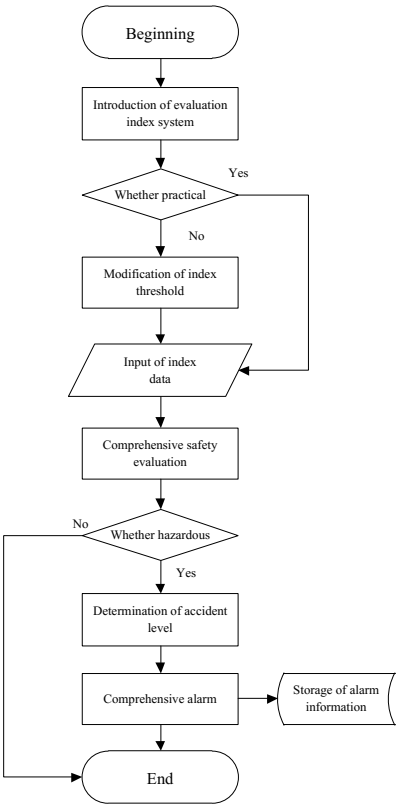


Fig. 5. Flow of Comprehensive Safety Evaluation Module

(3) Early warning and alarm module

This module mainly consists of micro early warning and comprehensive early warning. Micro early warning refers to the single-value early warning signal sent when the video monitoring recognizes any abnormal situation or the sensor detects any data beyond the threshold. It alarms by sound and dialog box. The dialog box mainly integrates with the real-time display module to provide the specific alarm information, such as the number of alarm device and extent of deviation from the threshold, through which the monitor may be aware of the alarm position and detailed information of relevant hazards. Comprehensive early warning refers to the comprehensive early warning signal sent according to the comprehensive safety evaluation and determined accident level completed in the comprehensive safety evaluation module. It alarms by sound and indicator lamp according to the hazard,

and may give different sounds or different colors of indicator lamp according to the level of hazard.

3.3 Emergency Processing Subsystem

(1) Emergency plan establishment module

Emergency plan establishment module may provide the emergency response preplan and establish the rescue plan according to the level of hazard and actual situation in case of any hazard accident or potential hazard. The monitor may understand the development and level of accident through the module, and establish the rescue plan in consideration with the rescue order from the superior authority.

(2) Rescue management module

Rescue management module mainly manages internal and external rescue resources to carry out the uniform deployment and elevate the rescue efficiency. The monitor may interlock with the local government and sectors of medical treatment and fire protection through the module to deploy internal and external available forces and resources uniformly to prevent the scattered rescue, thus improving the effect of rescue.

Flow of emergency processing subsystem is demonstrated in Figure 6:

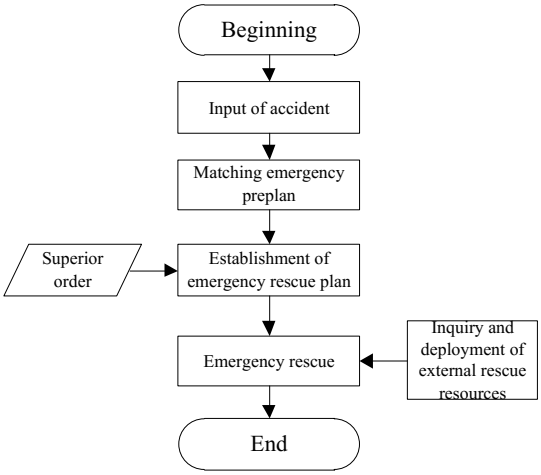


Fig. 6. Flow of Emergency Processing Subsystem

3.4 Statistics and Analysis Subsystem

(1) Information acquisition module

Statistics and analysis module mainly conducts statistics and analysis of the historical monitoring data, hardware status data, early warning data and accident data, and has them stored. The analysis of such data may obtain the information of current operation safety, equipment maintenance and accident frequency.

(2) Information inquiry module

This module provides the inquiry of the historical monitoring data, hardware status, personnel information, early warning information and accident processing information. The inquiry result is displayed in both the list and statistical chart. The information inquiry module interlocks with the complete-circuit monitoring system to improve the sharing of information.

(3) Decision support

Depending on the statistics and analysis of various analysis models, it may conduct multi-dimensional and more complicated comprehensive analysis and calculation to discover various tendencies and abnormities, obtain important details and find out internal disciplines to provide practical and effective assistance to the decision-making by the personnel in relation to the transportation.

4 Conclusion

This paper mainly introduces the concept and design frame of dynamic monitoring system of railway tank cars of gaseous dangerous goods, analyzes key technologies involved in the system and provides the design plan of the system.

Dynamic monitoring system of railway tank cars of gaseous dangerous goods may improve the safety and efficiency of the transportation of gaseous dangerous goods. It may also integrate with the existing safety management system of railway transportation to realize the full-process and integrated safety management monitoring of railway transportation.

Acknowledgment

The authors thank all the reviewers and editors for their valuable comments and works. This paper is supported by the Supported by the Fundamental Research Funds for the Central Universities (No. 2016JBM027) and Natural Science Foundation of China (NSFC) (No. 51408237).

References

1. Wang Zhe, Study on Integration, Optimization and Safety Monitoring Information System of Railway Stations for Dangerous goods, Doctoral Paper of Southwest Jiaotong University, 2010.
2. Hai Tao, Study on Theories and Key Technologies of Railway Tank Cars of Dangerous goods, Doctoral Paper of Beijing Jiaotong University, 2010.
3. Wang Yunpeng, et al. Transportation Safety Monitoring System of Dangerous goods Based on FTA, Journal of Jilin University (Engineering Version), 2010 (4).
4. Qingni Shen; Yahui Yang; Zhonghai Wu; Dandan Wang; Min Long. Securing data services: a security architecture design for private storage cloud based on HDFS, International Journal of Grid and Utility Computing, 2013 Vol. 4 No. 4 P 242-254.
5. Shu Guo; Haixia Xu. A secure delegation scheme of large polynomial computation in multi-party cloud, International Journal of Grid and Utility Computing, 2015 Vol. 6 No. 1 P 1-7.
6. Shuaishuai Zhu; Xiaoyuan Yang. Protecting data in cloud environment with attribute-based encryption, International Journal of Grid and Utility Computing, 2015 Vol. 6 No. 2 P 91-97.
7. Madoka Yuriyama, Takayuki Kushida. Integrated cloud computing environment with IT resources and sensor devices, International Journal of Space-Based and Situated Computing, 2011 Vol. 1 No. 2/3 P 163-173.
8. Cristina Dutu; Elena Apostol; Catalin Leordeanu; Valentin Cristea. A solution for the management of multimedia sessions in hybrid clouds, International Journal of Space-Based and Situated Computing, 2014 Vol. 4 No. 2 P 77-87.

Improvement of Study Logging System for Active Learning Using Smartphone

Noriyasu Yamamoto[1]
[1] Department of Information and Communication Engineerig
Fukuoka Institute of Technology
3-30-1 Wajiro-Higashi, Higashi-Ku, Fukuoka 811-0295, Japan
E-mail: nori@fit.ac.jp

Abstract. In general, it is difficult for the professors of universities to provide all needed information to satisfy the students because they have different knowledge and interests. To solve this difficulty, our previous work suggested an interactive learning process, which use the students' study records in order to improve the students' learning motivation and the self-learning time. To improve the progress speed of a lecture and the efficiency of the Self-learning, we proposed an Active Learning System (ALS) using smartphone. In addition, we performed experimental evaluation of our ALS and showed that the proposed ALS can improve student self-learning and concentration. Learning system that has been proposed so far, was only to give a feedback a record of student's learnings to lecturer. In this paper, we propose a mechanism to enhance the learning effects by logging a record of students' learnings in the whole process of learning system.

1 Introduction

In the conventional lecture style, the knowledge was transmitted from a lecturer to a student using one-way communication. Recently, in many universities the information terminals such as note PC, workstations, servers and mobile phones are used successfully during the lectures. Also, there are many e-learning systems [1-9] that use these devices. The amount of information that a lecturer transmits to the participant during the lecture can be improved by using different information terminals. However, in a lecture, if participants' motivation for learning is low, the learning results are not good. However, in universities is required to increase the learning motivation using information terminals.

Usage of the desk-top computers and notebook PCs for lecture may be inconvenient and they occupy a lot of space. Therefore, it will be better that students use small and lightweight terminals like Personal Digital Assistant (PDA) devices. Also, because of wireless networks are spread over university campuses, it is easy to connect mobile terminals to the Internet and now the students can use mobile terminals in many lecture rooms, without needing a large-scale network facility.

We considered a method of acquiring / utilizing the study record using smartphone in order to improve the students learning motivation [10]. During the lecture the

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_81

students use the smartphone for learning. The results showed that the proposed study record system has a good effect for improving students' motivation for learning.
For the professors of the university, it is difficult to offer all necessary information to the students. In addition, they cannot provide the information to satisfy all students because the quantity of knowledge of each student attending a lecture is different. Therefore, for the lectures of a higher level than intermediate level, the students should study the learning materials by themselves.

In our previous work, it was presented an interactive learning process in order to increase the students learning motivation and the self-learning time [11]. However, the progress speed of a lecture was not good. To solve this problem, we proposed an Active Learning System (ALS) for student's self-learning [12, 13]. Also, to improve student's self-learning procedure, we proposed some functions for the ALS [14]. We performed experimental evaluation of our ALS and showed that the proposed ALS can improve student self-learning and concentration [15]. Although, the self-learning time and the examination score were increased, the number of student that passed examination didn't increase significantly. To solve this problem, we proposed the group discussion procedure to perform discussion efficiently [16].

Learning system that has been proposed so far, was only to give a feedback a record of student's learnings in lecture to lecturer. In this paper, we propose a mechanism to enhance the learning effects by using a record of students' learnings in the whole process of learning system.

The paper structure is as follows. In Section 2, we introduce the related work on

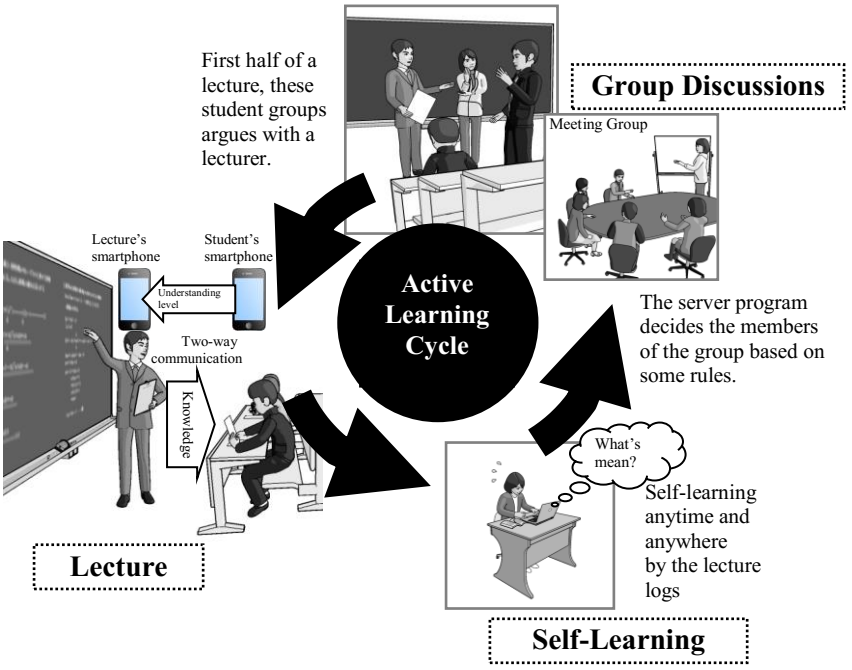


Fig. 1. ALC using a smartphone in the previous work

learning systems and present ALS. In Section 3, we indicate learning data of students. Then, in Section 4, procedure on recording and using that learning data is shown. Finally, in Section 5, we give some conclusions and future work.

2 ALS: Active Learning System

In Figure 1, we show ALC (Active Learning Cycle) of the proposed ALS. The student and the lecturer confirm the movement by setting each cycle and information by the smartphone.

As a beginning of ALC, the lecturer performs the interactive lecture by confirming the understanding degree of the student using their smartphone in real time. Therefore, a lecturer can transmit the knowledge to the student effectively.

After the lecture is finished, the student can read the lecture log by their smartphone. The student can review the lecture content using the lecture log anywhere and anytime.

Then, the student discusses the lecture content in small groups. In the group discussion, the students discuss what they did not understand in the lecture. Then, they submit the result of the group discussion to a lecturer as the reports. The formation of the group and the schedule of the group discussion is done automatically by smartphone application.

At the beginning of the next lecture, the student groups and the lecturer carry out open discussions based on the submitted repots and solve the problems that students may have. After the open discussion, the lecturer performs the next interactive lecture.

In this ALC, by adding the group learning and the open discussion the understanding is increased and the lecturer can keep a fixed progress speed of the lecture.

Table 1. Learning data of students

Item	Content	Creator/Editor
identification number (ID)	Study Point ID	ALS/invalid (Logging System)
Title	Title of Study Point	Lecturer/Lecturer
Content	Overview of Study Point	Lecturer/Lecturer
related small examinations	Related examinations and it's results	Lecturer/invalid
understanding level	Understanding level of Study Point (5 levels)	Student/Student
lecture speed	Lecture speed for students	Student/Student

3 Learning Data of Students

Learning data of students are given in Table 1. Learning data is classified based on the “study point” set by lecturers into identification number (ID), title, content, related small examinations, understanding level (5 levels) and lecture speed. The learning system automatically create the identification number. Title, content and related small examination items are set by lecturers. The understanding level and lecture speed is recorded with the result of evaluation decided by students during the lecture.

4 Procedure on Recording and Using the Learning Data

The learning system we are proposing divides the learning process into ‘Lecture’, ‘Self-Learning’ and ‘Group Discussions’. In this section, we explain here how to record and renew the learning data of students in each process (see Fig. 2).

4.1 Procedure of the Lecture for Study Logging

In the process of ‘Lecture’, the learning data of students should be created in the following procedure.

(Flow 01) The lecturer uses his smartphone application to connect to the server.
(Flow 02) The attendance registration for a student is started.
(Flow 03) The students use smartphone application and login to a server.
(Flow 04) A server notifies the attendance situation to the lecturer's smartphone application.
(Flow 05) After that the lecturer pushes the start button to start the lecture.
(Flow 06) The "study point" shift to a student who makes questions for the exam of a selection type problem is directed to a server.
(Flow 07) Next "study point" or selection type examination is transmitted to a student. Then, the "study point" or a selection type examination is recorded.
(Flow 08) At the moment of "study point" shift, for a defined period of time the student replies the lecture difficulty and lecture speed of progress. When a selection type examination is decided, the student answers in a defined period of time.
(Flow 09) The server collects the information of (Flow 08) and sends it to the lecturer's smartphone terminal application. The total information is saved as study record in the server.
(Flow 10) The total data of lecture difficulty and lecture speed of progress that the student answered is displayed. Then when the selection type examination is set by (Flow 06), the problem answer rate is displayed.

(Flow 11) When shifting to the next "study point", it returns to (Flow 06).
(Flow 12) In order to end the lecture, the lecturer connects to the server.
(Flow 13) The end of a lecture is transmitted to student application and the recording process is ended.
(Flow 14) Logoff. The end of a lecture is displayed.

4.2 Self-learning after the Lecture

After the end of a lecture, a student can access the “Study Point” for his/her learning after the lecture. In the process of ‘Self-learning’, students conduct self-study based on the results of the level of understanding the learning points set by lecturers (5 levels) and the small examinations, and based on the outcomes of it, the learning data should be renewed.

4.3 Group Discussion

After the self-learning, the group discussion procedure for ALS is as follows. In the process of ‘group discussions’, students are to be divided into groups based on their learning data. Try to have each group to have about 50/50 of the students who understood the class and who did not.

(Group Discussion Flow 01) After the self-learning, the students make the self-evaluation by five steps and register the evaluation results in the server.

(Group Discussion Flow 02) In the next lecture, the server program decides the members of the group based on some rules. In the study group, the students who understood the study points try to teach the other students who have low level of understanding. Then the students after improving their knowledge summarize the result of the group discussion to the lecturer. When all members of the group did not understand the study points, they summarize all points that they did not understand.

[Rule 01] Based on the self-learning results, the students are divided in two groups: the student with high level of understanding and low level of understanding.
[Rule 02] One or more students are selected as students with high level of understanding in a group.
[Rule 03] The degree of understanding of the students is different.
[Rule 04] When the students have high degree of understanding the study group is not created.
[Rule 05] When the students have low degree of understanding, the group is organized with the students.

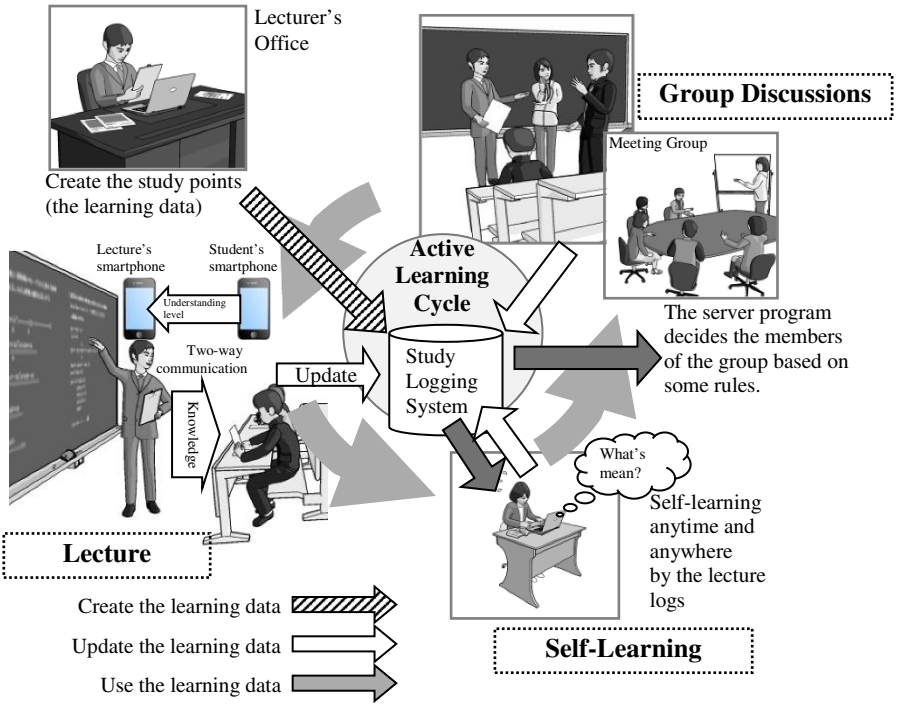


Fig. 2. Procedure on recording and using the learning data on ALS

5 Conclusions

We introduced an interactive learning process in order to increase the students learning motivation and the self-learning time. Then, we presented an active learning cycle for student's self-learning. In the proposed system, the students and the lecturer confirm the movement by setting each cycle and information by the smartphone.

Learning system that has been proposed so far, was only to give a feedback a record of student's learnings in lecture to lecturer. In this paper, we propose a mechanism to enhance the learning effects by using a record of students' learnings in the whole process of learning system. As shown in this procedure, we consider that, with our learning system, our ALS can help students learn more effectively by making use of the students' learning data.

In our future work, we will carry out extensive experiments to evaluate the proposed system for different lectures.

References

1. Underwood, J. and Szabo, A.: Academic Offences and e-Learning: Individual Propensities in Cheating. British Journal of Educational Technology, Vol. 34, Issue 4, pp.467–477 (2003)
2. Harashima, H.: Creating a Blended Learning Environment Using Moodle. The Proceedings of the 20th Annual Conference of Japan Society of Educational Technology, September, pp. 241–242 (2004)
3. Brandl, K.: Are you ready to "Moodle"?. Language Learning and Technology, Vol. 9, No. 2, pp. 16–23 (2005)
4. Dagger, D., Connor, A., Lawless, S., Walsh, E. and Wade, V. P.: Service-Oriented E-Learning Platforms: From Monolithic Systems to Flexible Services. Internet Computing, IEEE, Vol. 11, Issue 3, pp. 28–35 (2007)
5. Patcharee, B., Achmad, B., Achmad, H. T., Okawa, K. and Murai, J.: Collaborating Remote Computer Laboratory and Distance Learning Approach for Hands-on IT Education. Journal of Information and Processing, 22(1), pp. 67–74 (2013)
6. Emi, K., Okuda, S. and Kawachi, Y.: Building of an e-Learning System with Interactive Whiteboard and with Smartphones and/or Tablets through Electronic Textbooks. Information Processing Society of Japan (IPSJ), IPSJ SIG Notes 2013-CE-118(3), 1-4, 2013-02-01 (2013)
7. Yamaguchi, S., Ohnichi, Y. and Nichino, K.: An Efficient High Resolution Video Distribution System for the Lecture Using Blackboard Description. Technical Report of IEICE, 112 (190), pp. 115–119 (2013)
8. Hirayama, Y. and Hirayama, S.: An Analysis of the Two-factor Model of Learning Motivation in University Students. Bulletin of Tokyo Kasei University, 1, Cultural and Social Science 41, pp. 101–105 (2001)
9. Ichihara, M. and Arai, K.: Moderator Effects of Meta-Cognition: A Test in Math of a Motivational Model. Japanese Journal of Educational Psychology, Vol. 54, No. 2, pp.199–210 (2006)
10. Yamamoto, N. and Wakahara, T.: An Interactive Learning System Using Smartphone for Improving Students Learning Motivation. Information Technology Convergence, Lecture Notes in Electrical Engineering Volume 253, pp. 305–310 (2013)
11. Yamamoto, N.: An Interactive Learning System Using Smartphone: Improving Students' Learning Motivation and Self-Learning. Proceeding of the 9th International Conference on Broadband and Wireless Computing, Communication and Applications (BWCCA-2014), November, pp. 428-431 (2014)
12. Yamamoto, N.: An Active Learning System Using Smartphone for Improving Students Learning Concentration. Proceeding of International Conference on Advanced Information Networking and Applications Workshops (WAINA-2015), March, pp. 199-203 (2015)
13. Yamamoto, N.: An Interactive e-Learning System for Improving Students Motivation and Self-learning by Using Smartphones. Journal of Mobile Multimedia (JMM), Appril, Vol. 11, No.1&2 (2015), pp.67-75 (2015)
14. Yamamoto, N.: New Functions for an Active Learning System to Improve Students Self-learning and Concentration. Proceeding of the 18th International Conference on Network-Based Information Systems (NBIS-2015), September, pp. 573-576 (2015)
15. Yamamoto, N.: Performance Evaluation of an Active Learning System to Improve Students Self-learning and Concentration. Proceeding of the 10th International Conference on Broadband and Wireless Computing, Communication and Applications (BWCCA-2015), November, pp.497-500 (2015)
16. Yamamoto, N.: Improvement of Group Discussion System for Active Learning Using Smartphone. Proceeding of the 10th International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS-2016), July, pp.143-148 (2016)

Identifying prime customers based on mobile usage patterns

Deepali Arora and Kin Fun Li

Abstract Increasing competition has led telecommunications service providers to develop marketing strategies to retain their customers and attract new customers. Big data analytics offers some promising solutions that can aid telecommunications service providers in raising their revenues as well as designing better business strategies. In this work, we illustrate how usage behaviors of mobile phone users can be used to identify customers that generate more revenue. Additionally, telecommunications companies can also find a number of people in a customer's network and can potentially use this information for their benefits. The companies can identify the users with a large network and encourage them to serve as ambassadors to promote their services within their networks in exchange for certain incentives. For experimentation in this paper, we used the strath/nodobo dataset (v.2011-03-23) [1] comprising of customers' phone usage records. Unsupervised classification technique (clustering) is employed to find both high and low mobile usage customers based on the frequency of their calls, messages and their average call duration. Furthermore, users with a large number of outgoing calls and who are connected to many other users are also identified.

1 Introduction

The mobile phone user base is expected to grow up to 4.6 billion worldwide by the end of year 2016 [2] though most of the established telecommunications service providers are still struggling to improve their revenues or prevent customers' from

Deepali Arora
Department of Electrical and Computer Engineering, University of Victoria, P.O. Box 3055 STN CSC, Victoria, B.C., CANADA, V8W 3P6, e-mail: darora@ece.uvic.ca

Kin Fun Li
Department of Electrical and Computer Engineering, University of Victoria, P.O. Box 3055 STN CSC, Victoria, B.C., CANADA, V8W 3P6, e-mail: kinli@ece.uvic.ca

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_82

churning [3], [4]. However, the use of big data analytics by many telecommunications companies now offers some promising solutions [5], [6], [7]. A number of techniques have been recently proposed for predicting customer churn [8], [9], [10], [11], [12]. Since the customers are an important asset of a company, analyzing their mobile usage behaviors can also provide meaningful insights that could be potentially used for making better business decisions and for exploiting additional ways of monetization [13]. For example, based on the mobile usage behaviors, telecommunications companies can take the following actions:

- identify their loyal customers and offer them customized packages based on their phone usage
- derive the total number of people in a user's social network [14] and offer referral based promotions
- recognize the leaders or influential users within a user's social network and use them as brand ambassadors to promote new deals
- determine the location of a user and offer location aware services.

Furthermore, telecommunications companies can also use their data to generate additional streams of revenues by packaging and selling customers' data to other organizations for targeted mobile advertising and campaigns [15], while ensuring that the security and privacy of a user is maintained. Understanding a customer's mobile usage behavior is essential for these business activities. However, with millions of customers, analyzing each individual customer's usage pattern becomes a challenging task and sophisticated data mining techniques need to be applied.

The use of data mining techniques for various applications in the telecommunications sector has been investigated in the past decade. For example, [16] used temporal as well as last-calls features to predict the future outgoing calls following machine learning approaches. [17] used machine learning based classification models for link prediction using both call data records and location-based social networking data. [18] used data mining and machine learning algorithms for detecting fraudulent usage of cellular telephones based on customer behavior profiles. [19] developed machine learning based model to predict a user's location from the cellular data. [20] used data mining techniques to increase user engagement in mobile applications. [21] developed a framework that allowed mobile devices to continuously recognize current and anticipate future user generated context. [22] used machine learning algorithms to make predictions of mobile data traffic usage.

The use of machine learning techniques, especially clustering, for analyzing mobile usage behavior in telecommunications sector has gained traction in recent years. For example, [23] used the K-means clustering algorithm to segment users based on their usage pattern. The customers were also placed in four groups based on their loyalty assessed by their lifetime value. [14] used the K-means clustering algorithm for social network analysis of users based on their mobile phone usage. The proximity between the users was based on their registration frequencies to specific cellular towers associated with their working places. [24] also used the K-means clustering algorithm to segment users based on their mobile phone usage. [25] used clustering techniques to segment users on a mobile SMS-based chat service. [26] combined

K-means clustering and Artificial Neural Networks (ANN) technique to develop a framework for personalized mobile content delivery to users. [27] carried out clustering of mobile customers based on the call detail records and using Self Organizing Maps for revenue maximization.

Similar to earlier studies, the K-means clustering approach is used in this research to segment users based on their usage patterns. The work presented in this paper is closely related to the ones presented in [14] and [23], both of which also used incoming/outgoing calls and SMS messages for mobile user behavior analysis. However, the work presented in this paper, in addition to segmenting users based on their high or low usage, also focuses on demonstrating how mobile usage behaviors can be used for identifying closely connected user groups, and also finding influential or popular users within a group. Identifying these users within a segment is useful for telecommunications service providers as these users may potentially lead to revenue improvement with targeted marketing strategies.

This paper is organized as follows: The methodology adopted for analysis is discussed in Section 2, the K-means clustering algorithm used for analysis is introduced in Section 3 and the results are presented and discussed in Section 4, and conclusions are made in Section 5.

2 Methodology

The following steps were carried out for the analysis presented in this paper:

1. Data selection and cleaning: The strath/nodobo dataset (v.2011-03-23) was used [27] for the model presented here. The dataset comprises of mobile phone usage records collected with Nodobo suite at the University of Strathclyde. These data were collected between September 2010 and January 2011. These data comprise of call, SMS, and proximity records of 27 participants over a 5-month period. For the analysis presented here, we have used 13,035 call records and 83,542 SMS records.
2. Feature selection: For each participant or user, selected features were obtained based on the information available in the dataset. Features derived for a user include:
    a. Total incoming calls received over the 5-month period from September 2010 and January 2011.
    b. Total outgoing calls over the 5-month period.
    c. Average call duration.
    d. Total messages received.
    e. Total messages sent.
3. Forming user clusters: Once the user features are obtained the next step is to form user clusters based on usage behavior. Clustering is an unsupervised machine learning technique used for finding a structure in a collection of unlabeled data. It is therefore the process of organizing objects into groups whose members are

similar in some way according to certain criteria or features [28]. The collection of objects that are similar form a cluster. Some of the commonly used clustering models include:

a. Connectivity based models: e.g., hierarchical clustering that builds models based on distance connectivity.
b. Centroid based: e.g., the K-means algorithm that represents each cluster by a single mean vector.
c. Distribution models: clusters are formed using statistical distributions such as multivariate normal distributions, e.g., expectation-maximization algorithm.
d. Density based models: e.g., DBSCAN and OPTICS that define clusters as connected dense regions in the data space.

For the analysis presented herein, the K-means clustering technique, discussed in detail next, is used to classify users into two groups: low and high usage.

3 K-means Clustering

K-means clustering is an unsupervised learning algorithm widely used for classifying the data into K number of clusters, fixed a priori with K centers called centroids, one for each cluster. Each data point is connected to the cluster with the nearest centroid. In general, given data points that have to be partitioned into K clusters, the goal is to find the positions within the clusters that minimize the distance from the data points to the cluster centers. K-means clustering algorithm aims at minimizing an objective function known as squared error function given by [29]:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} (||x_i - v_j||)^2 \tag{1}$$

where $||x_i - v_j||$ is the Euclidean distance between x_i and v_j, c_i is the number of data points in i^{th} cluster and c is the number of cluster centers.

Let $X = x_1 \cdots, x_n$ be the set of data points and $V = v_1, \cdots, v_c$ be the set of centers. The K-means algorithm comprises of the following steps:

1. First c points are randomly chosen as cluster centers or centroids.
2. Next, the distances between every data point and centroids are computed.
3. Each data point is then assigned to the nearest cluster center. This gives the minimum Euclidean distance calculated for each data point and the nearest centroid.
4. The cluster center positions are then adjusted by using the following expression

$$v_i = \frac{1}{c_i} \sum_{j=1}^{c_i} x_j \tag{2}$$

where c_i represents the number of data points in the i^{th} cluster.

5. The process is repeated from step 2 onwards until the desired convergence is achieved when the K-means clustering algorithm is considered successful in partitioning the data into clusters around their respective centroids.

The advantages of the K-mean clustering algorithm include:

- It is fast, robust and easy to understand.
- It yields very good results for datasets that have data points which are well separated from each other.

The disadvantages of the K-mean clustering algorithm are:

- The learning algorithm requires a priori specification of the number of cluster centers.
- For highly overlapping data points, the K-means algorithm will not be able to perform efficiently.
- Choosing initial cluster centers randomly can yield very poor results.
- It is applicable only when the mean is defined, i.e., it will not work with categorical data.
- It is unable to handle noisy data and outliers.

4 Results and discussion

Figure 1 shows the clusters of users with high and low mobile usage based on the number of calls and SMS messages sent. Additionally, the duration of call was also considered for analyses but is not shown here. Figure 1 shows that using K-means clustering it is possible to identify customers with high and low mobile phone usage. Here high usage customers are considered as those users who make a lot of outgoing calls, with each call lasting for more than five minutes, and send a lot of SMS messages. Similarly, users who hardly make any calls or send any messages are considered as low usage customers.

Many telecommunications companies typically offer plans with minimum or no charge for incoming calls. The major portion of their revenue is generated mainly from outgoing calls. To find prime customers who make more outgoing calls than receiving incoming calls, the ratio of outgoing and incoming calls was derived. Figure 2 shows the bar graph of outgoing and incoming call ratio for the 27 participants. From Figure 2, it is evident that users 12, 19, and 26 are the top three users with the highest outgoing and incoming call ratios.

Although from Figure 1 and Figure 2 we can identify users that generate more revenues than others but identifying those customers alone is not enough. To improve revenues, telecommunications companies should try to identify users that are connected to a lot of other users. Figure 3 shows the bar graph of the number of people within a user's network with whom the user exchanges calls on a regular basis. Figure 3 shows that users 16, 21, 23, 24, and 25 are the top five users with the maximum number of connections. These users could potentially serve as brand

ambassadors for telecommunications companies and help them to promote their products or services to friends within their network. This can help companies in reaching to a wider audience in a fast and effective manner.

In this paper we have demonstrated by using classification techniques, how users' phone usage behavior can be used to gain business insights for telecommunications companies. For example, based on the number of calls and messages a user exchanges with his friends, a company can find its prime customers that generate more revenues than others based on usage. Telecommunications companies should adopt strategies to prevent these prime customers from churning. Additionally, telecommunications companies can also find the number of people in a user's network and therefore potentially use this information for their benefits. The idea is that if a user has a large network of friends then offering a special promotion to that user can help in spreading the message quickly. Therefore these users can serve directly or indirectly as a brand ambassador for the company. Here, direct marketing refers to the strategy in which a company approaches their prime users for promoting their products to people who participate in their social networks in exchange for some referral incentives. Indirect marketing here refers to a marketing strategy where a user unknowingly mentions to her friends about the special plans that she may have received from the company. Marketing departments of these telecommunications companies can focus on designing better marketing plans for word-of-mouth advertising. Additionally, telecommunications companies can also identify customers

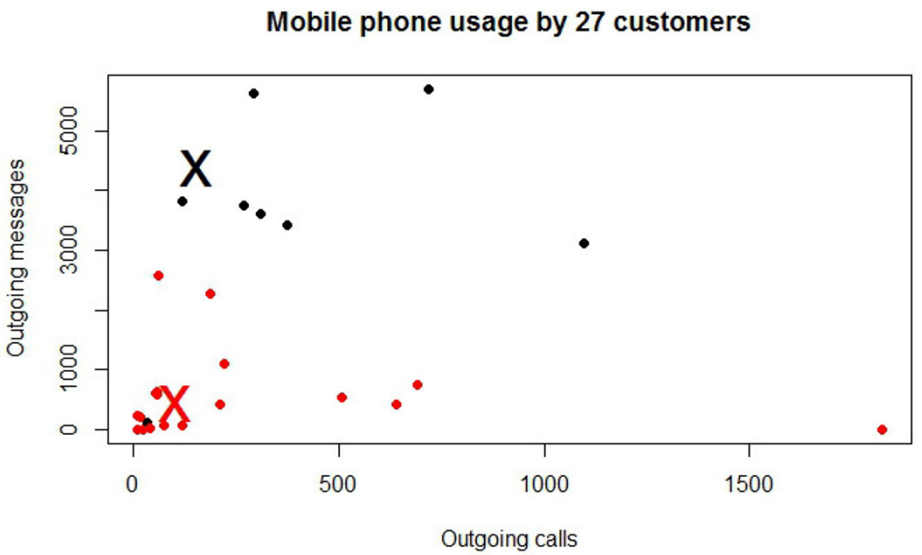


Fig. 1 Cluster demonstrating high and low mobile usage based on customers' incoming and outgoing calls and SMS messages.

who have subscribed their services but rarely use them. This may indicate that these consumers are likely to churn soon, and the company should make a decision to either retain or let go of these customers.

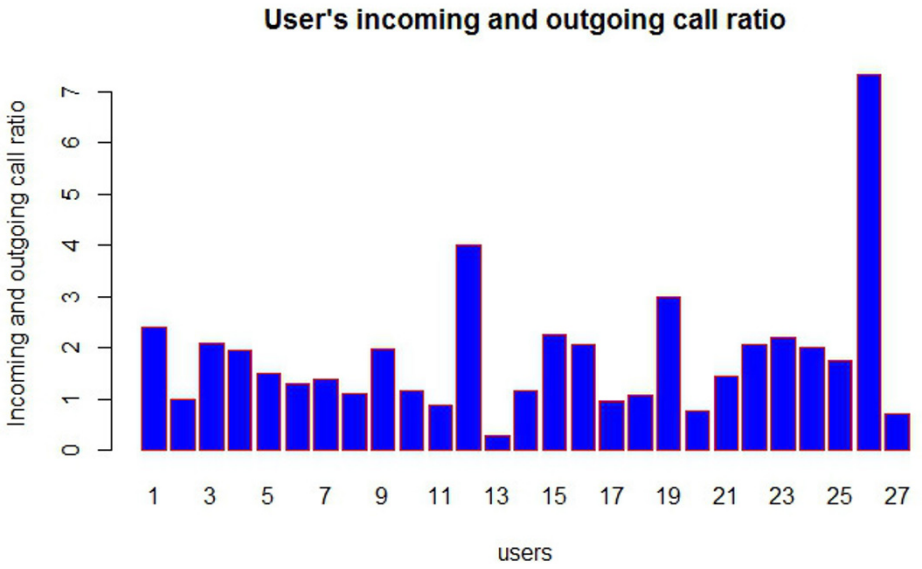


Fig. 2 Incoming and outgoing call ratio.

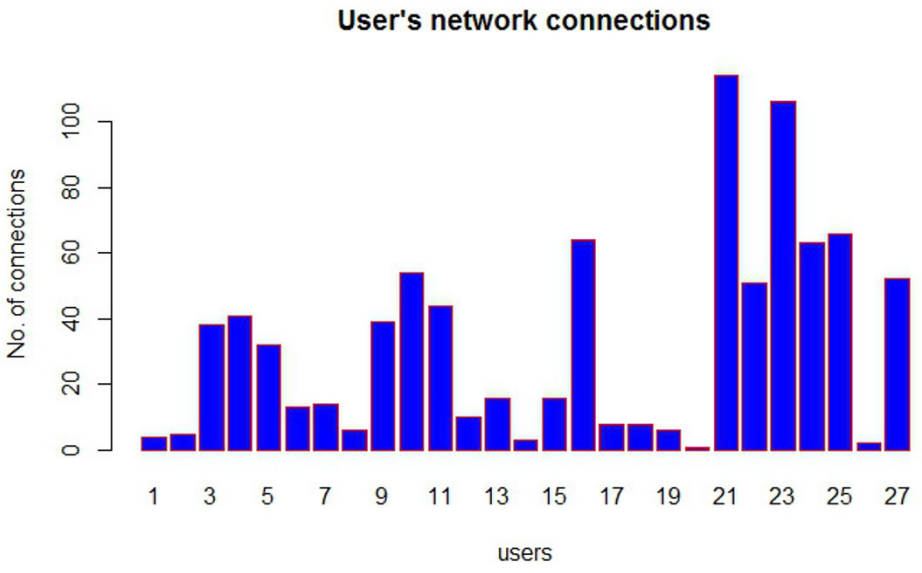


Fig. 3 Bar graph of number of people in each user's network.

5 Conclusions

In this paper the strath/nodobo dataset (v.2011-03-23) dataset was used for analyzing mobile phone usage behaviors of 27 customers. The unsupervised classification technique, clustering, was used to find both high and low mobile usage customers based on the frequency of calls and messages sent and the average duration of the calls. Additionally, users who make a lot of outgoing calls and are connected to many people in their social networks were also identified. In this work, simply by using a small dataset comprising of phone usage records, we have demonstrated how telecommunications companies can use customers' phone usage behaviors to improve their business and design better marketing strategies.

References

1. Bell S, McDiarmid A, Irvine J: **Nodobo: Mobile Phone as a Software Sensor for Social Network Research**. In *Vehicular Technology Conference (VTC Spring), 2011 IEEE 73rd* 2011:1–5.
2. **Number of mobile phone users worldwide from 2013 to 2019** 2016, [http://www.statista.com/statistics/274774/forecast-of-mobile-phone-users-worldwide/].
3. TechTarget: **Customer churn in the telecom industry**[http://searchcrm.techtarget.com/answer/Customer-churn-in-the-telecom-industry].
4. Dobardziev A: **Quality of mobile broadband the main reason for consumers to change provider**[https://www.ovum.com/press-releases].
5. **Can Advanced Analytics Help Telecom Businesses Reduce Customer Churn?**[http://www.digitalistmag.com/industries/telecommunications-industries/2015/03/03/advanced-analytics-help-telecom-reduce-customer-churn-02315792].
6. **Analytics solutions increase margin for US wireless provider**[https://www.accenture.com/ca-en/success-us-wireless-data-provider].
7. Arora D, Malik P: **Analytics: Key to Go from Generating Big Data to Deriving Business Value**. In *Big Data Computing Service and Applications (BigDataService), 2015 IEEE First International Conference on* 2015:446–452.
8. Lu N, Lin H, Lu J, Zhang G: **A Customer Churn Prediction Model in Telecom Industry Using Boosting**. *IEEE Transactions on Industrial Informatics* 2014, **10**(2):1659–1665.
9. Forhad N, Hussain MS, Rahman RM: **Churn analysis: Predicting churners**. In *Digital Information Management (ICDIM), 2014 Ninth International Conference on* 2014:237–241.
10. Li Y, e Xia G: **The Explanation of Support Vector Machine in Customer Churn Prediction**. In *E-Product E-Service and E-Entertainment (ICEEE), 2010 International Conference on* 2010:1–4.
11. Wu Y, Qi J, Wang C: **The study on feature selection in customer churn prediction modeling**. In *Systems, Man and Cybernetics, 2009. SMC 2009. IEEE International Conference on* 2009:3205–3210.
12. Gopal RK, Meher SK: *Customer Churn Time Prediction in Mobile Telecommunication Industry Using Ordinal Regression*, Berlin, Heidelberg: Springer Berlin Heidelberg 2008 :884–889.
13. **Winning the Intensifying Battle for Customers**[https://www.accenture.com/Accenture-Communications-Next-Generation-Customer-Analytics-Big-Data.pdf].
14. Yang X, Wang Y, Wu D, Ma A: **K-means based clustering on mobile usage for social network analysis purpose**. In *Advanced Information Management and Service (IMS), 2010 6th International Conference on* 2010:223–228.
15. **Analytics: Real-world use of big data in telecommunications**[http://www-935.ibm.com/services/multimedia/Anaytics.pdf].

16. Nasim M, Rextin A, Khan N, Muddassir Malik M: **On Temporal Regularity in Social Interactions: Predicting Mobile Phone Calls**. *ArXiv e-prints* 2015.
17. Mengshoel OJ, Desai R, Chen A, Tran B: **Will we connect again? Machine learning for link prediction in mobile social networks**. In *Eleventh Workshop on Mining and Learning with Graphs* 2013.
18. Fawcett T, Provost F: **Combining Data Mining and Machine Learning for Effective User Profiling**. AAAI Press 1996:8–13.
19. Mantoro T, Olowolayemo A, Olatunji SO: **Mobile user location determination using extreme learning machine**. In *Information and Communication Technology for the Muslim World (ICT4M), 2010 International Conference on* 2010:D25–D30.
20. Van T, Duong T, Tran DQ: **A fusion of data mining techniques for predicting movement of mobile users**. *Journal of Communications and Networks* 2015, **17**(6):568–581.
21. Mayrhofer R, Radi H, Ferscha A: *Recognizing and predicting context by learning from user behavior*. na 2003.
22. Kumar P, Kumar U, Rao M, Kaladhar D: **Prediction of Effective Mobile Wireless Network Data Profiling Using Data Mining Approaches**. *Journal of Algorithms Research* 2013, **2**:18–23.
23. Aheleroff S: **Customer segmentation for a mobile telecommunications company based on service usage behavior**. In *Data Mining and Intelligent Information Technology Applications (ICMiA), 2011 3rd International Conference on* 2011:308–313.
24. Lin Q, Wan Y: **Mobile Customer Clustering Based on Call Detail Records for Marketing Campaigns**. In *Management and Service Science, 2009. MASS '09. International Conference on* 2009:1–4.
25. Oliveira R, Brando WC, Marques-Neto HT: **Characterizing User Behavior on a Mobile SMS-Based Chat Service**. In *Computer Networks and Distributed Systems (SBRC), 2015 XXXIII Brazilian Symposium on* 2015:130–139.
26. Paireekreng W, Wong KW: **Mobile Content Personalisation Using Intelligent User Profile Approach**. In *Knowledge Discovery and Data Mining, 2010. WKDD '10. Third International Conference on* 2010:241–244.
27. Arora A, Vohra DR: **Segmentation of Mobile Customers for Improving Profitability Using Data Mining Techniques**. *International Journal of Computer Science and Information Technologies* 2014, **5**(4):5241–5244.
28. **Clustering**[https://en.wikipedia.org/wiki/Clusteranalysis].
29. **K-means clustering**[https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm].

NetFlow: Network Monitoring and Intelligence Gathering

Vivek Ratan and Kin Fun Li
Department of Electrical and Computer Engineering
University of Victoria, P.O. Box 3055 STN CSC, Victoria, B.C., Canada
University of Victoria
vivekratan2012@gmail.com; kinli@uvic.ca

Abstract. Network utilization, monitoring, and fault isolation are important aspects of modern day network management. Recently introduced technologies such as software defined networks, Internet of Things, and network virtualization necessitate the tracking of various physical and virtual devices in the organization. In this work, we examine the popular network monitoring tool NetFlow. Its key components and operations are introduced. Case studies are presented to illustrate how data mining and machine learning techniques can be utilized to derive high-level information from NetFlow data, facilitating business and operational decision makings in organizations.

1 Introduction

The rapid increase of network traffic has made network monitoring a monumental task for network administrators. The health of the network must be monitored constantly to ensure the bandwidth is utilized to its capacity without congestions and without heavy uses by application hogs. Of course, real-time network data can also provide network and system security management a wealth of information to detect and deter potential threats and attacks.

Among the many existing tools for network traffic monitoring, Cisco's NetFlow [1] is widely used in industry. In addition to supporting business and operation decision-makings, many organizations value NetFlow's information source to detect attacks, lower cost of network security monitoring, and enhance internal network visibility.

This work introduces NetFlow's components and operations in Sections 2. Applications and case studies in the realm of data mining and machine learning using NetFlow data are presented in Section 3. Finally, conclusions are made in Section 4.

2 NetFlow Components

NetFlow has various core components that enable network administrators to implement monitoring tasks easily with salient features [2].

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_83

2.1 Cache

The NetFlow cache can be used to store traffic flows that have been collected by the network monitor. There are two main types of cache:

- Temporary Cache: Timers are used to regulate the ageing and flushing of the cache. Contents flushed are sent to the exporter. This information can be used to track partial flows, which potentially can be used in the detection of SYN flood attack.
- Permanent Cache: The flows are stored permanently, which can be used in auditing and accounting. The network engineer can specify the cache size and once the cache is full, all flows will be dropped.

2.2 Flow Export Timers and Events

Flow data are stored in the cache memory. These data, originated from various routers, are merged for further analysis. Thus, it is important to have the flow records sent to the collector. The sending of the cache data takes place when the cache is recycled at various points in time:

- Inactive timer: cache entry has expired since it does not match any of the ingress traffic for a designated period of time.
- Active timer: cache entry matches the packets but it has exceeded its own time limit.
- Full cache: earlier entries are aged and must be flushed to create room for new entries.
- Flow termination: when either FIN or RST is set in TCP.

2.3 Flow Record

This is a set of attribute values used to describe the flows stored in the cache. A flow record is generated by examining the packet headers and adding the packet description to the cache.

2.4 Exporter

This is the router where the flows are captured and exported to the collector. NetFlow data can be accessed via a command line interface or a network reporting tool. Administrators can define the path to the exporter where the data are sent. Based on the specified parameters on the exporter, many optional settings are available including billing for device traffic and data tables.

2.5 Sampler

The sampler is used to define the packet sampling rate at the interface level when NetFlow is activated. Each packet sent is a sample out of N packets in the cache. N is

set such that the sampling process would not interfere with normal processing and other operations, and thus not affecting the overall performance of the system.

2.6 Monitor

The monitor keeps track of essential information including the types of traffic, the type of cache and its size, cache time out intervals, and the exporter used. Monitoring is done via the command line interface on the router.

2.7 Collector

This entity collects flow packets that are exported by the exporter. It then formats these flows into graphic objects to facilitate viewing and analysis. Data are collected from various exporters. The collection of data can then provide the network administrator an overall, centralized view of network usage, which is useful for capacity planning, billing, and security analysis.

2.8 NetFlow Operations

The operations of NetFlow are performed distributively at various components and levels [3]. The LANs and WANs perform traffic analysis within the respective networks, which are then captured by the routers via their interfaces. These interfaces are configured as monitors. The monitors send the flows to a router that also acts as an exporter. The router categorizes the traffic flows based on their source and destination IP addresses, ports, and class of service, etc. These classified traffic flows are grouped together in the cache for the network administrator to carry out further analysis.

3. Mining Data and Knowledge with NetFlow

Since NetFlow's introduction, network engineers and administrators have used the data mostly for operational purposes. However, some researchers and data scientists have dug deeper and attempted to mine 'knowledge' from NetFlow data. A wealth of information could be derived and used for making business decisions as well as operational decisions beyond traffic monitoring and control. A brief survey is presented here to show data mining activities associated with NetFlow in recent years.

The amount of NetFlow data is enormous and one needs to process these big data efficiently and effectively. In [4], a big data architecture is proposed for security monitoring. Intrusion detection and forensic analysis are done by mining DNS data in NetFlow records, HTTP traffic and honeypot data. This architecture also provides a distributed storage system to accommodate different data and manage changes. Based on the architecture and storage system, data correlation schemes for security are proposed and evaluated.

Since network devices produce enormous traffic data and these have to be monitored continuously, an adaptive estimation procedure is proposed in [5], partly to relieve the burden on network analysts. Current traffic pattern is given priority and is scanned for both false positives and any other missed out traffic streams.

Induction data mining techniques are used with NetFlow in [6] to counteract malware attacks that are intelligent enough to modify their own codes to avoid detection by regular scanners. The proposed architecture captures NetFlow data via monitors and creates a prediction model in the collector. Based on prediction and aggregation of newly created training sets, various induction based methods such as Naïve Bayesian Classifier, Decision Tree, and Support Vector Machine, are used to detect malware and classifies websites as either safe or malicious.

Brute force attacks are handled by analyzing traffic using NeFlow data in [7]. Using machine learning approaches, discriminative features in the traffic are extracted and used to prevent SSH brute force attacks.

NetFlow data works well within the FloVis system, which is a suite of visualization tools [8]. The system shows various views of network data for identifying rogue devices and malicious activities. Such a holistic and transparent view of the network enables the visualization of connections between hosts and between networks. Various tools including bundle diagrams, NetBytes viewer, and activity plots are used to determine interactions among various networking devices and to detect intrusions.

With the large volume of NetFlow data, big-scale software such as Hadoop and Spark are needed to analyze the captured traffic. To aggregate various traffic patterns, SequenceFile format and MapReduce are utilized to facilitate the reduction of space usage and access time by compressing the files in memory [9].

Data mining techniques are also used to improve network operations. Clustering traffic flows using NetFlow-type records to identify similar HTTP sessions allows redirection of traffic to the correct service, thus improving speed and performance of the network [10]. In [11], association rules and time-based event correlation are employed to classify dependencies among components in a network. The dependency model of the network provides valuable support in IT management and network reengineering.

An illustrative case study shows how NetFlow examines large volumes of real-time traffic and acts as a macro analytical tool [12]. Information such as the duration of traffic flow, the rules negotiated for data transfer, and the amount of data transferred, is invaluable for network management in the following aspects [12]:

- Detection of Denial-of-Service attacks
- Detection of unauthorized WAN traffic
- Reduction in peak WAN traffic
- Validation of QoS parameters
- Evaluation of the total ownership cost for applications

4. Conclusions

NetFlow is an indispensible tool in network planning and management. It has many benefits as described in this work, and can be easily adapted to various corporate network designs. A transparent view and improved visibility of the network facilitates troubleshooting and isolation for network administrators, thereby minimizing network down time and maximizing revenues.

It is anticipated that NetFlow will be incorporated into new tools that derive information using machine learning techniques for network security, planning, management, and also upgrades without affecting existing deployment.

References

[1] CISCO: Introduction to Cisco IOS NetFlow, White Paper (2012) 1-16
[2] Peddu, B.: ASR9K NetFlow White Paper. Cisco Support Community, https://supportforums.cisco.com/document/11933991/asr9k-netflow-white-paper [accessed: September 1, 2016]
[3] Pandora FMS: Documentation NetFlow. http://wiki.pandorafms.com/index.php?title=Pandora:Documentation_en:Netflow [accessed: September 1, 2016]
[4] Marchal, S., Jiang, X., State, R., Engel, T.: A Big Data Architecture for Large Scale Security Monitoring. IEEE International Congress on Big Data (2014) 56-63
[5] Bodenham, D.A., Adams, N.M.: Continuous Monitoring of a Computer Network Using Multivariate Adaptive Estimation. IEEE 13th International Conference on Data Mining Workshops (2013) 311-318
[6] Hsiao, H.-W., Chen, D.-N., Wu, T.J.: Detecting Hiding Malicious Website Using Network Traffic Mining Approach. 2nd International conference on Education Technology and Computer (2010) v5-276-280
[7] Najafabadi, M.M., Khoshgoftaar, T.M., Calvert, C., Kemp C.: Detection of SSH Brute Force Attacks Using Aggregated Netflow Data. IEEE 14th International Conference on Machine Learning and Applications (2015) 283-288
[8] Taylor, T., Paterson, D., Glanfield, J., Gates, C.: FloVis: Flow Visualization System. Cybersecurity Applications & Technology Conference for Homeland Security (2009) 186-198
[9] Zhou, X., Petrovic, M., Eskridge, T., Carvalho, M., Tao, X.: Exploring Netflow Data Using Hadoop. ASE Big Data / Social Com / Cybersecurity Conference (2014) 1-10
[10] Torres, L.M., Magafia, E., Izal, M., Morato, D.: Identifying Sessions to Websites as an Aggregation of Related Flows. XVth International Telecommunications Network Strategy and Planning Symposium (2012) 1-6
[11] Caracas, A., Kind, A., Gantenbein, D., Fussenegger, S., Dechouniotis, D.: Mining Semantic Relations using NetFlow. 3rd IEEE/IFIP International Workshop on Business-driven IT Management (2008) 110-111
[12] CISCO: NetFlow gives Network Managers a Detailed View of Application Flows on the Network, Case Study (2003) 1-13

Company Web Site Analysis based on Visual Information

Kosuke Takano and Yurika Moriwaki
Kanagawa Institute of Technology, 1030 Shimo-ogino, Atsugi, Kanagawa, Japan
takano@ic.kanagawa-it.ac.jp, s1021152@cce.kanagawa-it.ac.jp

Abstract. Many web search engines rank web documents based on textual information of the documents, where a user inputs query terms into a search engine, and receives the ranking results through a web browser interface based on the similarity score. However, we think that it is significant to consider the readability of web documents in the rank score, not only textual information, for ranking web documents. In order to obtain web pages with user's preference color objectively, we have developed a color-based ranking method for web documents. In this study, we present a method for web site analysis of companies based on the visual information using our color-based ranking method for web documents.

1 Introduction

Presently, most companies have a web site, and it is indispensable to our lives to find desired information according to our preference. Many web search engines rank web documents based on textual information of web documents, where a user inputs query terms into a search engine, and receives the ranking results through a web browser interface based on the similarity score.

However, we think that it is significant to consider the readability of web documents in the rank score, not only textual information, for ranking web documents. For example, when a user receives news articles from information recommendation systems [3], it would satisfy the user if the system provides news articles with a comfortable design and layout as well as good readability [4]. In addition, it is important for users who have particular color vision to obtain web documents with their preferable colors from a search engine. For example, for users who have the difficulty to perceive red color, it is desirable to exclude web pages that contain many red color fonts and backgrounds from the web search result, instead, to rank higher web pages with their acceptable colors.

In order to obtain objective web pages with user's preferred color, we have developed a color-based ranking method for web documents [5]. In our approach, we leverage color information of a web document, as well as textual information for

© Springer International Publishing AG 2017

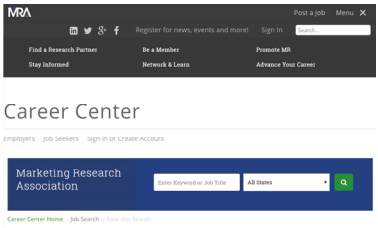
F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_84

ranking web documents. By using user's preferred color, which is automatically profiled into a database, our ranking method allows a user to obtain web documents according to the user's color taste. In this study, we present a method to analyze a company's web site based on visual information using our color-based ranking method for web documents.

2 Motivating Example

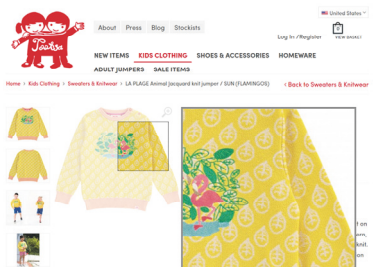
Although a great variety of designs and layouts for difference web sites are distributed on the web, it is deemed that many companies design their web sites based on the purposes and intents that they would like to convey to their consumers. For example, Figure 1 shows an example of web sites designed with a company's image. In Figure 1 (a), right and left figures show snapshots of job seeking sites for woman and focusing on the job area of market research, respectively. The woman's job seeking site uses light purple color that might give an impression of intelligence and style. On the contrary, the job seeking site for market research uses black and white colors, and we might get some rational and sensible impression for this design. Similarly, Figure 1 (b) shows snapshots of web sites of clothing companies. We can see that the child's clothing site uses red, yellow, and green colors to represent a child design, and the business casual clothing site uses dark gray and dark blue colors for a sophisticated and clever image.

[Woman's job]

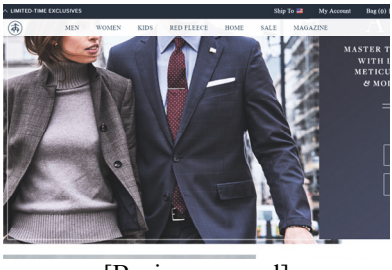


[Market research]

(a) Web sites of job seeking company



[Child's]



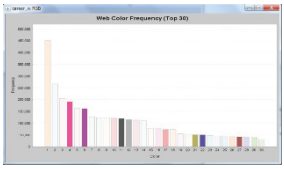
[Business casual]

(b) Web sites of clothes company

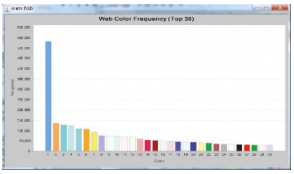
Fig. 1. Example of web sites designed with company's image

Meanwhile, there is a possibility that a user has preference of color and layout used in a web page, and is consciously or unconsciously browsing web pages based on the user's preference of web page design. For example, if a user is looking for interior accessory and household goods that appear warm for winter season, the user might frequently browse online shop sites that display items using warm tone colors such as orange and brown. On the contrary, if a user feels discomfort for web sites that contain loud colors and many advertisements, the user would skip those web sites upon the initial view.

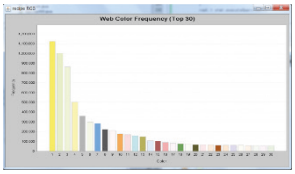
Our color-based web search engine stores a user's color preference of web sites as a color profile by capturing web sites that the user is browsing. Figure 2 shows an example of user's color profile for different genres of web sites for two users 1 and 2. From this example, we can see that users 1 and 2 have difference color preferences for each genre. For instance, since user 1 often browses job seeking sites with light colors, it is deemed that user 1 might look for jobs with bright image. On the contrary, user 2 browses job seeking sites with soft gray and blue colors, we can infer that this user prefers job seeking sites with a serious image.



[Job seeking]

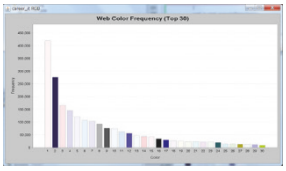


[Study]

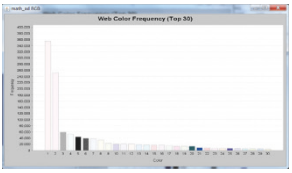


[Hobby]

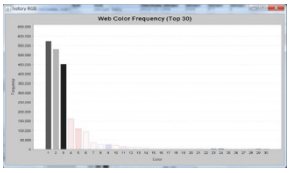
(a) User 1



[Job seeking]



[Study]



[Hobby]

(b) User 2

Fig. 2. Example of user's color profile

Figure 3 shows the basic idea of our method to analyze and improve the design of company's web site based on a user's color preference of web sites. First, a company web server captures a user's browsing behavior, and then extracts the user's color preference by accessing the user's personal color profile if the user permits. When many users visit a company's web site, but, leave in a short time, for example, it is deemed that they had interests in the company but lost them at a glance of the web site due to some reasons. In this case, if there is a gap between the web site colors and acceptable colors for many users, we can suggest an improvement to the design of the company's web site by changing colors. For example, as shown in Figure 3, if many users visit a web site "A" and leave in a short time, our method suggests the company to change colors of the web site "A" from green color to purple and pink colors that the users are likely prefer.

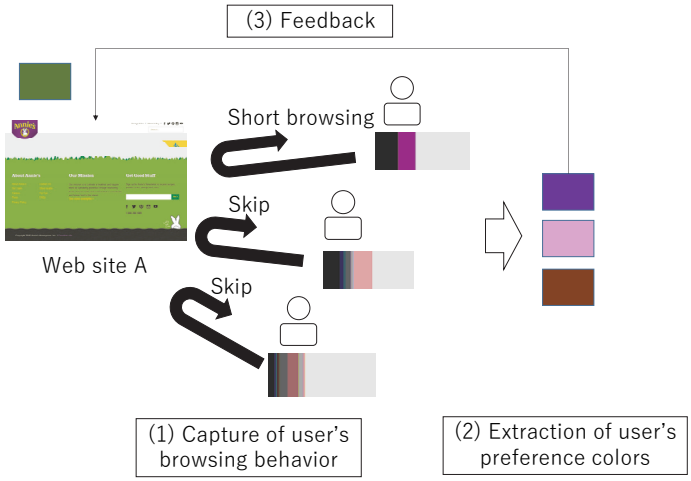


Fig. 3. Basic idea of web site analysis based on user's color profile

3 Color-based web search engine

Figure 4 shows an overview of our color-based web search engine [5]. Firstly, a user's color profile is constructed from the color information of web pages that the user is browsing through a web browser. Meanwhile, a web crawler with a color extractor indexes web documents using color metadata as well as term metadata. Using the user's color profile and the color metadata attached to each web document, our color-based search engine ranks web documents according to user's color taste.

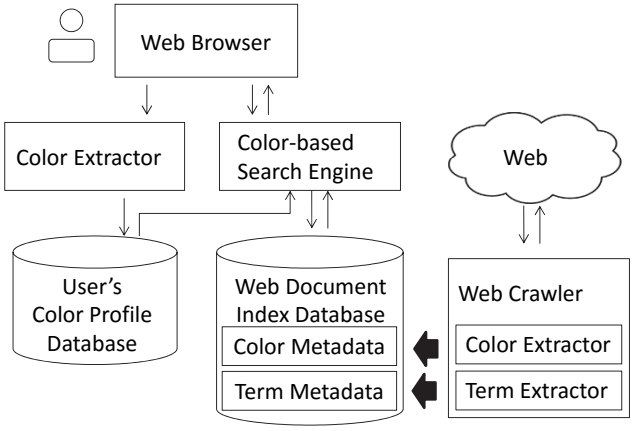


Fig. 4. Overview of color-based web search engine

3.1 Construction User's Color Profile

Through continuous browsing of web documents, the color information of the web documents that a user has been viewing is accumulated into a user color profile. The user's color profile is represented as a color-based feature vector $\mathbf{c}_u$, where n-colors definition (c_1, c_2, …, c_n) is used as the schema of feature vector, as follows:

$$\mathbf{c}_u = \lfloor\, c_{u1}, c_{u2}, \cdots, c_{un} \rfloor \tag{1}$$

3.2 Metadata Extraction from Web Documents

Color metadata is extracted from each web document by a color extractor within a web crawler. Using the same color scheme of $\mathbf{c}_u$, metadata color $\mathbf{c}_x$ of web document d_x is represented as a feature vector. All colors in d_x are reduced into n-colors schema by assigning a color to the similar color c_i.

$$\mathbf{c}_x = \lfloor\, c_{x1}, c_{x2}, \cdots, c_{xn} \rfloor \tag{2}$$

where c_{xi} is a pixel number of the color c_i used in d_x.

In order to extract term metadata from web documents, our method extracts term sets T_x from each web document d_x. T_x includes the m terms appearing in d_x.

$$T_x = \{\, t_{x1},\ t_{x2},\ t_{x3},\ \cdots,\ t_{xm} \,\} \tag{3}$$

where t_{xi} (i = 1, 2, …, m) is a term included in T_x. The total term set T of all terms appearing in each T_x is represented as follows:

$$T = \bigcup_{i=1}^{x} T_i \tag{4}$$

By using each term included in the term set T and each web document d_x, our method creates a document vector space [1] and a vectorized d_x.

$$\mathbf{d}_x = \lfloor\, u_{x1}, u_{x2}, \cdots, u_{x|T|} \rfloor \tag{5}$$

where u_{xi} is a term frequency of t_i appeared in a web document d_x.

3.3 Ranking of Web Documents

In our ranking method, firstly, the color-based similarity score S_c between c_u and c_x is calculated using the cosine measure [1] as follows:

$$Sc(c_u, c_x) = \frac{\mathbf{c}_u \cdot \mathbf{c}_x}{|\mathbf{c}_u||\mathbf{c}_x|} \tag{6}$$

Then, the term-based similarity score S_t between user's query q and d_x is calculated as follows:

$$St(q, d_x) = \frac{\mathbf{q} \cdot \mathbf{d}_x}{|\mathbf{q}||\mathbf{d}_x|} \tag{7}$$

Using both S_c and S_t, the ranking score S for each web document d_x is calculated,

$$S = \alpha \cdot S_c + (1-\alpha) \cdot S_t \tag{8}$$

where α $(0 \le \alpha \le 1)$ is a parameter to balance S_c and S_t.

Finally, our method ranks each web document based on the score S so that the user can obtain web documents according to his/her color taste.

4 Analysis of web sites based on the visual information

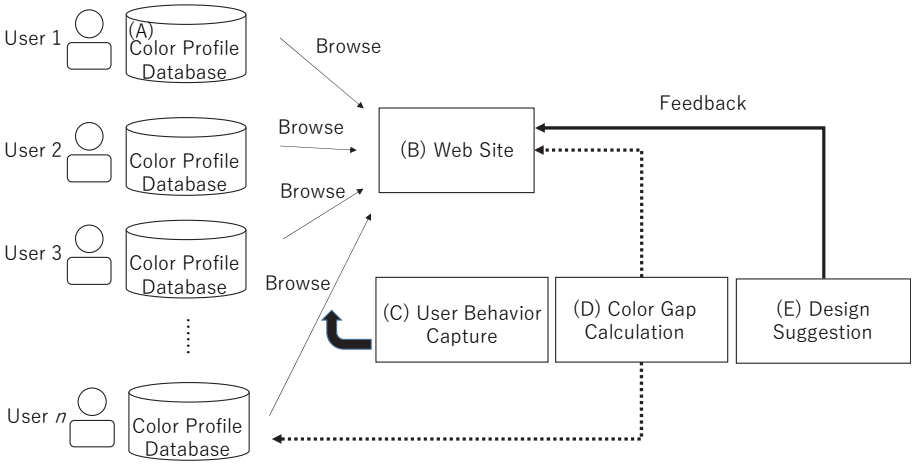


Fig. 5. Overview of analysis method of web site based on the visual information

By extending our color-based web search engine, we design a method of analyzing web sites based on the visual information. Figure 5 shows an overview of our method. It mainly consists of (A) color profile databases from different users, (B) the web site of a company, and three functions: (C) capture of user's browsing behavior, (D) calculation of color gap between colors used in the web site and summarized colors from different users' color profile, and (E) suggestion of design for the web site.

Our method is executed in the following steps:

Step-1: User's browsing behavior to the company's web site is captured, and k users U whose browsing time t is less than a threshold T_1 are extracted.

$$U = \left[u_1, u_2, \cdots, u_k \right] \tag{9}$$

Step-2: Color profiles from n users are summarized into color information $\mathbf{C}$, and the color gap G between colors used in the web site and the summarized color information C is calculated.

$$\mathbf{C} = \frac{1}{k} \sum_{u=1}^{k} \mathbf{c}_u \tag{10}$$

For example, G is calculated using the Jaccard similarity.

$$G = \frac{\left| \mathbf{C} \cap \mathbf{c}_x \right|}{\left| \mathbf{C} \cup \mathbf{c}_x \right|} \tag{11}$$

Step-3: In Step-2, when color gap G is less than a threshold G_1, difference color information $\mathbf{C}_{diff}$ are extracted. Colors with top-n values are extracted from $\mathbf{C}_{diff}$, and are used to suggest improvement the design of the web site.

$$\mathbf{C}_{diff} = \mathbf{C} - \mathbf{c}_x \tag{12}$$

5 Validation and Conclusion

In this paper, we have presented a method for company web site analysis based on the visual information using our color-based ranking method for web documents. Presently, we are implementing our prototype, and we will show the feasibility and the effectiveness of the proposed system in our future work.

References

1. Baeza-Yates, R., and Ribeiro-Neto, B.: Modern Information Retrieval, Addison Wesley, (1999)
2. Deerwester, S., Dumais, S., Furnas, G.W., Landauer, T.K., and Harshman, R.: Indexing by latent semantic analysis, Journal of the American Society for Information Science, Vol. 41, No. 6 (1990) 391–407
3. Jannach, D., Zanker, M., Felfernig, A., and Friedrich, G.: Recommender Systems: An Introduction, Cambridge University Press (2010)
4. Hearst, M. A.: Search User Interfaces, Cambridge University Press (2009)
5. Moriwaki, Y. Honda, H., and Takano, K.: Color-based Ranking Method for Web documents, Proceedings of the 2nd Asian Conference on Information Systems (2013) 430–433
6. Liu, W., Meng, X., and Meng, W.: ViDE: A Vision-Based Approach for Deep Web Data Extraction", IEEE Transactions on Knowledge & Data Engineering, Vol. 22, No. 3 (2010) 447–460
7. Li, K.F., and Takano, K.: Modelling User Behaviour in Recommender Systems Based on Content, Structure, Layout, and Schema, in Learning Structure and Schemas from Documents, Biba, M. and Xhafa, F. (Eds.), Springer Series in Studies in Computational Intelligence, SCI 375, Chapter 14 (2011) 289–313

A Study of the Standardization of Fire Forces Equipment Procurement

Zhang jie
Department of Military Management, Chinese People's Armed Police Forces Academy, Xi Huan Road #220, An Ci Section,065000,Langfang, Hebei Province, China;
xiaowangniuniu33@163.com

Abstract. This paper probes into the measures of achieving the standardization of fire forces equipment procurement supported by the harmonious management theory, based on the characteristics and existing problems of the procurement of the fire-fighting equipment.

1 Introduction

With the development of social economy, the functions of the fire forces are increasingly diversified, the expenditure of fire-fighting equipment procurement steadily grows, all kinds of fire-fighting and rescue equipment continuously upgrades, and the variety and technology of fire-fighting equipment is becoming comprehensive and complex day by day. Facing the complex fire-fighting equipment market, how to effectively use the funds of fire-fighting equipment procurement and standardize the procurement behavior is the priority pf the logistics management of the fire forces.

2 The characteristics and problems of equipment procurement

2.1 The characteristic of equipment procurement: complexity

Due to the particularity of equipment procurement activities and the complexity of equipment manufacturing, there are many difficulties in the process of military equipment procurement. Equipment, as public goods, whose quality is evaluated by a group, it is difficult to achieve subjective satisfaction; as a high-tech product which highly contains the knowledge of intangible value, it is difficult to assess the quality according to the perception of customers; although the equipment is mainly obtained through transactions as same as the general commodities, the markets are Imperfect competitive markets, and the quality cannot rely on market competition. In short, the equipment procurement is a complex process based on its argumentation of demand, procedure, incentive, evaluation, and the internal supervision of the forces. The complexity and its sources are shown in the following table:

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_85

Process	Argumentation of demand	Procedure	Incentive, evaluation, internal supervision of the forces
Subject	Government Army	Equipment supplier Army	Purchasing center Audit and supervision department Equipment using department
Sources Of Complexity	Total expenditure Market institution Technical environment Laws and regulations	Market structure Laws and regulations Information distortion	Supervision mechanism Incentive mechanism Evaluation mechanism Information feedback mechanism

2.2 Non-standard problems in the process of equipment procurement of fire forces

To effectively deal with the complexity of the above aspects, it is necessary to strictly regulate the process of fire-fighting equipment procurement. At present, because of the problems of the equipment procurement system of fire forces, the standardization management has not been effectively carried out, and the main drawbacks are as follows.

In terms of materials, there're non-standard problems existing in the institutions, structure, and procedure. From a macro perspective, the laws and regulations of fire forces equipment procurement are still not thorough. The "Government Procurement Act" and the "Tendering and Bidding Law" only stays in the level of laws and regulations, there's no specific basis and standard for equipment procurement, the planning effort of top-level design is not enough, the thought of system construction has not been established, the formulation of supporting regulations is slow, the role of centralized leadership needs to be further strengthened, etc. From a market perspective, the implementation of market access system is not strict, the equipment procurement market is in disorder, the phenomenon of "Bad money driving out good money" appears, the procurement of equipment standardization has just begun, the formulation and implementation of standards remains to be strengthened; finally, equipment procurement process is in its framework, specific operating procedures has not been established and its situations are various. And it shows in following aspects: Firstly the determination of procurement method is not refined; secondly, the procurement process is relatively disordered; thirdly, the supervision and management system of the procurement activities is still not sound; fourthly, the supplier selection management is not enough.

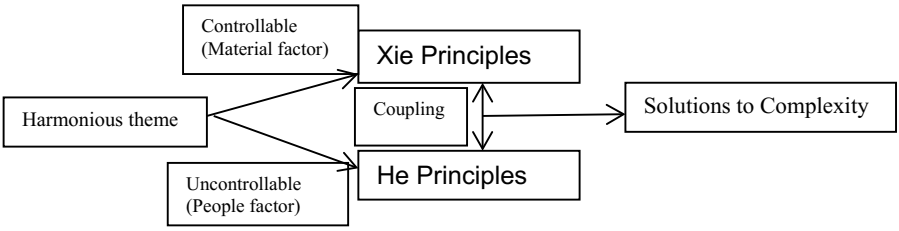
3 Harmonious management theory: the theory of coping with the complexity of equipment procurement flexibly

The basic opinions of the harmonious management theory include the following three parts.

Firstly, any reform of organization system should be based on specific issues, with the purpose of solving the problem, that is, to define the harmonious theme of the stage.

Secondly, the complexity of social system mainly comes from two aspects, that is, the complexity of people and materiel. Therefore, complex problems should be dealt with in two aspects.

Thirdly, there is close connection between the above-mentioned "Xie Principles" and "He Principles" which is chosen according to the harmonious theme, that is, the "Harmonious coupling" in harmonious management. Harmonious coupling is the adjustment of the relationship of "Xie Principles" and "He Principles" based on the harmonious theme, and also is the process of adaption and evolution in the different levels of the organization of "Xie Principles" and "He Principles" based on the harmonious theme. The key to harmonious coupling is to reveal the organizational problems emerge in specific environment and how to finish the task by using the "Xie Principles", "He Principles" and their combination. The basic framework of harmonious management theory is as follow.



4 Establishing a standardized management system of fire forces equipment procurement based on harmonious management theory

4.1 The harmonious theme of military equipment procurement

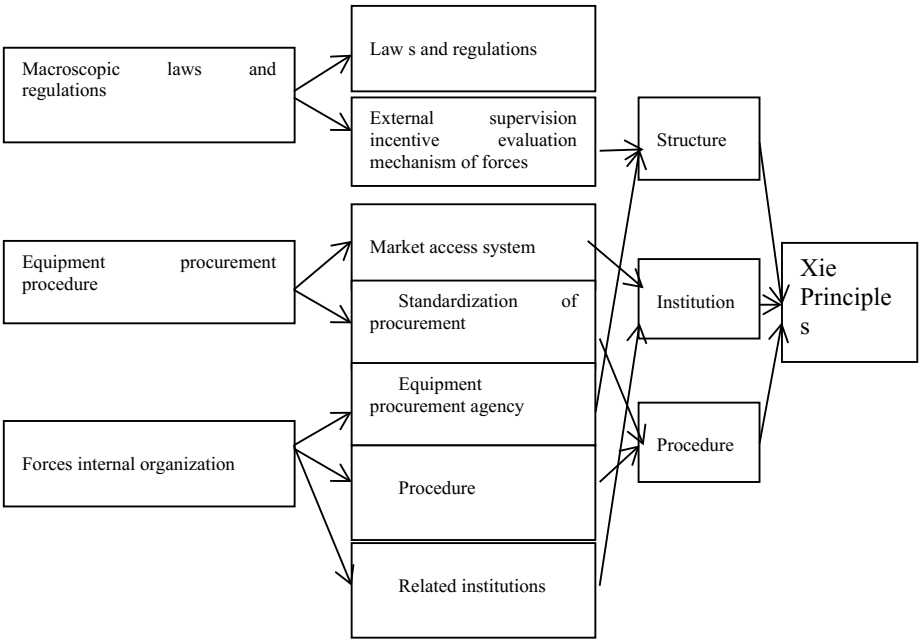
The harmonious theme of military equipment procurement: standardized management**.**

4.2 Realizing the "Xie Principles" in the standardized management of fire forces equipment procurement

The "Xie Principles" reduces uncertainty mainly by prescribed line, and the specific tools and methods includes the construction and improvement of structure, procedure,

institution, and so on. In the process of fire forces equipment procurement, the uncertainty mainly comes from three aspects, because that this kind of prescriptive design includes three levels. First is the formulation and perfection of institution, laws and regulations from the macro level of country. The second is the improvement of the standardization of equipment procurement and the construction and perfection of market access system from the meso level. The third is the improvement of internal structure, procedure and institution of the organization from the micro level.

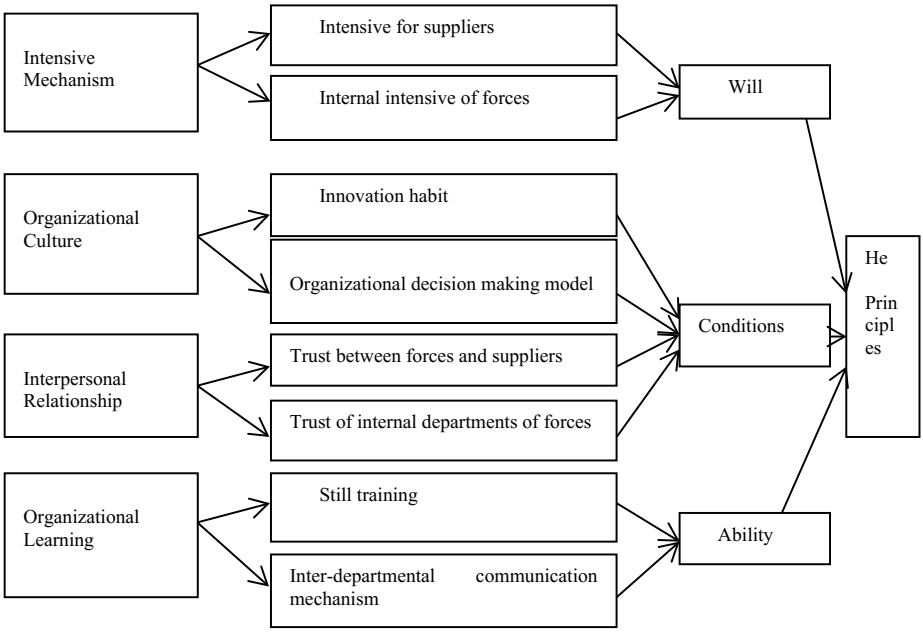
In the whole, although there are differences among above-mentioned optimal control mechanisms, their basic principles are similar. They all use the optimization and design thought originated from the "Xie Principles" of harmonious management theory, and achieve the purpose of reducing uncertainty with the method of regulate the behavior route of the actor. And these behaviors are basically to establish the related structure, procedure and institution. The framework of the "Xie Principles" of fire-fighting equipment is as follows.



4.3 Achieving the "He Principles" in the standardized management of fire forces equipment procurement

The above-mentioned "Xie Principles" in equipment procurement management mainly cope with the optimizable management problems through prescribed line, but there're problems caused by uncertainty are not suitable to be solved by prescribed line. Because of the lack of flexibility in certain programme, it cannot adopt the volatile and frequent innovation in equipment procurement. Besides, due to the bounded rationality of human, there're many lines and regulations of matters that

can't be designed because of the limited cognitive level, such as the incentive of the procurement relating staff, the build of organization atmosphere and the trust among the members of different organizations. These "social problems" are all caused by the uncertain of people. And according to the theory of harmonious management, it is supposed to be solved by the "He Principles". The uncertainty can be reduced by improving skill level (i.e. ability). The limitation of the ability in equipment procurement mainly embodies in the ability of assessment experts and the managing ability of procurement staff. Study shows that the main approach of the improvement of members' ability is organizational learning, and the most common ways are training and inter-departmental communication. The main method of reducing uncertainty is the support of organizational conditions by using the "He Principles" in harmonious management theory, which mostly includes cultural atmosphere and trust relationship. The framework of the "He Principles" in fire-fighting equipment procurement is shown in following table.



4.4 Realizing the "harmonious state" of standardized management of fire forces equipment procurement

The two above-mentioned mechanisms gives the solutions to the complexity caused by material factor and people factor, but the problems in practice cannot be easily divided into two types. They're intertwined and changing, and this kind of dynamic nature requires the constantly change of the combination of the two mechanisms, especially the relationship between the two mechanisms, that is, "harmonious

coupling", and eventually achieving the "harmonious state" of equipment procurement –the standardized management.

The key to harmonious coupling is the suitable chosen in particular situation. The basic thought of choosing is that dealing with the problems which can be clarifies by certain regulated line by the thought of design and optimization; dealing with the uncertain problems which caused by people's cognition, will , ability and can't be solved by prescriptive line by reducing and using the uncertainty in the way of inducing. Such as designing and adopting incentive contract, setting incentive compatibility as the principle, formulating the content of bidding document scientifically, considering the suppliers' profit and encouraging the suppliers to innovate in technology development and produce the equipment with good performance by improving the technical function parameters when designing the base price. The main function of the "He Principles" is stimulating people's initiative and creativity. But only initiative and creativity is not enough, certain procedure and regulations are required to make sure that the initiative can be used rightly and effectively. That's the reason of the demand of the "Xie Principles". For example, providing more chances of promotion and performance to those procurement staff that have both ability and political integrity by the design of professional promotion incentive mechanism, and forming a good positive stimulus by providing a wider spectrum of career opportunities.

References

1. Qi Zhuang, The National Fire Forces Equipment Goes to a New Level [J], Fire-fighting Weekly, The people's public security newspaper, 5 March 2010.
2. Yu Zongming, Ding Kemou, Jiang Zhenduo, Zhu Hongshan, Enlightenment of Foreign Weapon Procurement Reform [J], Military Economic Research, 2009, (2)
3. Liu Baohong, Management of Procurement and Supply Chain [M], Chinese Machine Press, Beijing, 2012.
4. Liu Xiaofeng, Mo Liangguang, A Study on Chinese Government Procurement Institution [J] , Journal of Central University of Finance and Economics, 2012, (10)
5. Chen Haiwu, A Study on Standard Operation of Military Procurement Institution [J], Military Economic Research, 2012, (4)
6. Ningping Sun; Soichiro Murakami; Hiroshi Nagaoka; Takuya Shigemoto. A correction algorithm for stereo matching with general digital cameras and web cameras DOI: 10.1504/IJSSC.2013.056408
7. International Journal of Space-Based and Situated Computing 2013 Vol. 3 No. 4.
8. Masafumi Takematsu; Kazunori Uchida; Junichi Honda. Investigation of observational plate length and diffraction point for DRTM DOI: 10.1504/IJSSC.2013.058369 193-206
9. Sayed Mohammad Almohri; Ghufran A. Al-Shiridah; Maytham Safar; Khaled Mahdi. Correlating feedback capacity with degree of diffusion in heterogeneous complex networks DOI: 10.1504/IJSSC.2013.058370 207-214
10. International Journal of Space-Based and Situated Computing 2011 Vol. 1 No. 1.

Part VIII
Workshop-SiPML 2016: Workshop on Signal Processing and Machine Learning

Monitoring of Cardiac Arrhythmia Patterns by Adaptive Analysis

José Elías Cancino Herrera[1], Ricardo Rodríguez Jorge [1], Osslan Osiris Vergara Villegas [1], Vianey Guadalupe Cruz Sánchez [1], Jiri Bila [2], Manuel de Jesús Nandayapa Alfaro[1], Israel U. Ponce[1], Ángel Israel Soto Marrufo[1], Ángel Flores Abad[1]
[1] Autonomous University of Ciudad Juarez
Emails: al94596@alumnos.uacj.mx,{ricardo.jorge, overgara, vianey.cruz, mnandaya, israel.ulises, angel.soto, anflores}@uacj.mx
[2] Czech Technical University in Prague
Emails: Jiri.Bila@cvut.cz

Abstract. In this paper, a study and development of a monitoring adaptive system based on dynamic quadratic neural unit are presented. The system is trained with a recurrent learning method, sample-by-sample in real time. This model will help to the prediction of possible cardiac arrhythmias in patients between 23 to 89 years old, age range of the electrocardiogram signals obtained from the Massachusetts Institute of Technology-Beth Israel Hospital arrhythmia database. By means of the implementation of this adaptive monitoring system the model is capable of processing heart rate signals in real time and to recognize patterns that predict cardiac arrhythmias up to 1 second ahead. The Dynamic Quadratic Neural Unit in real time has demonstrated presenting greater efficiency and precision comparing with multilayer perceptron-type neural networks for pattern classification and prediction; in addition, this architecture has demonstrated in developed research, to be superior to other different type of adaptive architectures.

1 Introduction

In recent years, investigations related to biological control have demonstrated a significant progress in the comprehension of the behavior of the biologic neural structure. These investigations are of great interest in the medical field because of the complex characteristics that present the biological neural model when processing perceived data. Recent mathematical models and architectures have been developed thanks to the progresses presented day-by-day in the neurological field, it allows reaching the aim of an architecture of processing efficiently by a methodology, systems naturally chaotic [1], [2].

Towards the goal of modeling a system with the biologic neural characteristics, a mathematical model named quadratic neural unit (QNU) has been developed [3], [4].

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_86

This model is considered a type of higher-order neural unit (HONU). It has been capable of processing systems that present linearities and non-linearities on spaces of continuous or discrete time. The QNU model is considered as a special class of polynomial neural networks. The works of [4], [5] and [6] have presented QNUs indicating the characteristics of the neural unit of providing a faster response comparing to a controller by linear states feedbacks as well as with controllers applied to nonlinear systems, unstable systems, and unknown non-linear dynamic systems.

The development of this type of higher-order neural units opens new perspectives in the solution of complex problems, such as pattern recognition and automatic control, both issues were addressed by Benes et al. [7], and Rodriguez et al. [8]. Banes focused in automatic control, while Rodriguez proposed the use of QNU to reduce the number of neural weights without sacrificing the superior neural unit performance.

The adaptive novelty detection algorithms are addressed with higher-order neural units (HONUs), in [9] where authors presented two algorithms of adaptive novelty detection based on supervised learning of a HONU and a self-organized map model both with extensions for adaptive monitoring of the existing closed-loops.

These new perspectives of HONUs are thanks to the characteristics that possess the neural units, towards to this type of systems classified as difficult to model because of its high complexity such as: flexibility, adaptability, efficiency, and time response that offer [1], [2].

Currently, the unusual pattern detection and prediction of dynamic systems is a topic of scientific interest. The representability by time series usually are used for the observability of real time monitoring systems. The complexity of heart rate variability signals can be considered as chaotic because of its instability and perturbations that present during the monitoring; because of that it is difficult the electrocardiogram (ECG) signal modeling and prediction. The variations affect the prediction accuracy and therefore the applications where it has been implemented are compromised. One of the examples that can be mentioned is the application in medical treatment. It is because of the above mentioned, that the ECG monitoring by adaptive models is proposed.

In this research work, a pattern monitoring system is constantly fed of new data obtained from the heart rate signals acquired from Massachusetts Institute of Technology- Beth Israel Hospital (MIT-BIH) arrhythmia database [10] The system has been implemented by a dynamic quadratic neural unit (D-QNU) with real time recurrent learning (RTRL), sample-by-sample; in this sense, the system error function decrease considerably and the algorithm is capable of modeling in real time, efficiently, the people cardiac behavior between 23 to 89 years old. In addition, by the implementation of the adaptive monitoring system, the algorithm is able of recognizing patterns that can predict a possible cardiac arrhythmia up to 1 second ahead.

The paper is organized as follows: the dynamic structure and the RTRL learning technique of the QNU are addressed in the section 2. In addition, in this section, the methodology of the adaptive monitoring system of heart rate variable signals for the prediction of possible cardiac arrhythmias is described as well. The section 3 presents the results and finally, the section 4 presents the conclusions and future work.

2 Adaptive methodology

There are several types of adaptive methodologies, such as multilayer neural networks, higher-order neural networks, adaptive self-organized maps, and adaptive radial basis function neural networks, among others. However, the main goal is in fact the same: fed input data to the QNU are processed to obtain the output capable of classifying or adjusting a desired function.

According to the above mentioned, the QNU is a type of polynomial neural network that is able to adjust any kind of function. It is possible with just having the proper values of the adjustable parameters. Such values are known as neural weights and they are used to save the knowledge of the QNU. The knowledge is acquired and saved by a training stage or learning, which consists on modifying the neural weights, with the aim of reaching a better performance in approximation or in the prediction or classification [11]. There are basically two types of training: supervised and unsupervised. In this paper, only the supervised training is applied, which consists on feeding the QNU with a data set as inputs and a desired output that the model should predict. The objective is to learn the second order correlation between the given data set. The learning algorithm will update the neural weights, in such a way that the neural output is as more similar as the training data set. Fig. 1 presents the adaptive methodology process.

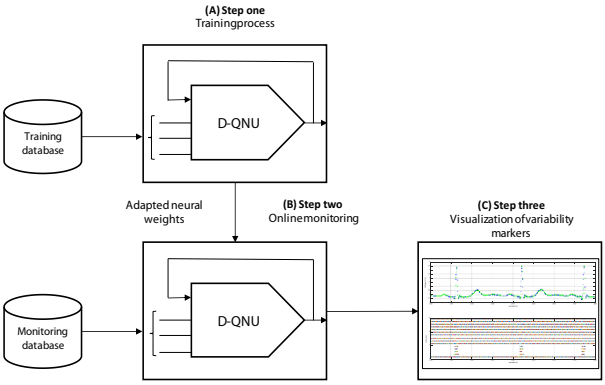


Fig. 1. Adaptive methodology process of the monitoring system.

Then, the learning, online monitoring and the variability markers visualization of the D-QNU are presented next. These stages are used in the arrhythmia patterns monitoring system.

2.1 Step one: training process

The QNU can be classified in two categories: static and dynamic, however, the static QNU is not considered in this paper. The D-QNU use a RTRL learning method and can be implemented in discrete or continuous real-time. In (1) the notation of D-QNU with RTRL is shown, where $\tilde{y}(k+h)$ is the predicted neural output and ^{T} represents

the transpose of the vector $\mathbf{x}$, which contains the feedback neural output values and the signal from the heart rate variability (HRV).

$$\tilde{y}(k+h)=\left[\sum_{i=0}^{nx-1}\sum_{j=0}^{nx-1}x_j\cdot w_{i,j}\cdot x_i\right]=\left(\mathbf{x}^T\cdot\mathbf{W}\cdot\mathbf{x}\right) \tag{1}$$

The upper triangular matrix of weights $\mathbf{W}$ with neural bias $w_{0,0}$ is defined as

$$\mathbf{W}=\begin{bmatrix} w_{0,0} & w_{0,1} & \cdots & w_{0,nx-1} \\ 0 & w_{1,1} & \cdots & w_{1,nx-1} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & w_{nx-1,nx-1} \end{bmatrix} \tag{2}$$

Equation (3) is the column of the augmented vector $\mathbf{x}$ of the neural input, $y(k)$ is the real value, n is the number of real values that feeds the neural input, k is the variable that describes the discrete time and h is the prediction horizon.

$$\mathbf{x}=\begin{bmatrix} 1 \\ \tilde{y}(k+h-1) \\ \tilde{y}(k+h-2) \\ \vdots \\ \tilde{y}(k+1) \\ y(k) \\ y(k-1) \\ \vdots \\ y(k-n+1) \end{bmatrix} \tag{3}$$

The architecture of the D-QNU is shown in Fig. 2, where $1/z$ represents the values of feedback, which are fed to the augmented vector $\mathbf{x}$.

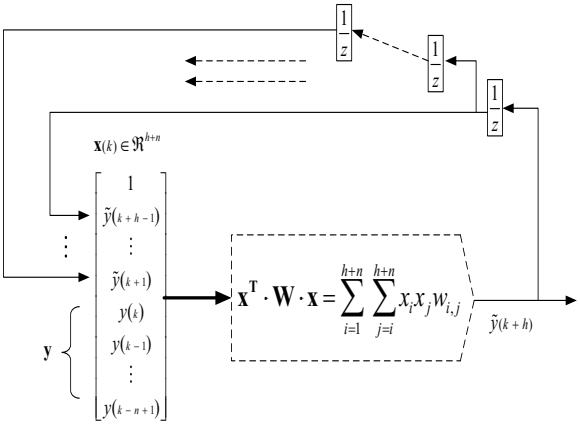


Fig. 2. Dynamic QNU architecture.

An error function is defined in such a way that, in each learning step indicates how close the solution is to minimize the error function (mean squared error function) of the D-QNU model; $\frac{1}{N}\sum_{k=1}^{N} e(k+h)^2$ in each prediction time $k+h$. In (4), $w_{i,j}(k+1)$ stands for neural weights adaptation in the new adaptation time of the model, which are the sum of the neural weights increment $\Delta w_{i,j}(k)$ from the adaptation of the model, with individual weights $w_{i,j}$ of each k discretized value from the HRV.

$$w_{i,j}(k+1) = w_{i,j}(k) + \Delta w_{i,j}(k) \tag{4}$$

The increment of the neural weights with RTRL is described in (5), where μ is the learning rate that determines in which proportion the neural weights are updated and the velocity of the learning process, $e(k+h)$ is the error in each discrete time $k+h$ that multiplies the partial derivative of the neural output with respect to the weights $w_{i,j}$ described in (6).

$$\Delta w_{i,j}(k) = \mu \cdot e(k+h) \cdot \frac{\partial \tilde{y}(k+h)}{\partial w_{i,j}} \tag{5}$$

$$\begin{aligned} &\frac{\partial \tilde{y}(k+h)}{\partial w_{i,j}} = \\ &\frac{\partial (\mathbf{x}^T \cdot \mathbf{W} \cdot \mathbf{x})}{\partial w_{i,j}} = \left(\frac{\partial \mathbf{x}^T}{\partial w_{i,j}} \cdot \mathbf{W} \cdot \mathbf{x} + \mathbf{x}^T \cdot \frac{\partial \mathbf{W}}{\partial w_{i,j}} \cdot \mathbf{x} + \mathbf{x}^T \cdot \mathbf{W} \cdot \frac{\partial \mathbf{x}}{\partial w_{i,j}} \right) \end{aligned} \tag{6}$$

In (7), $\mathbf{jxw}_{i,j}$ is described as the partial derivative of the augmented vector $\mathbf{x}$ of the neural inputs with respect to the neural weights.

$$\begin{aligned} &\frac{\partial \mathbf{x}}{\partial w_{i,j}} = \mathbf{jxw}_{i,j} = \\ &\left[0 \; \frac{\partial \tilde{y}(k+h-1)}{\partial w_{i,j}} \; \frac{\partial \tilde{y}(k+h-2)}{\partial w_{i,j}} \cdots \frac{\partial \tilde{y}(k+1)}{\partial w_{i,j}} \; 0 \cdots 0 \right]^T \end{aligned} \tag{7}$$

Meanwhile, in (8) the representation of a full matrix is shown, which is a convenient adaptive representation of a D-QNU, where $\mathbf{r}\Delta\mathbf{W}(k)$ is a vector with all increased weights and $\mathbf{jxw}_{i,j}$ is the Jacobian matrix evolving recurrently [12].

$$\begin{aligned} \mathbf{r}\Delta\mathbf{W}(k) = \mu \cdot e(k+h) \cdot \\ \left((\mathbf{jxw}_{i,j}(k)^T \cdot \mathbf{W} \cdot \mathbf{x})^T + \mathbf{r}\{\mathbf{x}^T \cdot \mathbf{x}\} + \mathbf{x}^T \cdot \mathbf{W} \cdot \mathbf{jxw}_{i,j}(k) \right) \end{aligned} \tag{8}$$

where $\mathbf{r}\{\mathbf{x}^T \cdot \mathbf{x}\}$ represents all the combination of the input $\mathbf{x}$. The Jacobian matrix is updated on a recurrent basis for the adaptation of the model.

2.2 Step two: online monitoring

The D-QNU proposed model is under research, the development of this methodology has shown to be useful in real-time evaluation processes of dynamic systems from real-world and theoretical systems of chaotic behavior.

The D-QNU system contains an adaptive graphic monitor, which is used for displaying the HRV in real time, is shown in Fig. 3. This system consists of detecting and visualization unusual patterns that can indicate possible cardiac arrhythmias in patients between 23 to 89 years old, by adapting the neural weights during the model adaptation. In the system, sample-by-sample adaptation was implemented, based on the gradient descent rule with RTRL.

The predictive model monitors online the records of electrocardiograms (ECG) obtained from MIT-BIH, a database freely-available [10]. This makes that the predictive model is able to identify patterns in time series and to achieve predicting values of HRV during the real-time monitoring. This monitoring is performed on different data set utilized on the training stage, it is illustrated on Fig. 3.

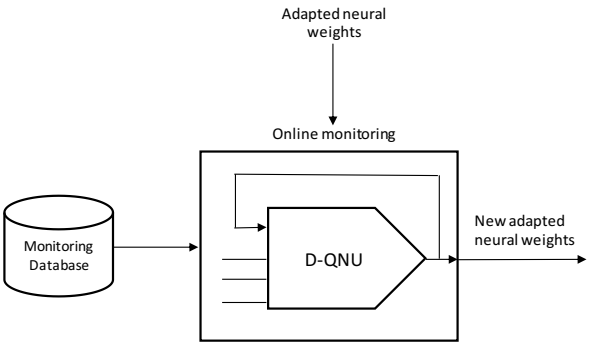


Fig. 3. Online monitoring.

2.3 Step three: visualization of variability markers

In the adaptive methodology the neural weights are initially random values from -1 to 1, such values are updated by the learning process in every iteration time, respecting to the system training intervals. During this process the sum of squared error is reduced to the vicinity of the global minimum. The predictive model acquires data continuously by the sample-by-sample adaptation. Unusual adaptive neural weights are evaluated and visualized by a graphical interface designed in Matlab. For the detection of changes in the signal, a sensitivity parameter is initialized experimentally as $\alpha = 2.95$. This sensitivity parameter is considered on the online monitoring to

influence a condition that allows drawing markers according to the evaluation of the increment parameters. To allow drawing the markers, a condition is considered as in (9).

$$\left|\Delta \mathbf{W}(k)\right| \geq \alpha \cdot \overline{\Delta \mathrm{W}(k)} \tag{9}$$

where the $|\cdot|$ stands for the absolute value, and $\overline{\Delta \mathrm{W}(k)}$ corresponds to the average of the weight increments. Accordingly, for the weight increments average, a window of a set of values, have been considered.

This adaptation of D-QNU converges singular and fast to the vicinity of global minimum thanks to the gradient descent methodology. The configuration of inputs and outputs of the D-QNU are nonlinear, however, the optimization of the neural weights is a linearly solvable task due to the D-QNU mathematical structure. It is important to point out that the mathematical structure does not present problems with local minima due to the adjustment made by the quality of the training data, however, improper data when training will affect significantly the predictive model.

3 Results

The MIT-BIH arrhythmia database was utilized to validate the adaptive monitoring system proposed in this article. The arrhythmia database contains sampled records to 360 Hz, with a resolution of 11 bits over a range of 10 mV; each record contains a duration of 30 minutes with 5.556 seconds [10]. For the adaptive monitoring, only the first channel of each record has been considered, because both channels describe the same signal behavior. Records contain beats with inverse polarity, low amplitude, ventricular ectopic with low signal noise-to-ratio, premature ventricular heartbeats and premature atrial heartbeats, which are duly identified during the real-time adaptive monitoring. It is important to point out that the adaptive model is capable of identifying variabilities of ECG arrhythmias showing different patterns for each cases.

The D-QNU has been presented in section (3) for predicting ECG signals. The inputs configuration of the model (see Fig. 2) was $(n+h)$, where n stands for the number of samples of the signal used to feed the model and $h-1$ is the number of feedbacks to feed the input model, as well. The bias $\mathbf{x}(k=1)=1$. The number of training times has been set to $epochs=50$, the learning rate has been configured to $\mu=0.0001$. In addition, a window for the average of the neural weights increments has been considered as $window=180$, the initial weights of the adaptive model have been initialized as random normalized values. Also, the sensitivity parameter has been set to $\alpha=2.95$, which will be used on the step three of the adaptive methodology.

Because the adaptive model is based on RTRL, sample-by- sample adaptation, it can be observed that the neural output predicts the behavior of unstable signal of the real cardiac signals, while markers indicate the relevant changes of the sampled signal, allowing to display patterns formed in each type of cardiac arrhythmia.

Fig. 4 shows the results of the adaptive monitoring of a patient with Premature Ventricular Contraction (PVC) ECG arrhythmias in the MIT-BIH arrhythmia

database. In subplot (a) the real signal is shown with blue line color, and the predicted signal in real time monitoring is superimposed in green line color. Also, the subplot (b) presents the patterns visualization of the variability markers obtained for the PVC arrhythmia. PVC arrhythmias are highlighted into the vertical red blocks.

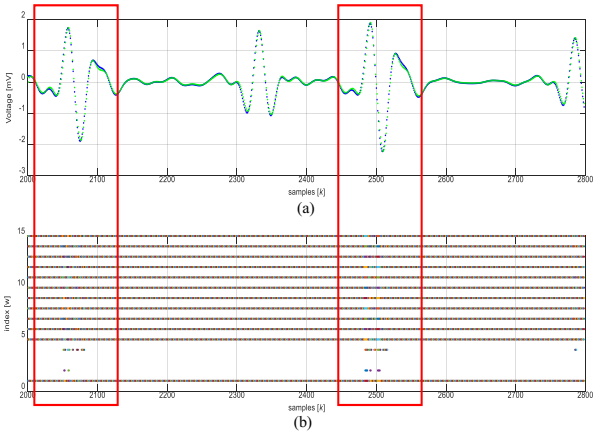


Fig. 4. Graphic of adaptive monitoring for the detection of ECG arrhythmias: (a) shows the sampled real signal (blue line color) of the HRV of a patient. The pridicted signal is shown in green line color; (b) presents the variability of the predicted signal on real time monitoring. Patterns highlighted into the signalized red blocks, clearly identify the PVC arrhythmias.

Fig. 5 presents the results of the adaptive monitoring of the ECG obtained from a patient with Atrial Premature Contraction (APC) ECG arrhythmias in the MIT-BIH arrhythmia database.

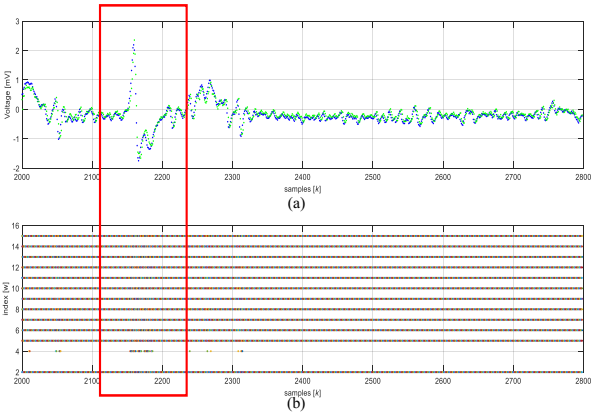


Fig. 5. Graphic of adaptive monitoring for the detection of APC ECG arrhythmia: (a) shows the sampled real signal, in blue line color, of the HRV from a patient. The predicted signal is shown in green color line; (b) presents the variability of the predicted signal on real time monitoring. The highlighted red block clearly identifies the APC arrhtyhmia variability pattern.

In subplot (a) the real signal is presented with blue color line, and the predicted signal in real time monitoring is shown in green color line. In addition, in subplot (b)

shows the variability patterns visualization with the markers obtained from the adaptive monitoring system. The APC arrhythmia is highlighted into the vertical red block in Fig. 5.

3.1 Discussion

The main contribution of this research is the prediction and monitoring of the HRV in order to provide a tool to facilitate medical diagnosis of various cardiac abnormalities in patients and therefore avoid premature death because of an improper medical diagnosis.

Because of the above, several records from the freely-available MIT-BIH database were utilized in the adaptive monitoring system for validation, which contain heartbeat records with inverse polarity, low amplitude, ventricular ectopic with low signal noise-to-ratio, premature ventricular heartbeats and premature atrial heartbeats. In Fig. 4, and Fig. 5 it can be demonstrated that the presented methodology meets the objective of identifying associated patterns with the different types of cardiac abnormalities during real-time monitoring as well as identifying related patterns to the variability of normal heartbeats. In addition, different data set was used during real-time monitoring than those used during the training phase of the system in order to monitor different data from those learned by the model. The variability markers are shown through the adaptive monitoring graphics which indicate the proper detection of the different patterns that might arise during the real-time monitoring.

The main advantages of the methodology of the D-QNU model with respect to other methodology types addressed by various authors consist that on this methodology a second order correlation with RTRL method is applied to capture nonlinear and complex behavior of the HRV. In addition, the mathematical notation remains relatively simple in comparison with other methodologies that because of their complexity require higher processing. Besides, D-QNU does not present problems with local minima during its optimization.

Currently, an automatic pattern recognition system is required, which enables an automatic classification of each cardiac abnormality shown above. In addition, a cloud service through a connection between a Matlab server and raspberry PI as data acquisition card will be carried out. With this cloud service, the cardiac arrhythmia patterns detection of the proposed model will be implemented into medical institutions.

4 Conclusions and future work

The proposed system is currently in the experimental stage, however, results of the monitoring system for prediction of cardiac arrhythmias, allow us to conclude according to the figures shown in section 3 that the adaptive model predicts the behavior of the HRV and presents the variability behavior on a monitor plot. The model of recurrent learning in real-time has proved to be able of providing the system with the necessary neural weights adaptation on which the values prediction of HRV is based. This, thanks to the algorithm of backpropagation of the error in simple-by-

sample adaptation during training. In its current stage the adaptive methodology process plots the variability markers and the patterns of variability are visualized. Future work will focus on increasing the prediction horizon without sacrificing the accuracy of the adaptive model. In addition, a pattern recognition method will be included to the monitoring system and an automatic recognition of different types of arrhythmias that may occur during signal monitoring will be included as well.

Acknowledgments. R. R. would like to thank his previous Ph.D. supervisor-specialist Doc. Ivo Bukovsky for his valuable suggestions. This work was supported by PRODEP grant No. DSA/103.5/16/10473 and by Autonomous University of Ciudad Juarez.

References

1. law Wozniak, S., Almási, A. D., Cristea, V., Leblebici, Y., & Engbersen, T. Review of Advances in Neural Networks: Neural Design Technology Stack.Proceedings of ELM-2014 Volume, 367.
2. Liu, Q., Wang, J., & Zeng, Z. (2016). Advances in Neural Networks, Intelligent Control and Information Processing. Neurocomputing, 198(C), 1-3.
3. Bataineh, M., Marler, T., Abdel-Malek, K., & Arora, J. (2016). Neural network for dynamic human motion prediction. Expert Systems with Applications, 48, 26-34.
4. Cordero, R., Suemitsu, W. I., & Pinto, J. O. P. (2015). Analysis and convergence theorem of complex quadratic form as decision boundary for data classification. Electronics Letters, 51(7), 561-562.
5. Cejnek, M., Bukovsky, I., Homma, N., & Liska, O. (2015, October). Adaptive polynomial filters with individual learning rates for computationally efficient lung tumor motion prediction. In Computational Intelligence for Multimedia Understanding (IWCIM), 2015 International Workshop on (pp. 1-5). IEEE.
6. Bukovsky, I., Homma, N., Ichiji, K., Cejnek, M., Slama, M., Benes, P. M., & Bila, J. (2015). A fast neural network approach to predict lung tumor motion during respiration for radiation therapy applications. BioMed research international, 2015.
7. Benes, P. M., Erben, M., Vesely, M., Liska, O., & Bukovsky, I. (2016). HONU and Supervised Learning Algorithms in Adaptive Feedback Control. Applied Artificial Higher Order Neural Networks for Control and Recognition, 35.
8. Rodriguez, R., Bukovsky, I., & Homma, N. (2012). Potentials of quadratic neural unit for applications. Advances in Abstract Intelligence and Soft Computing, 343.
9. Oswald, C., Cejnek, M., Vrba, J., & Bukovsky, I. (2016). Novelty Detection in System Monitoring and Control with HONU. Applied Artificial Higher Order Neural Networks for Control and Recognition, 61.
10. Moody, G. B., & Mark, R. G. (2001). The impact of the MIT-BIH arrhythmia database. IEEE Engineering in Medicine and Biology Magazine, 20(3), 45-50.
11. Ekonomou, L., Christodoulou, C. A., & Mladenov, V. (2015). An artificial neural network software tool for the assessment of the electric field around metal oxide surge arresters. Neural Computing and Applications, 1-6.
12. Bukovsky, I., Ichiji, K., Homma, N., Yoshizawa, M., & Rodriguez, R. (2010, July). Testing potentials of dynamic quadratic neural unit for prediction of lung motion during respiration for tracking radiation therapy. In The 2010 International Joint Conference on Neural Networks (IJCNN) (pp. 1-6). IEEE.

Pakistan Sign Language Recognition and Translation System using Leap Motion Device

Nosheela Raziq[1], Seemab Latif [2]
[1] Department of Computer Sofware Engineering, National University of Science and Technology, Islamabad,Pakistan
Nosheela@gmail.com
[2] Department of Computer Sofware Engineering, National University of Science and Technology, Islamabad,Pakistan
seemab@mcs.edu.pk

Abstract. The recent innovation in the technology and introduction of new gesture control based devices, like leap motion controller, have revolutionized computing. Technology that can track finger and hand gestures accurately at very low cost can reshape the gesture control industry. Leap motion controller, a little USB device, has an ability to obtain accurate description of the hand and figures pose that can be used for Recognition of Pakistan Sign Language. Sign Language is very crucial to facilitate deaf and mute children to communicate with the society. This paper proposes a gesture based recognition system for Pakistan Sign Language using leap motion device. Our proposed system compromises of two modules training module and communication module. In training module motion data received from leap motion controller is used to train the system for Pakistan Sign Language. Than in communication module motion data is acquired through leap motion controller and fed into correlation algorithm to detect and recognize the sign and then convert it into text form. The proposed system provides an efficient and cost effective solution to communication problem of deaf and mute children.

1 Introduction

As today's technology is all about mobile computing, gesture based environments and cloud computing. Different smart phones have voice recognition system, touch system as well as different laptops have touch and voice recognition system. World is leaping into gesture control era. Different companies like Microsoft, Google and leap motion working on gesture control based environments introducing devices like Kinect, Google Glass and Leap Motion controller. Therefore this advancement in technology can be used for the benefit of deaf and mute children.

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_87

When it comes to deaf children, communication with others is far longer struggle for them. They are unable to communicate with normal persons properly. They can't blend with society due to their inability to communicate normally. Unwillingly Deaf and Mute individuals and children are unable to show good performance in many areas of life. When it comes to education, the learning environment for deaf and mute children is not similar as the normal child. They face difficulties in finding jobs and living normal life like others.

The only way for them to communicate is through Sign Language. But their sign language is only understood by sign language experts or their fellows. It creates a huge gap between deaf and normal person. Sign Language is a gesture based language composed of different hand and finger postures. Different signs are made up of hand and combining fingers together.

In existing frameworks and research for Sign Language Translation and recognition system both image based and sensor based approaches are used. In image based approaches, cameras struggle to deal with different tricky background or illumination and distance issues. In senor based approaches, for example, data gloves sign language detection and recognition achieves good results but has its own challenges like wearing data gloves and also a very expensive device to popularize. Latest research focuses on Hand gesture recognition systems due to its application in HCI, Robotics, Game Based Learning and sign language recognition applications and systems. Different approaches and algorithms from computer vision community have been used. Then Microsoft Kinect cameras have been used for gesture recognition. Machine learning techniques have been used mostly to recognize the gestures.

Therefore, with invention of new device Leap Motion Controller by leap Motion Company the tracking of hand and figure postures can be done accurately at very low cost.

In this paper we propose a gesture based recognition system for Pakistan Sign Language using device Leap Motion Controller. By using Leap motion controller hand and finger movement's recognition will be done and then by using motion track data of leap motion controller, system will be trained according to Pakistan Sign Language. The communication module of proposed system will acquire data through leap motion controller and will apply correlation pattern recognition algorithm on that data to recognize the sign and will then display it on the screen in the form of text.

The structure of our paper is as follow. Section II discusses the related work of Pakistan Sign language recognition system. Section III describes the framework and algorithm for proposed system. Section IV explains Implementation of application. Section V explains the Experimentation and Results. Sections VI discusses the Conclusion and Future Work.

2 Related Work

According to Bushra and Bashir [1]'s estimation 2.49% of the population is disabled. Out of this 7.40% disabled population are deaf. In Pakistan, children with disabilities have been educated in segregated special schools. These special schools are reason behind their isolation with their normal peers. This segregated system has restricted all their possibilities of social interaction and detaches them from daily educational and social emotional experiences of normal children.

According to Husain [2], in Pakistan very rare schools have got professional trainers that include special education teachers, psychologists, physiotherapists and speech therapists. Hearing aids are quite expensive and not everyone is able to afford these hearing aids [3]. According to Saeed and Majeed [4] most of the trainers and Principals are not qualified enough to educate these children. They don't have relevant field experience.

Marin et.al [5] have discussed their approach for detection of hand gestures by using two innovative gesture control based devices leap motion controller and Microsoft Kinect. They divided hand identification into 5 vectors which are divided into two sets, 3 in set of leap motion and the other 2 in Kinect and then fed these computed sets into SVM classifier to recognize the performed gestures. They have done the comparison between two devices which shows the accuracy of each device recognizing the gestures. According to them by combining both device's feature sets, high accuracy recognition system can be achieved in real time.

Chai et.al [6] in their paper presented Kinect based sign language recognition system. According to them existing methods using data gloves for recognition gives the better result for large sign vocabulary but it is too costly to popularize. They have presented a system in which 3d trail was to be matched for a gesture. The system is based on two modes. . In "Translation Mode", the gestures were recognized as words and sentences and in "Communication Mode" an avatar transform Sign Language Recognition into texts so that the normal hearing people could easily understand.

Raees et.al [7] presented an image based Pakistan Sign Language Recognition system. Input data from thumb is extracted by applying Template matching technique while recognition of rest of the four fingers has been done through deep pixels-based analysis. For recognition of signs a model has been made including seven phases identified by algorithm. The proposed system achieved accuracy level of 84.2% with 180 digits and 240 alphabets of Pakistan Sign Language.

3 Pakistan Sign Language Recognition and Translation System

Detailed analysis of the literature review and gathered requirements highlighted the usability requirements of proposed system. Accordingly requirements research methodology has been defined. This research will overcome the communication problem of deaf and hearing impaired children in Pakistan. This application is

specifically designed for children under the age of 15. The system can be used both in school and at home.

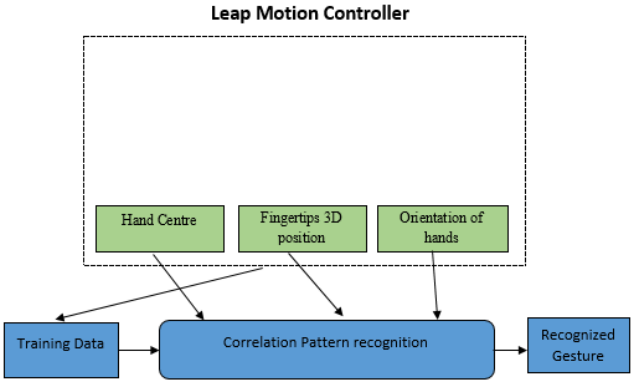


Fig. 1. Simple Pipeline of Proposed System

Fig.1 describes the simple pipeline of Proposed System. The approach that has been used is introduction of new Gesture recognition device Leap Motion Controller (LMC) as our input interface to get data from fingers positions, hands and palm movements for performing signs. Motion track data acquired from LMC is then integrated with the system that after applying correlation pattern recognition algorithm, recognize the gesture and display it on the screen.

2.1 Architecture

Leap motion controller acquires data from hand and figure while performing signs and then utilize acquired data for training and detecting signs.

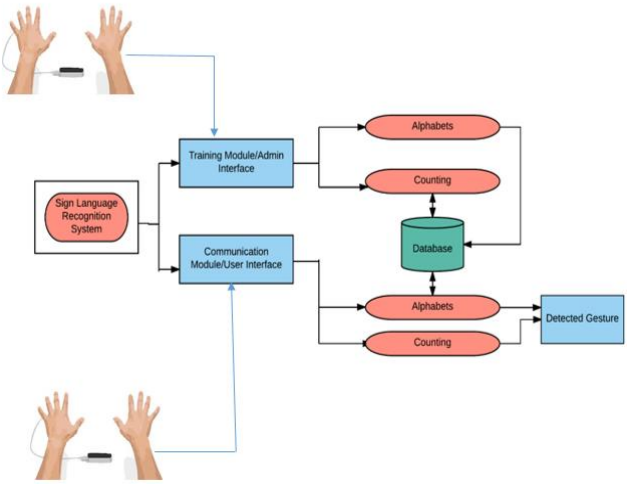


Fig. 2. Architecture of Proposed System

Fig.2. shows the main architecture of Proposed System. Application comprises of two main Modules. These Modules are associated with the training of Pakistan Sign language and then detecting the signs and displaying the matching sign as an output on the main screen.

Training/Admin Module

This module focuses on training part. User/Admin trains the system according to Pakistan Sign language's needs. Different signs can be trained and stored into database from the data acquired by Leap Motion Controller for further use. User can train new signs, also it can remove and modify the signs. User can have three tries to train the sign once he is sure he can save the sign in database. Artificial neural network algorithm has been used of training the data sets.

User/Communication Module

This module presents an option to make a sign, when user makes the sign with the help of leap motion device, in the backhand Correlation Pattern Recognition Algorithm has been used to match and detect the Signs, the matched sign are then displayed on the screen in the form of text. User interface also provides the facility to record your signs in the form of video while pressing recording button.

4 Implementation

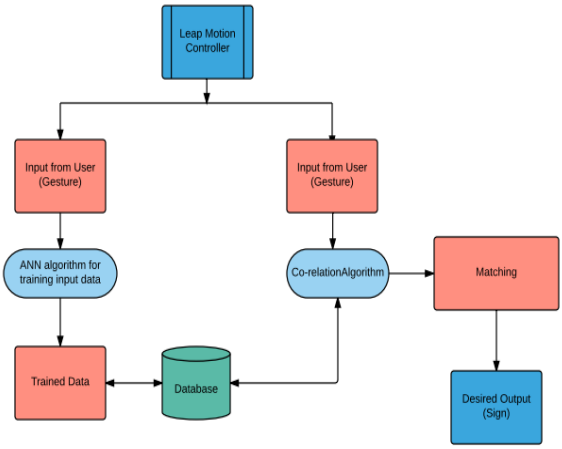


Fig. 3. Schematic Presentation of System

4.1 Data Captured

Fig.3. highlights the schematic presentation of proposed system. A program was written in Java for gathering of data. It was then integrated with the main system. Main GUI allows the user to record the data. Artificial Neural Network algorithm has

been used to train the data into Pakistan Sign Language, 26 Alphabets of Pakistan Sign Language has been recorded and trained. For each sign user needs to train the sign total of three times.

4.2 Pattern Recognition

Leap motion controller's API is integrated with the main system. For recognition of PSL Correlation pattern recognition algorithm have been used. In the main interface or communication interface the input data is captured through leap motion controller, pattern recognition algorithm correlates the input data with trained data and in case of matching give the desired output. Fig.4 and Fig.5 shows the main and Training interface of the proposed system.

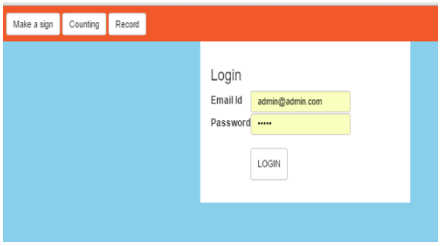


Fig. 4. Main Interface of System

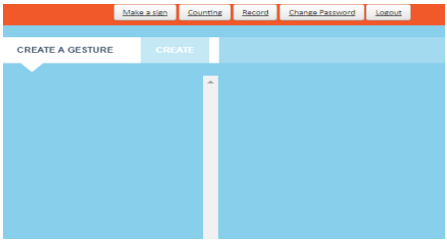


Fig. 5. Training of Signs

5 Experiments and Results

For testing the effectiveness of Application confusion matrix approach has been used. The proposed method using correlation classification algorithm managed to classify 1400 out of 1500 samples correctly. Confusion matrix is created to determine the misclassification cases of 6 PSL alphabets. For 6 PSL Alphabets A, B, C, D, E, F 20 samples of each alphabet has been collected. These 20 samples are then compared with trained set of data. Four possibilities of classifier (TP, FP, TN, and FN) have been found for all the letters. Then overall accuracy, precision and recall of the system has been measured. Then for each letter accuracy, precision and recall has been calculated separately. Therefore the summary of performance measures for 6 PSL Alphabets are shown in Table 1.

Table 1. Performance measures of proposed system

Performance Measures %	A	B	C	D	E	F
Accuracy	94.9%	98%	95.7%	98.2%	99.1%	98.2%
Precision	81.8%	95%	90%	95%	100%	100%
Recall	90%	95%	85.7%	95%	95%	90%
F-Score	85.71	95%	87.8%	95%	97.4%	94.7%
Overall Accuracy of whole system=92.5%						

Table 5 shows that for 6 PSL Alphabets overall accuracy of 92.5% have been achieved. While for each letter different quality measures like accuracy and precision has been measured.

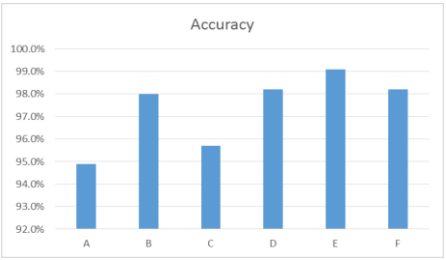


Fig. 6. Accuracy Measure

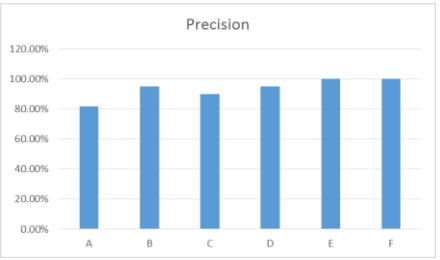


Fig. 7. Precision Measure

After undergoing different experiments the system has shown 92.5% average recognition rate. Most classes of alphabets have 0 to 1 misclassification. Fig.6 and fig.7 shows the accuracy and precision measure for 6 Alphabets. Letter A, C, and F have more misclassification cases as the accuracy they have achieved is less than letters B, D and E. These three alphabets are misclassified due to similar shapes and camera orientations. Some misclassifications were faced because of leap motion controller's limitations of detecting closed fingers.

6 Conclusion and Future Work

The experimental results revealed that application has ability to transform the communication level for deaf and mute children. This desktop application can enable them to talk to normal person. The proposed application can provide a cost effective solution to their communication problem. The main focus of this proposed system was on PSL Alphabets. The system was able to achieve 92.5% accuracy for 6 Alphabets of PSL. But has potential to recognize all the 26 letters. Some of the letters that have same shapes of signs were misclassified. But by altering the light conditions and changing the device position more accurate results can be produced. That proposed system can be enhanced further. The devised algorithm can be altered for PSL (Urdu).

References

1. Akram, Bushra, and Rukhsana Bashir. "Special Education and Deaf Children in Pakistan: An Overview." Journal of Elementary Education 22.2 (2012): 33-44.

2. Husain, Z. (2003). A study of self-image of deaf community in Lahore. Unpublished Master's Thesis: Department of special education; University of Punjab, Lahore.

3. Khatoon, A. (2003). A historical and evaluative study of special. Doctoral Thesis: Special Education Department, University of Karachi, Karachi.
4. Majeed, Z., & Saeed, A. (1998). Evaluation of the audiological services available in special schools in Pakistan: Journal of Special Education, 1 (1), 33-40.
5. G. Marin, F. Dominio and P. Zanuttigh, "Hand gesture recognition with leap motion and kinect devices," 2014 IEEE International Conference on Image Processing (ICIP), Paris, (2014), pp. 1565-1569. doi: 10.1109/ICIP.2014.7025313
6. Chai, Xiujuan, et al. "Sign language recognition and translation with kinect." IEEE Conf. on AFGR. (2013).
7. Raees, Muhammad, Sehat Ullah, Sami Ur Rahman, and Ihsan Rabbi. "Image based recognition of Pakistan sign language." Journal of Engineering Research 4, no. 1 (2016).

Identifying stable objects for accelerating the classification phase of k-means

A. Mexicano[1], S. Cervantes[2], R. Rodríguez[3], J. Pérez[4], N. Almanza[4], M. A. Jiménez[1], A. Azuara[1]

[1] Technological Institute of Ciudad Victoria, Cd. Victoria, Tamaulipas
mexicanoa@gmail.com

[2] University Center of Ciudad Valles, Guadalajara University
Ciudad Valles, Jalisco

[3] Autonomous University of Ciudad Juarez, Ciudad Juarez, Chihuahua
ricardo.jorge@uacj.mx

[4] National Centre of Research and Technological Development,
Cuernavaca, Morelos
jpo_cenidet@yahoo.com.mx

Abstract. Abstract—This work presents an improved version of the K-Means algorithm, this version consists in a simple heuristic where objects that remains in the same group, between the current and the previous iteration, are identified and excluded from calculi in the classification phase for subsequent iterations. In order to evaluate the improved version versus the standard, three synthetic and seven well-known real instances of specialized literature were used. Experimental results showed that the proposed heuristic spends less time than the standard algorithm. The best result was obtained when the Transactions instance was grouped into 200 clusters, achieving a time reduction of 90.1% regarding the standard version, with only a grouping quality reduction of 3.97%.

1 Introduction

Grouping is the process of assigning a set of objects into subsets, named groups, such that, objects in the same group are similar and different to the rest of objects in other groups [1][2]. This process has been studied long time ago, because grouping can be applied in a wide variety of domains. For example in computational science has been applied in image segmentation, machine learning, artificial intelligence, pattern recognition; in the science of life has been applied in genetics, biology, microbiology, paleontology, psychiatry, pathology; in earth sciences has been applied in geography, and geology; in social sciences, in sociology, psychology, archeology, education; amongst others [3][4][5][6]. In addition, the emergence of Big Data requires the development of efficient methods or technologies to analyze and process data [7], being the clustering a highly used method in order to reduce data dimensionality or the cost of search in data for example [8]. In specialized literature

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_88

there are many grouping algorithms, however, the K-Means algorithm [9] is one of the most used, because its ease of implementation [10]. Nevertheless, one of the drawbacks of K-Means is its high computational cost due to the number of calculations that the algorithm requires in the classification phase. In this sense, this work presents a new heuristic for accelerating the classification phase of the K-Means algorithm. The heuristic is based on the idea that major changes in clusters occurs into the first iterations, then objects into clusters begin to stands into the same cluster iteration after iteration. The idea is to identify the objects that remain stables between the current iteration and the past one, in order to remove them for subsequent calculus in the classification phase. It allows reducing the execution time of the algorithm. Several tests were executed in order to define the number of iterations after applying the proposed heuristic. Experiments with 2, 4, 6, 8, and 10 iterations after applying the proposed heuristic were conducted. The election of applying the heuristic after the eight iteration was done considering that the quality of grouping is not affected significantly regarding the standard version.

The paper focuses on improving the performance of the K-Means algorithm in terms of reducing the number of calculations of the algorithm, without affecting significantly the quality of grouping.

The rest of this article is structured as follows, Section II presents a briefly description of related work, Section III shows the main idea for reducing the number of calculations in the classification phase of the K-Means algorithm, Section IV presents the conducted tests in order to validate the proposed heuristic; finally, Section V show the conclusion and future work.

2 Related Work

In specialized literature, amount of work is devoted to improve the performance of the K-Means algorithm offering different approaches. Improvements can be classified according to the one the next three phases of the algorithm: initialization, classification or convergence. In the initialization phase, the objective has been to improve the cluster quality of the algorithm being this one of its drawbacks, some approaches are presented in [11], [12], [13], [14], [15], [16], and [17]. In the convergence phase, improvements are oriented to reduce the execution time of the algorithm without affecting significantly the quality of the grouping, examples in this area are [18], [19], and [20]. Finally, in the classification phase there is amount of work, because in this phase the calculi of each object to each centroid are required. In this sense, there are approaches that identify objects with high possibility of cluster change [21], [22], [23], and [24], or define thresholds to identify areas of clusters that can be modified after an specific iteration [25] and [26] or identify stable clusters which are removed from classification phase avoiding computations of whole clusters in subsequent iterations [27] and [28]. In contrast to the related work, the presented heuristic consists in identifying the objects that after a given iteration remains in the same cluster in order to remove them from calculi in the classification phase.

3 Identifying stable objects in the classification phase

The proposed heuristic is simple, it consists in identifying objects that after a given iteration, they remain in the same cluster, these objects are called as stable objects. Each time that an object is identified as stable, the object is removed from the classification phase in subsequent iterations. Fig. 1 shows a motivation example, where an execution of the proposed approach is presented, in the example four iterations are required in order to get the final grouping. However, if stable objects are detected after second iteration, it is possible to observe that 35 of 36 objects remains in the same cluster (objects rounded by a circle at iteration 3) leaving only one for taking it in account in the four iteration. It means that a high reduction in the number of object-centroid calculations can be applied; in the example, 105 calculations are omitted in the last iteration. In Fig 1, the iteration 4 shows that only the calculi from one object to the three centroids are required. The identification of stable objects contributes in accelerating the stopping criterion when it is related to the changes in clusters.

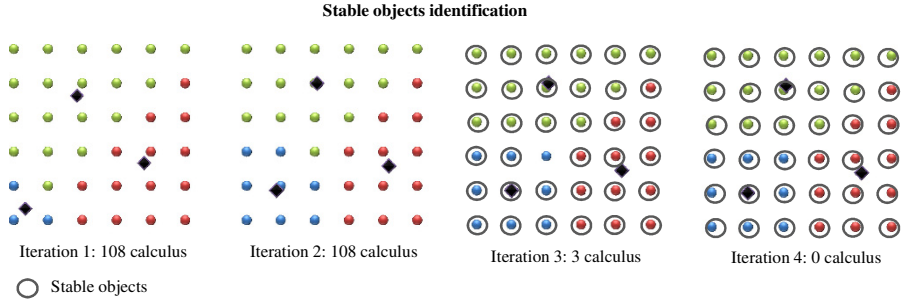


Fig. 1. Motivational example using a two dimensional synthetic instance. The example shows the execution of the proposed heuristic, applied after iteration 2. In this case, in the iteration three, 35 objects are identified as stables (objects circled in gray) and at iteration four, only an object-centroids computation is required, when iteration 4 finishes, all the objects are identified as stables and the algorithm stops.

Due to the ease of this heuristic, only it was needed to verify, after a given iteration; if each of the objects; after classification phase, remains or not in the same cluster. When an object reminded in the same previous cluster, it was marked as stable using a simple flag, for avoiding calculating the distance from this object to the centroids. After several experiments, using real and synthetic datasets, it was noted the convenience of applying the heuristic after the eight iteration, in order to do not affect significantly the quality of the grouping, regarding the standard version. The incorporated condition was based on the idea that the major movements between centroids and clusters occurs in the first iterations of the K-Means algorithm, after that, many of the objects in the grouping process remains in the same cluster, because their high similarity. Then, identifying these stable objects is able to discard the items for calculations in subsequent iterations at classification phase. It is important to say that objects discarded from the classification phase are not discarded in the rest of phases of the algorithm.

3 Computational experiments

In order to compare the proposed version versus the standard, both algorithms were implemented in C language, a computer with Intel (core) i7 processor, 3.5 GHz, 8 GB RAM and Ubuntu 14.04 operative system was used. Seven sets of instances were used for experimentation, three synthetic datasets (2500U, 10000U, and 40000U) with 2,500, 10,000, and 40,000 2-dimensional objects distributed uniformly. The *Skin segmentation* real dataset extracted from UCI Machine learning [29], with 245,057 3-dimensional objects. The *Transactions* dataset, extracted from KEEL-dataset [30], with 284,284 3-dimensional objects. *Paris* and *New York* datasets with 414,528 and 657,308 2-dimensional objects, extracted from Flickr [31]. All instances were grouped into 100, 200, 400, and 800 clusters. Results are the average of 30 executions using random centroids at each time.

Tables I to IV show the obtained results from executing both, the standard and the proposed version. In tables, the most representative results are highlighted in bold. Tables I and II present results in terms of execution time. The first column corresponds to the evaluated dataset. The second column corresponds to the cluster size. The third column represents the execution time required for each experimentation, this column is divided into three subcolumns where the results for the standard algorithm and the proposed version are presented; subcolumn $\tau\%$ corresponds to the percentage of time reduction achieved by the proposed version (expressed in milliseconds) regarding the standard K-Means (see eq. 1).

$$\tau\% = \frac{(T_s - T_h)\times 100}{T_s} \tag{1}$$

Where T_s denotes the average running time of the standard K-Means and T_h the average running time of the proposed version.

Table I shows the results obtained for synthetic instances. For the 40000U instance, in all cases, the instance reached the best results, being the best a reduction in execution time of 89.96%, regarding to the standard version, when the instance was grouped into 100 clusters. Fig. 2.a shows graphically the obtained results in Table I, the reduction time is better perceived with the growth of the instances.

Table 1. Execution time after evaluating the standard versus the proposed version for Synthetic datasets.

Dataset	Cluster size	Time (ms)		
		K-Means	*Proposed Heuristic*	τ %
2500U	100	160.46	70.36	56.15
10000U		1800.01	253.54	85.91
40000U		**10013.72**	**1005.46**	**89.96**
2500U	200	229.76	128.66	44.00
10000U		2546.99	493.81	80.61
40000U		**19363.19**	**1971.19**	**89.82**
2500U	400	278.04	231.69	16.67

10000U		2883.94	977.43	66.11
40000U		**34519.33**	**3904.99**	**88.69**
2500U		372.74	329.95	11.48
10000U	800	3909.05	1940.87	50.35
40000U		**46614.85**	**7740.09**	**83.40**

Table II presents the results for Real instances. In this case, the best result was obtained for the *Transaction* dataset when 200 and 800 clusters were generated; the reduction time reached was of 90.10% in both cases. However, in the rest of experiments, reductions of 90.01 and 90.07% were obtained, when the instance was grouped into 100 and 400 groups, respectively. It is important to note that using real instances, the less reduction time achieved was of 84.24%, for the *New York* Instance, when it was grouped into 100 clusters. In all experiments, the *New York* instance achieved the less reduction time. Fig. 2.b shows graphically the results in Table II, in all cases the proposed version outperforms the standard.

Table 2. Execution time after evaluating the standard versus the proposed version for Real datasets.

Dataset	**Cluster size**	**Time (ms)**		
		K-Means	***Proposed Heuristic***	***τ %***
Skin	100	68230	8643.12	87.33
Transactions		**100344.6**	**10019.99**	**90.01**
Paris		73884.67	10416.94	85.90
New York		105083.84	16558.6	84.24
Skin	200	154390.36	17035.37	88.97
Transactions		**199065.4**	**19703.99**	**90.10**
Paris		140926.74	20442.19	85.49
New York		211691.95	32401.62	84.69
Skin	400	324096.91	33803.14	89.57
Transactions		**393022.55**	**39038.6**	**90.07**
Paris		287024.28	40462.25	85.90
New York		463844.02	64129.79	86.17
Skin	800	664166.96	67197.67	89.88
Transactions		**784311.02**	**77636.62**	**90.10**
Paris		581785.18	80286.9	86.20
New York		854867.63	127192.47	85.12

Tables III and IV present results in terms of quality of grouping, using the Mean Squared Error as metric (see eq. 2), third column is divided into three subcolumns where results for the standard, the proposed version, and the percentage of reduction in the quality of grouping , are presented. The percentage of reduction in the quality of grouping ε% was calculated using the eq. 3.

$$E = \sum_{l=1}^{k} \sum_{i_j \in C_l} \left\| i_j - \mu_l \right\|^2 \tag{2}$$

where E stands for the Mean Squared Error, $I = \{i_1, \ldots, i_n\}$ is the set of objects, $C = \{C_1, \ldots, C_k\}$ is the set of clusters, and μ_l is the mean of objects in C_l.

$$\varepsilon\% = \frac{(\mathrm{E}_h - \mathrm{E}_s)\times 100}{\mathrm{E}_s} \tag{3}$$

where E_s denotes the mean squared error of the standard K-Means and E_h the mean squared error of the proposed version.

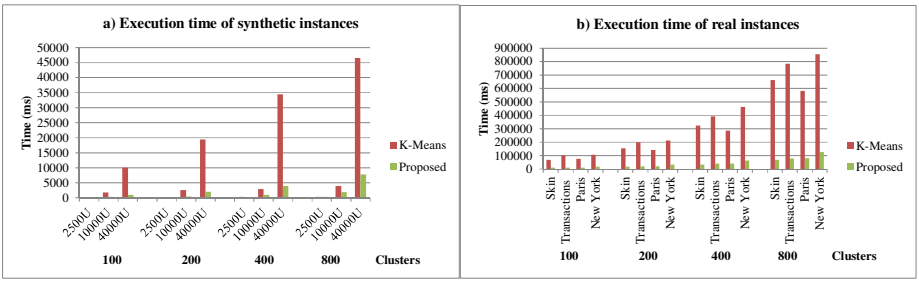


Fig. 2. Execution time comparison between K-Means versus the proposed version for: a) synthetic and b) real instances.

Table III shows that for synthetic instances the quality of the grouping was not significantly affected, the highest reduction of quality was obtained for the 40000U in all experiments, this instance achieved a 1.99, 1.77, 1.35 and 0.91% of reduction when was grouped into 100, 200, 400, and 800 clusters, respectively. For instance 2500U only a 0.09% in reduction of quality was obtained when it was grouped into 400 clusters, being this the less reduction reached in all the experiments. Fig. 3 shows a graph where the mean squared error obtained for the synthetic instances, in all cases is possible to observe that the quality is not affected significantly.

Table 3. Comparative results evaluating the standard versus the proposed version for Synthetic datasets.

Dataset	Cluster size	Mean Squared Error		
		K-Means	*Proposed Heuristic*	*ε %*
2500U	100	160.46	70.36	56.15
10000U		1800.01	253.54	85.91
40000U		**10013.72**	**1005.46**	**89.96**
2500U	200	229.76	128.66	44.00
10000U		2546.99	493.81	80.61
40000U		**19363.19**	**1971.19**	**89.82**
2500U	400	278.04	231.69	16.67
10000U		2883.94	977.43	66.11
40000U		**34519.33**	**3904.99**	**88.69**
2500U	800	372.74	329.95	11.48
10000U		3909.05	1940.87	50.35
40000U		**46614.85**	**7740.09**	**83.40**

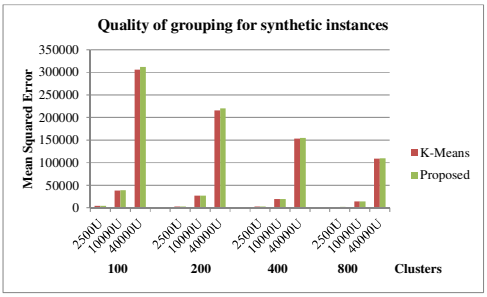


Fig. 3. Execution time comparison between K-Means versus the proposed version for real instances.

In Table IV results for real instances regarding cluster quality shows that in general the reduction quality is in the range of 1.96 to 14.33%. The worst result occurred when the *Paris* instance was grouped into 100 clusters, despite its reduction in execution time of 85.9%; the instance achieved a 14.33 in quality reduction. For the *New York* instance, the reduction in quality was of 9.44% when it was grouped into 400 clusters; and of 8.93 and 7.32% when 800 and 200 clusters were generated. The instance, which in all the experiments obtained the less quality reduction, was *Transactions,* obtaining 2.92, 2.85, 1.93, and 3.97% when the instance was grouped into 100, 200, 400, and 800 clusters, respectively.

Table 4. Comparative results evaluating the standard versus the proposed version for Synthetic datasets.

Dataset	Cluster size	Mean Squared Error		
		K-Means	***Proposed Heuristic***	***ε %***
Skin	100	2323473.35	2368970.33	1.96
Transactions		**38857629.34**	**39994101.62**	**2.92**
Paris		1676.23	1916.38	14.33
New York		5448.46	5829.16	6.99
Skin	200	1754980.92	1835018.72	4.56
Transactions		**27192270.99**	**27967311.36**	**2.85**
Paris		1251.06	1330.1	6.32
New York		4030.46	4325.41	7.32
Skin	400	1360522.39	1438305.75	5.72
Transactions		**18722741.45**	**19084099.08**	**1.93**
Paris		898.02	955.44	6.39
New York		3027.2	3312.85	9.44
Skin	800	1104532.16	1172307.96	6.14
Transactions		**11741893.65**	**12208240.23**	**3.97**
Paris		645.93	675.28	4.54
New York		2329.87	2537.96	8.93

Due to the differences into values of mean squared errors, graphs were divided into *Skin* and *Transactions* (Fig. 4.a), and *Paris* and *New York* (Fig. 4.b). In both graphs, the tendency in quality is the same. In the case of *Paris* instance, despite the 14.33%

in quality reduction the graph shows that the tendency also is similar to the standard K-Means.

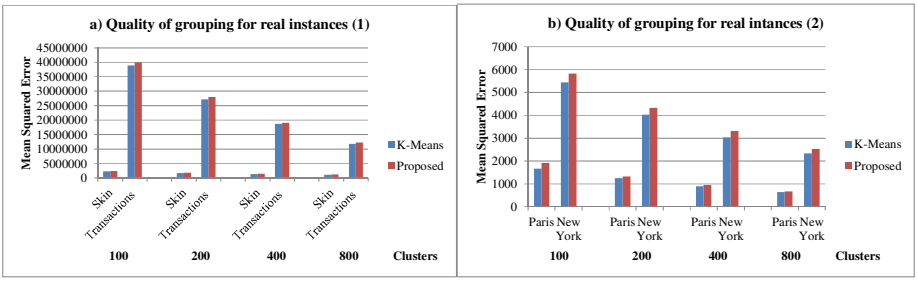


Fig. 4. Grouping quality comparison between K-Means versus the proposed version for instances: a) Skin and Transactions, and b) Paris and New York.

Computational experiments and graphs show that identifying stable objects allows reducing the execution time of the standard version for synthetic and real datasets. The proposed algorithm reaches a reduction time up to 90.10%. Despite the high reduction time, it is important to aim towards the quality reduction, in this sense, the use of a threshold for determining the consolidation of each cluster will be proposed in future work.

3 Conclusion

Due to the necessity of the development of better methods or processing large amounts of data, in this work, a new heuristic for reducing the execution time in the classification phase of the K-Means algorithm is presented. The heuristic consists in identifying objects that in the actual and in the previous iteration remain grouped into the same cluster. After several experiments it was noted the convenience of applying the heuristic after the eight iteration, in order to do not affect significantly the quality of the grouping. The incorporated condition was based on the idea that the greatest movements between centroids and clusters occur in the first iterations of the K-Means algorithm. After that, many of the objects in the clustering process remains in the same cluster because their high similarity, then identifying these objects is able to discard items avoiding calculating the distances object-centroids in subsequent iterations. Computational experiments show for synthetic instances the proposed version reduced up to 89.96% the time required by K-Means with only a 1.99% in quality reduction, when 100 clusters were generated using the 40000U instance, however this instance achieved the best results in the four different experiments (100, 200, 400, and 800 clusters). In the case of real datasets the Transactions instance as the 40000U obtained the best results in all the experiments; the achieved reductions on time and quality were 90.01, 90.10, 90.07, and 90.10%; and 2.92, 2.85, 1.93, and 3.97; respectively. The worst result was obtained using the Paris instance when generating 200 clusters, despite reaching an 85.9% of time reduction, this instance reduced in 14.33% the quality of grouping. In future work we try to find a threshold that helps to reduce the loss of quality and apply the new version in image segmentation.

References

1. Jain, A., and Dubes, R.: Algorithms for Clustering Data, Prentice Hall, Englewood Cliff, Nueva Jersey (1988)
2. Junjie, W.: Advances in K-Means Clustering A Data Mining Thinking, Doctoral Thesis, Tsinghua University, China, Springer (2012)
3. Scoltock J.: A survey of the literature of cluster analysis," The Computer Journal, 25 (1982) 130-134
4. Al-Zoubi, B., Hudaib, A., Huneiti, A., and Hammo, B.: New Efficient Strattegy to Accelerate K-Means Clustering Algorithm, American Journal of Applied Sciences, 5:9 (2008) 1247-1250
5. Xu, R. and Wunsch II, D.: Survey of clustering algorithm, IEEE Transactions on Neural Networks, 16:3 (2005) 645-678
6. Everitt, B. S., Laudau, S., Leese, M., and Stahl, D.: Cluster Analysis. John Wiley and Sons, Inc., London, United Kindom (2011)
7. Chen, M., Mao, S., Zhang, Y., Leung, V.: Big data: related technologies, Challenges and future prospects, Springer (2014)
8. Li, K.C., Jiang, H., Yang, T. L.: Big Data: Algorithms, Analytics, and Applications, CRC Press Taylor and Francis Group, New York (2015)
9. MacQueen, J.: Some Methods for Classification and Analysis of Multivariate Observations, in Fifth Berkeley Symposium on Mathematics, Statistics and Probability, University of California Press, Berkeley, Calif., (1967) 281–296
10. Wu, X., Kumar, V., Quinlan, J.L., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B, Yu, P.S., Zhou, Z., Steinbach, M., Hand, D.J., and, Steinberg, D.: Top 10 algorithms in data mining, Journal of Knowledge and Information System, 14 (2008) 1-37
11. Khan, S. S., and Ahmad, A.: Cluster center initialization algorithm for K-Means clustering, Pattern Recognition Letters, 25 (2004) 1293–1302
12. Redmond, S. J. and Heneghan, C.: A method for initializing the K-Means clustering algorithm using kd-trees, Pattern Recognition Letters, 28: 8 (2007) 965–973
13. Zalik, K. R.: An efficient K-Means clustering algorithm, Pattern Recognition Letters, 29 (2008) 1385–1391
14. Li, C. S.: Cluster Center Initialization Method for K-Means Algorithm over Data Sets with Two Clusters, Procedia Engineering, 24, (2011) 324–328
15. Eltibi, M. F. and Ashour, W.M.: Initializing K-Means Clustering Algorithm using Statistical Information, International Journal of Computer Applications, 29:7 (2011) 51–55
16. Agha, M. E. and Ashour, W. M.: Efficient and Fast Initialization Algorithm for K-Means Clustering, International Journal of Intelligent Systems and Applications, 1:1 (2012) 21–31
17. Kaur, N., Kaur, J., and Kaur, N.: Efficient K-Means clustering algorithm using ranking method in data mining, International Journal of Advanced Research in Computer Engineering & Technology, 1:3 (2012) 85–91
18. Perez, J., Pazos, R., Cruz, L., Reyes, G., Basave, R. and Fraire, H.: Improving the Efficiency and Efficacy of the K-Means Clustering Algorithm through a New Convergence Condition, in Computational Science and Its Applications - ICCSA, Kuala Lumpur, Malaysia (2007) 674–682.
19. Yu, S., Tranchevent, L. C., Liu, X., Glänzel, W., Suykens, J. A. K., Moor, B.D., and Moreau, Y.: Optimized Data Fusion for Kernel K-Means Clustering, IEEE Transactions on Pattern Analysis and Machine Intelligence, 34:5 (2012) 1031–1039
20. Mexicano, A., Rodríguez, R., Cervantes, S., Montes, P., Jiménez, M., Almanza, N. and Abrego, A.: The early stop heuristic: A new convergence criterion for K-means, in AIP Conf. Proc. 1738, ICNAAM2015, Rhodes Greece (2016) 3100031–3100314.
21. Lai, J.Z.C. and Liaw, Y.C.: Improvement of the K-Means clustering filtering algorithm, Pattern Recognition, 41(2008) 3677–3681

22.Fahim, A. M., Salem, A. M., Torkey, F. A. and Ramadan, M. A.: An Efficient Enhanced K-Means Clustering Algorithm, Journal of Zhejiang University-Science, 7:10 (2006) 1626–1633
23.Sheeba, A., Mahfooz, S., Khusro, S. and Javed, H.: Enhanced K-Mean Clustering Algorithm to Reduce Number of Iterations and Time Complexity, Middle-East Journal of Scientific Research, 12:7 (2012) 959–963
24.Pérez, J., Martínez, A., Almanza, N., Mexicano, A., and Pazos, R.: Improvement to the K-Means algorithm by using its geometric and cluster neighborhood properties, in Proceedings of ICITSEM 2014, Dubai, UAE (2014) 21–26.
25.Pérez, J., Pires, C. E., Balby, L., Mexicano, A. and Hidalgo, M.: Early Classification: A New Heuristic to Improve the Classification Step of K-Means, Journal of Information and Data Management, 4:2 (2013) 94–103
26.Mexicano, A., Rodriguez, R., Cervantes, S., Ponce, R. and Bernal, W.: Fast means: Enhancing the K-Means algorithm by accelerating its early classification version, in AIP Conf. Proc. 1648, ICNAAM2014, Rhodes Greece (2015) 8200041–8200044
27.Pérez, J., Pazos, R., Hidalgo, M., Almanza, N., Díaz-Parra, O., Santaolaya, R., and Caballero, V.: An improvement to the K-Means algorithm oriented to big data, in AIP Conf. Proc. 1648, ICNAAM2014, Rhodes Greece (2015) 8200021–8200024
28.Pérez, J., Pazos, R., Olivares, V., Hidalgo, M., Ruiz, J., Martínez, A., Almanza, N., and González, M.: Optimization of the K-Means algorithm for the solution of high dimensional instances, in AIP Conf. Proc. 1738, ICNAAM2015, Rhodes Greece (2016) 3100021–3100214
29.Merz, C., Murphy, P., and Aha, D.: UCI Repository of Machine Learning Databases. Department of Information and Computer Science, University of California, http://www.ics.uci.edu/mlearn/MLRepository.html, 2016.
30.http://sci2s.ugr.es/keel/datasets.php, Knowledge Extraction based on Evolutionary Learning, KEEL-dataset, last view: July 2016.
31. http://www.flickr.com/map/,Photography's repository, last view: July 2016.

Expert System To *Engage* CHAEA Learning Styles, ACRA Learning Strategies and Learning Objects into an E-Learning Platform for Higher Education Students

José Angel Montes Olguín, Francisco Javier Carrillo García, Ma. de la Luz Carrillo González, Antonia Mireles Medina, Julio Zenón García Cortés
Instituto Tecnológico Superior Zacatecas Norte
anxelm@gmail.com, carrillogarciafj@yahoo.com.mx, luzcarrillog@gmail.com, mirelesmed_7@hotmail.com, profejulioz@hotmail.com

Abstract. The present work proposes to increase a student learning style on which he is weaker and to accelerate his learning process through a solution based on an expert system (ES) to engage CHAEA learning styles, ACRA learning strategies and learning objects (LO), into an e-learning platform for higher education students. The main expert system activity is to select a learning objects block that allows shaping a learning strategy and present this strategy in an e-learning platform, taking as a reference the CHAEA questionnaire. To develop our solution we use set theory and conceptual graphs (CG) for designing the knowledge base (KB) and the inference engine (IE), to solve learning objects interoperability we take the "Norma mexicana para la interoperabilidad entre entornos de objetos de aprendizaje", finally we construct the learning strategy considering a previous work which recommends a block of ACRA learning tactics according to the CHAEA learning styles.

1 Introduction

It is important first to understand each key concept according to this research. We agree with [1] who make it evident develop the concept of "learning to learn" and effective training in the use of learning strategies at university level, among our students.

1.1 Expert System (ES)

[2] defines it as a computer program that contains subject-specific-knowledge of human experts, [3] as a computer program that contains the erudition of a human specialist versed in a particular field of application, for [4] is a computer system that simulates human experts in a given area of expertise, [5] as a computer-based system that performs functions similar to those normally performed by a human expert, finally [6] says, ES simulates the judgment and behavior of a human that has expert knowledge and experience in a particular field. Based on the above definitions it is allowed to define an expert system as a computer program that simulates the expert judgment and the expert behavior of a human expert in an application particular field.

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_89

1.2 Learning object (LO)

Also known by different authors as digital learning object or virtual learning object.

For [7] is the minimum independent structure containing a target, a content, a learning activity, a metadata and an evaluation mechanism, which can be developed with info-communication technologies to enable reuse, interoperability, accessibility and durability over time, [8] defines it as small training components that allow great flexibility of access via web for lifelong learning, one of the most popular definitions has been given by the IEEE Standard [9] and says that is any entity-digital or non-digital- that may be used for learning, education or training, in the same manner [10] brings that this entity, can be used, re-used, or referenced during technology-supported learning, to end [11] shows a learning object as a small instructional components that can be reused in different learning contexts. Therefore and taking e-learning as a target of this investigation too, a learning object can be defined as any interoperable and independent digital entity that contains: an objective, a content, a learning activity, metadata and an evaluation mechanism, and can be used, re-used, accessed, or referenced during technology-supported learning.

1.3 Learning strategies (LS)

Even today there is no consensus with the concept of learning strategy, however there are many favorable coincidences to this work.

[12] says that is a process of decision making, this one is conscious and intentional, also in this process the student selects and retrieves a coordinated manner, the conceptual, procedural and attitudinal skills needed to complete a given objective, always depending on the conditions of the educational situation when the action occurs, whereas [13] see them as a purposive activities are reflected in the four major stages of information processing, on the other hand [14] defines it as operations performed by the thought when it has to face the task of learning, for [15] is a sequence of procedures for accomplishing learning, and the specific procedures within this sequence are called, learning tactics, [16] defines it as "an individual's approach to a task. It includes how a person thinks and acts when planning, executing, and evaluating performance on a task and its outcomes", finally [17] define them as the thoughts and actions we engage in, consciously or not, to learn new information.

As you can see, the above concepts are very varied and these ones vary depending of the researcher point of view. Hence here are defined as the tactics, operations, thoughts, actions, acts, approaches or techniques inside a sequence of procedures that the person uses, consciously or not to complete learning tasks.

1.4 E-Learning

Computer-aided instruction, just-in-time learning, and intelligent systems can be included in the term "e-learning" [18]. [19] considers that electronic learning (or eLearning) is a learning that occurs through the resources provided via the Internet and that includes learning solutions that go beyond the traditional paradigms of teaching, with a growing emphasis on learning type cooperative, collaborative, personalized and as an alternative to solve problems related to education demand, for [20] is the electronically supported learning; all kinds of learning with the use of digital media for the presentation or distribution of learning materials and/or the support of communication, [21] affirm that is "the appropriate application of the Internet to support the delivery of learning, skills, and knowledge in a holistic approach not limited to any particular course, technologies, or infrastructures", [22]

concludes that is "electronically facilitated, enhanced and managed learning", last is necessary to mention a new concept named E-Learning 2.0 which goes very relevant with the previous learning object concept and is describe for [23] as a platform for personal learning through interoperable tools that enable reusable content to be authored, repurposed, mixed and shared according to students' particular needs and interests. Taking into account the aforementioned e-learning can be defined as an electronically supported learning acquisition, distributed and facilitated primarily by electronic means usually online for its presentation and distribution that enable reusable content to be authored, repurposed, mixed and shared according to the students' particular needs and interests, in addition this system consisting of a variety of channels (e.g., wireless, satellite), and technologies (e.g., cellular phones, tablets) as they are developed and adopted to improve the quality of learning.

2 Background

2.1 CHAEA (Learning styles questionnaire from Honey-Alonso)

[24] define them as a questionnaire on learning styles that consists of eighty questions to there to answer expressing agreement or disagreement. Similarly they mention that the reliability and validity of CHAEA has been demonstrated on the evidence relevant statistics to analyze the learning styles of a sample of 1,371 students from 25 faculties of the Complutense and Politécnica of Madrid.

The characteristics of the styles are described as [25]: (1) Active.- people who are characterized by being cheerleaders, improvisers, discoverers, spontaneous and risky (2) Reflective.- people who are careful, receptive, analytical and exhaustive, (3)Theoretical.- people who are methodical, logical, objective, critical and structured, (4) Pragmatic.- Includes the experimenters people, practical, effective and realistic.

2.2 ACRA Learning Strategies Scales (Acquisition, Coding, Retrieval and Support)

[27] explain, 4 scales of ACRA. A general description of the scales is set out below:

Acquisition. Evaluates information procurement strategies [27]. **Coding.** This scale assesses the coding information strategies [27]. **Retrieval.** The cognitive system needs, have the resilience or memory of that knowledge stored in the LTM (Long Term Memory) [1]. **Support.** Helps the acquisition, coding and retrieval scales to increment motivation and attention.

2.3 CHAEA and ACRA together

[28] present a review of learning styles, learning strategies and academic performance during the 2000-2011 period. They conclude that the most used for measuring learning styles were the CHAEA and LSI (Learning Styles Inventory) instruments; meanwhile, for measuring learning strategies are the ACRA and Lassi.

On the other hand, the data presented by [29] shows how using innovative methodologies, focusing on learning, improvements occur in strategies for student learning. [30] present data that demonstrate the need for an explicit and systematic

teaching study skills to undergraduates, as seems to be the best route to certify that their learning is the most successful.

2.4 Mexican standard for interoperability between learning objects environments

We decided to build this project based on the [31] standard because it permits to link the learning objects and e-learning successfully to our work.

OA properties. OA must have a minimum set of properties that facilitate its location, use, storage and sharing.

Unreachable, Reusable, Portable, Durable, Serializable, Recoverable, Educational, Autocontenible.

Metadata. Metadata can describe the OA characteristics, such as aspects of thematic content, learning objectives, represent a key element for achieving interoperability between environments for learning objects, and hence to reuse of educational resources.

OA composition. The OA composition is given by the set of elements that are part of the resources that enable the use in diverse instructional contexts, and help achieve the goal of learning for the OA was designed

OA environment (OAE). They are information systems that are used for storaging and delivering an OA.

2.5 Expert system as a solution

As [32] says, expert systems represent the knowledge of experts and typically query and guide users during a decision making process. Also [33] say that the knowledge of an expert should be represented by using symbols, creating a knowledge base, then a program inference that manipulates symbolic information stored in the database.

Problem-Solving Methods. [34] specifies which actions inference must be carried out to solve a given task, determines the sequence in which these actions have to be activated and also called roles of knowledge determine which role plays the mastery of knowledge in every action of inference. [35] mentioned that different types of knowledge give rise to different types of knowledge-based systems, including the rule-based systems (RBS).

Knowledge representation system rule-based. It is one in which knowledge is represented in a particular domain, so that such representation is processable by a computer program. AI is divided into two areas that complement themeselves, symbolic and subsymbolic. The symbolic aspects, represents knowledge based on mathematical logic through symbols establishes a representation of objects or ideas from the real world [36]. Our work belongs to the symbolic based. [37] Mentions the following four models, (1) triplets Object-Attribute-Value, (2) semantic networks and (3) frames and predicate logic. In this paper we use semantic networks considering the work of [38] who defines conceptual graphs (CG) as a graph representation for logic based on the semantic networks of artificial intelligence and the existential graphs of Charles Sanders Peirce.

3 Related work

On their study [39] identify the learning style preferences of first year statistics undergraduates and their relation to gender and academic. They conclude that for effective instruction, instructors need to broaden their range of presentation styles to help effective learning environment for all students. [40] assert that as a teacher, it is important to understand the different learning styles of the students in acquiring the information, and hence one can make the necessary changes that best match the learning style of the students. Finally conclude that none of the students learn only by one method, all the multimodal learners will have a predominance of a particular learning style. Although [41]´s study was published on 1988 their discoveries are current, the next statements are some proof of this. (1) Learning styles of most engineering students and teaching styles of most engineering professors are incompatible in several dimensions. (2) These mismatches lead to poor student performance, professorial frustration, and a loss to society of many potentially excellent engineers. (3) In this way a teaching style that is both effective for students and comfortable for the professor will evolve naturally and relatively painlessly, with a potentially dramatic effect on the quality of learning that subsequently occurs.

Now we present the work of [42] where they choose to analyze a sample of 64 students enrolled in the first year of nutritional biology utilizing two questionnaires to find indicators relate to the teaching/learning activities and to define a plan of guidance and strategies for effective teaching and tutoring. The study of learning styles is useful as a starting point for improve strategies in the field of learning/teaching and to stimulate the students' self-evaluation.

In their paper [43] compare the learning styles and learning strategies of students in different university degree programs. They concluded that students specializing in the same field had a similar learning style, indicating the predominance of one style over another depending on the specialization type. [44] deducted educational implications regarding the type of learning and study techniques employed by students. The influence of university specialties in teaching learning strategies are demonstrated. [45] has adapted ILS with e-Learning method because e- Learning is an efficient technology that particularly focuses on learners who wish to study anywhere and anytime. Their framework evaluation results indicate that the framework can help learner more effectively to learn the lesson by their learning styles. [46] proposed a system that takes under consideration learners' answers to a certain questionnaire, as well as classification of learners who have been examined before. As a result, factors such as cultural environment will add value to the learning style estimation. Finally [47] established an online diagnostic system for Learning Style Estimation that contributes to the adaptation of learning objects, they propose an easily applicable expert system founded on Bayesian Networks. Their research contributes to the economy of AEHS functionality, by reducing the range of LOs metadata characterizations concerning LSs.

4 Methodology

We coincide with [1] and [48] and consider the importance of evaluating CHAEA questionnaire and ACRA Learning Strategies, as potential tutoring tools can be used to identify psycho educational characteristics of tutored students entering a public university. Then we propose to present the students learning objects, which are integrated on strategy blocks depending of his learning style. These recommendations are opposite to his learning style preference. E.g. if a student is very low pragmatic

then we recommend high pragmatic learning strategies, for incrementing his pragmatic style.

Following is the process for the allocation of learning objects a student depending on their learning style.

We create a function that depends on Active, Reflective, Theoretical and Pragmatic learning styles, it is identified as function (1).

$$E(x)=x \tag{1}$$

Where x=1,2,3,4 and 1 represents Active, 2 Reflective, 3 Theoretical y 4 Pragmatic.

We generate a new function that depends of the learning style preference identified.

$$P(y)=y \tag{2}$$

Where y=1,2,3,4,5 y 1 represents very low, 2 low, 3 moderate, 4 high and 5 very high, then we obtain a function in $[x,y]$ detonated with (3).

$$EP(x,y)=xy \tag{3}$$

The above function gives us a couple that contains the learning style and the learning style preference.

We generate a new function that involves the above couple and now we add the recommended learning strategy level identified as function (4).

$$Z(xy)=z \tag{4}$$

Where z=1,2,3, and 1 represents low, 2 moderate, 3 high. We receive xy and then according the learning style preference we assign the z value.

Below we show the conditions.

If last term of (3) is 1 (very low) or 2 (low), Then we assign 3 (high) to z,

If last term of (3) is 3 (moderate), Then we assign 2 (moderate) to z,

If last term of (3) is 4 (high) or 5 (very high) Then we assign 1 (low) to z

Therefore function (xz) (5) is given below.

Then we get a function (xz) denoted by (5) that gives us a couple containing learning style and the recommended learning strategy level.

$$ER(x,z)=xz \tag{5}$$

Now we obtain a function to denote (6) the learning objects block in its general form.

$$O(xz)=w \tag{6}$$

Where w is the recommended learning objects block. This set is assigned directly taking on account the Table 1.

Table 1. Example of learning objects classification according learning strategies.

Learning objects classification according to learning style and learning strategy.	
1 (Adquisición)	2 (Adquisición)
1. Exploración	14. Repaso mental
2. Subrayado idiosincrático	15. Repaso reiterado
3. Repaso en voz alta	16. Epigrafiado
1 (Codificación)	2 (Codificación)
4. Nemotécnicas	17. Relaciones Intracontenido
5. Relaciones Intracontenido	18. Relaciones compartidas
6. Relaciones no compartidas	19. metáforas
7. Metáforas	20. Autopreguntas
8. Aplicaciones	21. paráfrasis
1 (Recuperación)	2 (Recuperación)
9. Búsqueda de codificaciones	22. Libre asociación
10. Búsqueda de indicios	23. Claves
1 (Apoyos)	24. Conjuntos
11. Contradistractoras	2 (Apoyos)
12. Interacciones sociales	25. Autocontrol
13. Motivación intrínseca y extrínseca	26. Contradistractoras
	27. Motivación intrínseca
	28. planificación

Next we assign learning objects block to each selected couple xz (here there is an example):

If xz=(1,1) the recommended set will be, w=1,4,5,9,11
If xz=(1,2) the recommended set will be w=1,2,4,5,6,9,11,12
If xz=(1,3) the recommended set will be w =1,2,3,…,13
If xz=(2,1) the recommended set will be w =14,17,18,22,25,26
If xz=(2,2) the recommended set will be w =14,15,17,18,19,22,23,25,26
If xz=(2,3) the recommended set will be w =14,15,16,…,28
If xz=(3,1) the recommended set will be w = 29,32,35,38

4.1 Conceptual graph

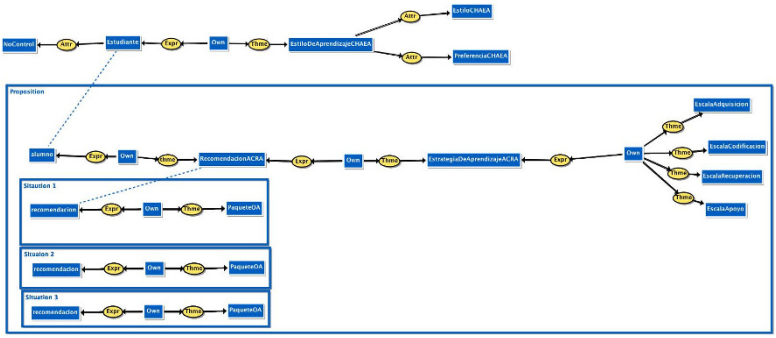


Fig. 1. First section represents the learning style identification and learning style preference for a given student. Second section represents the proposition which shows recommended strategy assignation under first section conditions. Third section shows three subsections or situations for assigning learning object sets under second section conditions.

5. Conclusions and discusion

According to the assumption considered, it is possible to develop an expert system that achieves engage learning objects, learning styles and learning strategies into an e-learning platform; it is concluded that it is possible however there are difficulties to achieve the objective because to relate learning styles with the strategies and learning objects is not possible to assign established mathematical operations as , i.e., it is not possible to use matrix, set theory and equations, since no operations can be performed between them.

We use conceptual graphs that allow accessible function symbology with manipulated attributes in different learning styles, learning strategies and learning objects for the relevant combinations and thus propose to the student a learning objects package according to his characteristics to acquire knowledge. Therefore a mathematical model with the desired characteristics was made to achieve the research objective. Analyzed expert systems only work with learning style, learning objects, e-learning or learning strategies, but they not link all of these.

There are intelligent platforms, which manage learning objects, however, do not consider the student's learning style, and they do not consider learning strategies. At universities no standardized platforms using learning objects and those that do not customize the objects according to the learning style of students. Universities do not standardize their learning objects and those that do, not personalize learning objects according to the student learning style. It is important not indiscriminately assign learning objects that provides the mathematical model, we recommend should do at random and not constant as now happens with our model. It is important to create a mathematical model that provides students all blocks of learning objects in one time and not in parts as until now occurs with our model. If the student has a “very high” preference learning strategy he will receive a great deal of learning objects, situation that will generate him confusion. Finally, it is not considered the teacher as a facilitator inside this context, and his role on the learning-teaching process.

Acknowledgments. The authors are grateful to Instituto Tecnológico Superior Zacatecas Norte authorities, for providing the means to carry out this research.

References

1 J.-C. C. Narvaez, «Estrategias de aprendizaje y rendimiento en inglés en alumnos universitarios,» Valadolid, 2015.

2 A. Benkő y C. Sik Lányi, «History of artificial intelligence,» *Encyclopedia of Information Science and Technology,* pp. 1759-1762, 2009.

3 R. P. Diez, A. Gómez Gómez y N. de Abajo Martínez, Introducción a la inteligencia: sistemas expertos, redes neronales artificiales y computación evolutiva, España: Universidad de Oviedo, 2001.

4 E. Castillo, J. M. Gutiérrez y A. S. Hadi, Sistemas expertos y modelos de redes probabilísticas, 1996.

5 Y. Duan y M. Xu, «Decision Support Systems in small businesses,» Encyclopedia od Information Science and Technology, pp. 974-977, 2009.

6 I. Atansova y J. Křupka, «Architecture and design of expert system for quality of life evaluation,» Informatica Economicaˇ, vol. 17, nº 3, pp. 28-35, 2013.

7 S. M. Massa, «Objetos de aprendizaje: Metología de desarrollo y evaluación de la calidad,» Buenos Aries, 2012.

8 M. F. Céspedes, R. T. Mora, L. Rodríguez, E. Toledo, A. Jiménez y M. R. Pacheco, «Investigación en la mejora docente mediante learning objects,» pp. 3096-3111, 2012.

9 Learning Technology Standards Committee of the IEEE, «Draft Standard for Learning Object Metadata,» New york, 2002.

10 P. Askar y U. Halici, «Diffusion of E-Learning as an Educational Innovation,» Encyclopedia of Information Science and Technology, pp. 1097-1100, 2009.

11 M. W. Dixon, J. M. Karlsson y T. J. McGill, «Data Communications and E-Learning,» Encyclopdia of Information Science and Tecnology, pp. 908-013, 2009.

12 C. S. J. Lugo, G. Rodríguez Hernández y M. Luna, «El cuestionario de estilos de apredizaje CHAE y la escala de estrategias de aprendizaje ACRA como herramienta potencial para la tutoría académica,» Revista Estilos de Aprendizaje, vol. 10, nº 10, pp. 1-28, 2012.

13 M. d. J. A. Vázquez y C. Doerfer, «Correspondencia de las estrategias de aprendizaje vs. desempeño académico,» vol. 1, nº 1, pp. 2205-2224, 2015.

14 J. A. B. llera, «Estrategias de aprendizaje,» Revista de educación, nº 332, pp. 55-73, 2003.

15 R. R. Schmeck, Learning Strategies and Learning Styles, Carbondalle, 1988.

16 N. Protheroe y S. Clarke, «Learning strategies as a Key to student success,» 2008.

17 A. Uhl Chamot, C. Foster Meloni, A. Bartoshesky y C. Keatley, Sailing the 5 Cs learning strategies a resourse guide for secondary foreifn language educators, National Capital Language Resoruce Center.

18 P. L. Rogers, «How teachers use instructional design in real classrooms,» Encyclopedia of Information Science and Technology, pp. 1777-1781, 2009.

19 S. Bermejo y M. E. Treviño, «Objetos de aprendizaje personalizados,» 2002.

20 A. Wiesner-Steiner, H. Wiesner y P. Luck, «Distance education teaching methods in childcare management,» Encyclopedia of Information Science and Technology, pp. 1168-1173, 2009.

21 K. E. Dooley, T. Pesl Murphrey, J. R. Lindner y T. H. Murphy, «Faculty competencies of information science and technology,» Encyclopedia of Information Science and Technology, pp. 1527-1531, 2009.

22 S. Fernando, «Issues of E-Learning in Third World Countries,» Encyclopedia of Information Science and Technology, pp. 2273-2277, 2009.

23 L. Walsh, «Cultural Issues in the globalisation of distance education,» Encyclopedia of Information Science and Technology, pp. 858-863, 2009.

24 C. M. A. García y D. J. Gallegos Gil, «Si yo enseño bien... ¿Por qué no aprenden los niños?,» Madrid, 2005.

25 L. M. E. Mayaute, «Análisis psicométrico del Cuestionario de Honey y Alonso de Estilos de Aprendizaje CHAEA con los modelos de la Teoría Clásica de los Tests y de Rasch,» Persona, nº 14, pp. 71-109, 2011.

26 C. M. A. García y D. J. Gallegos Gil, «Cuestionario Honey- Alonso de Estilos de Aprendizaje,» 10 Abril 2008. [En línea]. Available: http://www.estilosdeaprendizaje.es/menuprinc2 .htm.. [Último acceso: 10 Abril 2016].

27 J. M. a. R. Sánchez y S. Gallego Rico, «ACRA escalas de estrategias de aprendizaje,» Madrid, Tea, 2008.

28 M. J. Bahamón Muñetón, M. A. Viancha Pinzón, L. L. Alarcón Alarcón y C. I. Bohórquez Olaya, «Estilos y estrategias de aprendizaje: una revisión empírica y conceptual de los últimos diez años,» Pensamiento psicológico, vol. 10, nº 1, pp. 129-144, 2012.

29 B. G. López, I. Morera Bertuomeu, S. Iborra Chornet, M. J. Climent Olmedo, S. Navalón Oltra y E. García Felix, «Metodología centrada en el aprendizaje. Su impacto en las estrategias de aprendizaje y en el rendimiento académico de los estudiantes universitarios,» Revista española de pedagogía, nº 259, pp. 415-435, 2014.

30 M. d. J. A. Vázquez y C. Doerfer, «Correspondencia de las estrategias de aprendizaje vs. desempeño académico,» vol. 1, nº 1, pp. 2205-2224, 2015.

31 L. Castro García, G. López Morteo, R. Morales Gamboa, J. Muñoz Arteaga, J. J. Contreras

Castillo, C. M. Curlango Rosas y R. Villa Angulo, Norma Mexicana para la Interoperabilidad Volumen 1. Marco Teórico, 2014.

32 N. Wickramasinghe, «Creating Superior Knowledge Discovery Solutions,» Encyclopedia of Information Science and Technology, pp. 795-799, 2009.

33 C. S. Cinca y J. L. Gallizo Larraz, «Sección Inteligencia Artificial,» ciberconta.unizar.es, [En línea]. Available: http://ciberconta.unizar.es/Biblioteca/0004/SerGall96.html. [Último acceso: 21 Febrero 2016].

34 R. Studer y v. R. Benjamins, «Knowledge Engineering: Principles and methods,» pp. 1-38.

35 N. M. Sánchez, M. M. García Lorenzo y Z. Z. García Valdivia, «Modelo para diseñar sistemas de enseñanza-aprendizaje inteligentes utilizando el razonamiento basado en casos,» Revista Avances en Sistemas en Informática, pp. 67-78, 2009.

36 J. M. F. Fernández, «Generación de sistemas basados en reglas mediantes programación genética,» Madrid, 2008.

37 J. V. Román, R. M. Crespo García y J. J. García Rueda, «Open Course Ware,» 2011. [En línea]. Available: " http://ocw.uc3m.es/ingenieria-telematica/inteligencia-en-redes-de-comunicaciones/material-de-clase-1/03-sistemas-basados-en-conocimiento/view . [Último acceso: 01 Julio 2016].

38 J. F. Sowa, «Conceptual Graphs,» de Handbook of knoeledge representation, 1st Edition ed., 2008, pp. 213-237.

39 T. S. Talaga y P. D. Talaga, «Learning style preferences of first year statisctics Undergraduates,» 5th International Statistical Conference 2014, 2014.

40 C. Thomas, P. K. Kodumuri y S. P., «How so medical student learn? a study from two medical colleges in south India-across sectional study,» International journal of medical research & health science, vol. 4, nº 3, pp. 502-505, 28 april 2015.

41 R. M. Felder y L. K. Silverman, «Learning and teaching styles in engineering education,» Engr. education , vol. 78, nº 7, pp. 674-681, 1988.

42 M. Angeletti, V. Polzonetti y G. G. Pulcini, «Learning styles as a tool to improve both guidance and tutoring actions,» de New perspectives in science education, 2016.

43 V. Garzuzi y M. Mafauad, «Estilos y estrategias de aprendizaje en alumnos universitarios. Learning styles and strategies in university students,» Revista de orientación educacional, vol. 28, nº 54, pp. 71-95, 2014.

44 F. Camarero Suárez, F. Martín del Buey y J. Herrero Diez, «Estilos y estrategias de aprendizaje en estudiantes universitarios,» Psicothema, vol. 12, nº 4, pp. 615-622, 2000.

45 L. N. Nongkhai y T. Kaewkirya, «Framework for e-learning recomendation based on index of learning styles model,» de 2015 7th International conference on information technology and electrical engineering (ICITEE), Chiang Mai, 2015.

46 S. Botsios, D. A. Georgiou y N. F. Safouris, «Learning style estimation using bayesian networks,» de WEBIST 2007, Barcelona, 2007.

47 S. Botsios, D. Georgiou y N. Safouris, «Contributions to adaptative educational hipermedia systems via on-line learning style estimation,» de International forum of educational technology & society (IFETS), 2008.

48 C. S. J. Lugo, G. Rodríguez Hernández y E. Luna Montijo, «El cuestinario de estilos de aprendizaje CHAEA y la escala de estrategias de aprendizaje ACRA como herramienta potencial para la tutoría académica,» Revista Estilos de Aprendizaje, vol. 10, nº 10, pp. 1-28, 2012.

49 H. Delugach, «CharGer conceptual graph and knowledge modeling tool,» Harry Delugach, 12 septiembre 2014. [En línea]. Available: http://charger.sourceforge.net/. [Último acceso: 01 julio 2016].

50 G. López, A. Celina Justo y B. L. Flores, La educación a distancia en el contexto de Baja California: Fundamentos y capacidades para la creación de programas de estudios unveristarios, México, 2015.

Part IX
Workshop-A2LS-2016: International Workshop On Analytics & Awareness Learning Services

Security Analysis of OpenID Connect Protocol with Cryptoverif in the Computational Model

Jinli Zhang [1], Jintian Lu [2], Zhongyu Wan[3], Jing Li[4], Bo Meng[5]
Correspondent author: Bo Meng
[1,2,4,5] School of Computer, South-Central University for Nationalities,
MinYuan Road #708, HongShan Section,430074,
Wuhan, Hubei, China;
[3]School of Netcenter,Jianghan University,
SanJiaoHu Road #8,CaiDian Section,430056,
Wuhan,Hubei,China;
[1]jinlysa@163.com; [2]ljt45@hotmail.com; [3]wanzhongyu@jhun.edu.cn;
[4]stevenlee710@sina.com ; [5]mengscuec@gmail.com

Abstract. OpenID Connect protocol is widely used today, and it is one of the newest Single Sign-On protocols in authentication. At present, a lot of people are deeply focused on the researches of the security analysis of it. In this paper, we aimed at analyzing the authentication of OpenID Connect protocol by getting the message term of it through its authentication message flow, then formalizing it with Blanchet calculus in computational model, and finally transforming the model into the syntax of CryptoVerif, generate the CryptoVerif inputs in the form of channels Front-end, and import the syntax into the mechanized tool CryptoVerif to analyze its authentication. The result shows that OpenID Connect protocol has no authentication between the End-User and Authorization Server, Token Endpoint can't authenticate Client, while Client can authenticate Token Endpoint.

1 Introduction

OpenID Connect [1] was published in 2014 as the latest OpenID identity authentication and distributed identity system standard, is one of the significant single sign-on authentication protocol standard. It is built on top of OAuth2.0 [2], and simply introduces identity layer, which is abstracted as an ID Token to transport authenticated user information to the Client from authorization server. Based on OpenID Connect, Client can verify the identity of the End-User based on the authentication performed by an Authorization Server, and obtain basic profile information about the End-User in an interoperable manner. Once published, OpenID Connect gained the support of Yahoo, Google, Microsoft and many other companies [3, 4, 5].

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_90

Nowadays, people have done a large number of researches on security, protocol, authentication, authorization [6, 7, 8, 9, 10]. D Feet et al. [11] discovered that OpenID Connect was easily attacked by 307 Redirect attack and IDP mix-up attack in practice. Li W P et al. [12, 13] also revealed that the authentication standards like OpenID Connect were vulnerable to web-spoofing and phishing attacks, and attackers could successfully login Client as the status of protected user. Mainka C et al. [14] detected Second-Order Vulnerabilities in OpenID Connect, which may break user authentication, enable Dos, SSRF and Client-side code injection. Today, there is no automatic analysis of the security of OpenID Connect based on computational framework. So this paper analyzes the message term of OpenID Connect, models it with Blanchet calculus, and uses the mechanized tool CryptoVerif [15] to analyze its authentication according to the result.

2 Review of OpenID Connect Protocol

OpenID Connect performs authentication for the End-User or to determine whether the End-User is already logged in or not, and returns the result of the Authentication performed by the Server to the Client in a secure manner, the Authentication result is returned in an ID Token.

Authentication in OpenID Connect can follow one of three paths: the Authorization Code Flow, the Implicit Flow, or the Hybrid Flow, the flows determine how the ID Token and Access Token are returned to the Client. This paper picks the path using Authorization Code Flow to analyze the OpenID Connect, and Fig.1 describes the procedure of the abstracted flow in OpenID Connect protocol.

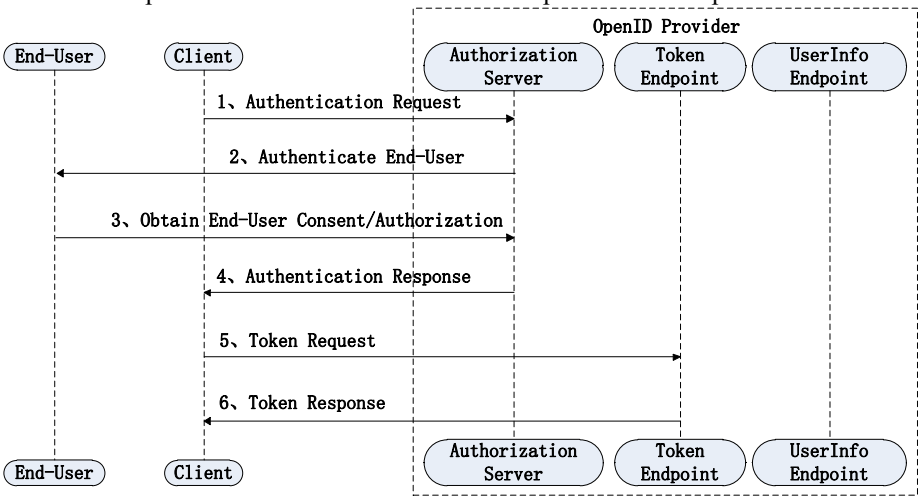


Fig. 1. Procedure of the abstracted flow in OpenID Connect protocol.

There are three parties in OpenID Connect protocol, Client, End-User, and OpenID Provider. Client is an application that attempts to request the access to resources

owned by protected End-User. End-User is the owner of resource, and he can permit the Client to have access to the resources protected by server through authorization. OpenID provider consists of three parts: Authorization Server, Token Endpoint, UserInfo Endpoint. Authorization Server can authenticate End-User, and send authorization code to Client when he successfully authenticates End-User and obtains End-User's consent. Token Endpoint receives the Token Request from Client, and sends ID Token with Access Token to the Client. UserInfo Endpoint is used to make response to the UserInfo Request from Client, and provide the Client with End-User's authorized information. When using the Authorization Code Flow, all tokens are returned from the Token Endpoint, authorization server sends an Authorization Code to the Client, which can then exchange it for an ID Token and an Access Token directly. Thus no tokens may be exposed to the User Agent and possibly other malicious applications with access to the User Agent.

3 Formalizing OpenID Connect Protocol with Blanchet Calculus

According to the procedure of the abstracted flow in OpenID Connect protocol, we formalize this protocol with Blanchet Calculus based on computational model. When modeling the protocol, the client and the End-User must finish registrations at the OpenID Provider, which is not the main content of OpenID Connect protocol itself, so we wouldn't show this process in detail in this paper.

3.1 The Declaration of Cryptographic Primitives and Channels Assertions

In Blanchet Calculus, channel is an important medium, all the communications between different parties are based on channels, so the definition of channels is needed first. In OpenID Connect protocol, ID Token must be signed with JWS, so when modeling, signature must be used. Here we choose a predefined SUF-CMA probabilistic signature scheme which is strongly unforgeable under chosen message attacks. Fig.2 describes the cryptographic primitives and channels assertions in formalizing OpenID Connect.

```
proba Psign.
proba Psigncoll.
expand SUF_CMA_signature(keyseed, pkey, skey, signinput, signature, seed,
                         skgen, pkgen, sign, check, Psign, Psigncoll).
channel start,c,c1,c2,c3,c4,c5,c6,c7,c8,c9,c10,c11,c12,c13,c14,c15,c16,c17,c18.
```

Fig. 2. Cryptographic primitives and channels assertions

3.2 Main Process

The main process receives a null message sent by the adversary on channel start, generates a randomly bitstring signseed in the type keyseed, generates OpenID provider's public key signpkey and private key signskey used for signature of ID token, sends the signpkey by the public channel c, models the execution of OpenID Connect with a parallel implementation of OpenID provider process OP_Process, Client process Client_Process and End-User process End_User_Process. Fig.3 describe the main process.

```
process
  start();
  new signseed:keyseed;
  let signpkey:pkey=pkgen(signseed) in
  let signskey:skey=skgen(signseed) in
  c<signpkey>;
  ((! iop n OP_Process)|(! iclient n Client_Process)|(! iend_user n End_User_Process))
```

Fig. 3. Main process

3.3 Client Process

When Client attempts to request the resources owned by protected End-User, he needs to send Authentication Request to the Authorization Server, just generates these must parameters responsetype, state_one, scope_one through new statement, then sends them with id_client and uri_client to the Authorization Server by the channel c7. To receive the response from the Authorization Server, Client gains code_from_OP, state_three by the channel c14. To have the access to the protected resource, the Client needs to send Token Request to the OpenID provider, namely generates granttype_one through new statement, and then sends granttype_one, code_from_OP, id_client, secret_client, and uri_client to the OpenID provider by the channel c15. If the OpenID Provider verified the information successfully, the Client can obtain message, sign_one_s, accesstoken, tokentype_one, expiresin_one, scope_three by channel c18. By checking the signature sign_one_s and message, the client can gain the ID Token, and the accesstoken is the Access Token, the Client gains the access to the resources of protected End-User only if he gains the Access Token. Fig.4 describe the critical part of Client process.

```
let Client_Process=
  c ();
  new responsetype:response_type;
  new state_one:state;
  new scope_one:scope;
  c7< scope_one,responsetype,id_client,uri_client,state_one >;
  c14(=OpenIDP_two,code_from_OP:code_authorization,state_three:state);
  new granttype_one:grant_type;
  c15< granttype_one,code_from_OP,uri_client,id_client,secret_client >;
  c18(message:signinput,sign_one_s:signature,accesstoken_one:access_token,
      tokentype_one:token_type,expiresin_one:expires_in,scope_three:scope);
  if check(message,signpkey,sign_one_s) then
    let combine5(idtoken_one:id_token)=message.
```

Fig. 4. Client process Client_Process

3.4 End_User Process

By channel c10, the End-User receives the string_from_OP . If the string_from_OP is ask_for_authentication, it means that the Authentication Server request for authentication. The End-User agrees to authorize and sends its user_name and user_password by channel c11. Fig.5 describes the critical part of End-User process.

```
let End_User_Process=
   c();
   c10(=OpenIDP_one,string_from_OP:bitstring);
   if string_from_OP=ask_for_authentication then
      c11< User_one,user_name,user_password >.
```

Fig. 5. End-User process End_User_Process

3.5 OpenID Provider Process

When OpenID Provider receives scope_two, responsetype_two,clientid_two, uri_two, state_two by the channel c8, Authorization Server must validate the request by verifying whether the clientid_two and uri_two are bound together through find statement. If it is successfully validated, and the value of responsetype_two is “code”, Authorization Server requests to authenticate the End-User by sending a message ask_for_authentication by channel c9.

When Authentication Server receives user, pass by the channel c12, he validates whether user has registered and whether user is bound to pass or not through two find statements. If the validation is successful, that means the End-User is successfully authenticated, and he has to give the Client a successful response, generates auth_code and binds the auth_code with uri_two, clientid_two, sends auth_code and state_two to the Client by the channel c13.

After the client received grant_type_two, code_from_client, client_uri_three, clientid_three, client_secret_three by channel c16, if the client_uri_three is identical to uri_two and the value of client_uri_three is “authorization_code”, OpenID provider must finish the following verifications through three find statements. Firstly, validate the credential of the Client, ensure that the clientid_three and client_secret_three are combined together. Secondly, verify whether the code_from_client exists, finally, verify that the code_from_client is combined with clientid_three, client_secret_three and client_uri_three to prove that the authorization code was generated for this Client. Only if all the validations are successful, the OpenID Provider generates idtoken, accesstoken, tokentype, expiresin, and seed_one through new statements, transforms the idtoken to signmessage and signs the signmessage with public key signskey as sign_one, and finally sends signmessage, sign_one, accesstoken, tokentype, expiresin, scope_two to the Client by the channel c17. The idtoken is used to send the authentication information from Authorization Server to Client, and the accesstoken is used for the Client to get the access to the protected resources of the End-User.

Fig.6 describes the critical part of OpenID Provider process.

```
let OP_Process=
    c8(scope_two:scope,responsetype_two:response_type,clientid_two:client_id,
    uri_two:redirect_uri,state_two:state);
    find k<=N suchthat defined (clientid_client_uri_succ[k])&&
    (clientid_client_uri_succ[k]=combine2(clientid_two,uri_two)) then
    if responsetype_two=code then
       c9<OpenIDP_one,ask_for_authentication>;
    c12 (=User_one,user:username,pass:password);
    find l<=N suchthat defined (login_username_succ[l])&&(login_username_succ[l]=user) then
    find m<=N suchthat defined (username_password_succ[m])&&
    (username_password_succ[m]=combine3(user,pass)) then
    new auth_code:code_authorization;
    let clientid_code:bitstring=combine4(clientid_two,uri_two,auth_code) in
    c13<OpenIDP_two,auth_code,state_two>;
    c16 (grant_type_two:grant_type,code_from_client:code_authorization,
         client_ uri_three:redirect_uri,clientid_three:client_id,client_secret_three:client_secret);
    if grant_type_two=authorization_code then
    if client_uri_three=uri_two then
    find p<=N suchthat defined (clientid_clientsecret_succ[p])&&
    (clientid_clientsecret_succ[p]=combine1(clientid_three,client_secret_three)) then
    find q<=N suchthat defined (auth_code[q])&&(auth_code[q]=code_from_client) then
    find r<=N suchthat defined (clientid_code[r])&&
    (clientid_code[r]=combine4(clientid_three,client_uri_three,code_from_client)) then
    new idtoken:id_token;new accesstoken:access_token;
    new tokentype:token_type;new expiresin:expires_in;new seed_one:seed;
    let signmessage:signinput=combine5(idtoken) in
    let sign_one:signature=sign(signmessage,signskey,seed_one) in
    c17<signmessage,sign_one,accesstoken,tokentype,expiresin,scope_two>.
```

Fig. 6. OpenID provider process OP_Process

4 Automatic Verification of Authentication in OpenID Connect with CryptoVerif

In order to verify the authentication in OpenID Connect protocol, the Blanchet calculus model needs to be transformed into the syntax of CryptoVerif and to generate the CryptoVerif inputs in the form of channels Front-end.

At first, we use non-injective correspondences to model authentication between different parties, and table 1 shows the goals of authentication.

Table 1. The goals of authentication.

Correspondences	Authentication
event UserAuth(x)==>Authentication(x)	Verify the authentication from End-User to Authorization server
event Authorization(x)==>User(x)	Verify the authentication from authorization server to End-User
event TokenEnd(x)==>Client(x)	Verify the authentication from Token Endpoint to Client
event IDTokenClient(y)==>IDTokenEnd(x)	Verify the authentication from Client to Token Endpoint

Then, we transmit the OpenID Connect protocol's Blanchet Calculus model into the syntax of CryptoVerif and add all events in appropriate place, then importing the syntax of CryptoVerif into the mechanized tool CryptoVerif, and the Fig.7, Fig.8, Fig.9, Fig.10 show the result of analysis of authentication in OpenID Connect.

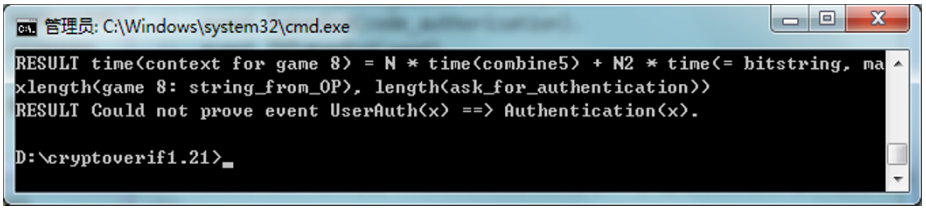


Fig. 7. Result of event UserAuth(x) ==>Authentication(x).

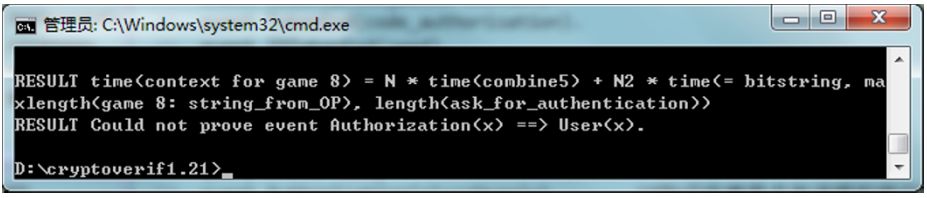


Fig. 8. Result of event Authorization(x) ==>User(x).

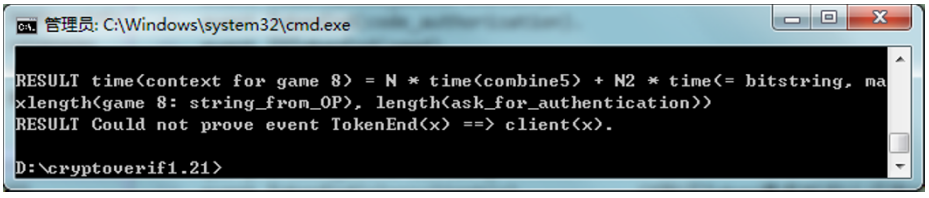


Fig. 9. Result of event TokenEnd(x) ==>Client(x).

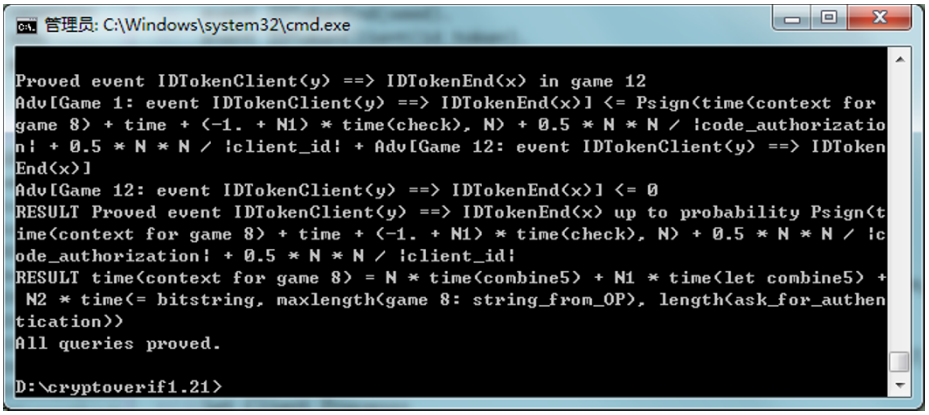


Fig. 10. Result of event IDTokenClient(y) ==>IDTokenEnd(x).

According to the result, it's easily to know whether there exists authentication between different parties. More specifically, "RESULT Could not prove event UserAuth(x)==>Authentication(x)" and "RESULT Could not prove event Authorization(x)==>User(x)" show that there is no authentication between the authorization server and the End-User, "RESULT Could not prove event TokenEnd(x)==>Client(x)" shows that Token Endpoint can't authenticate Client, while the result "All queries proved" of enent IDTokenClient(y)==>IDTokenEnd(x) shows that Client can authenticate Token Endpoint.

After analyzing the result, we know that the reason why Client can authenticate Token Endpoint is that the signature is used in the Token Response from Token Endpoint to Client, while no security mechanisms are used in message transmission between other parties. So, we can add some security mechanisms like signature in the message transmission to ensure the authentication between different parties.

Conclusion

In order to verify whether there exists authentication between different parties in OpenID Connect, this paper gets the message term of OpenID Connect by analyzing its authentication message flow, models it with Blanchet calculus in computational model, and finally transforms the Blanchet calculus model into the syntax of CryptoVerif, generates the CryptoVerif inputs in the form of channels Front-end, and imports the syntax into the mechanized tool CryptoVerif to analyze its authentication. The result shows that, without the support of TLS, OpenID Connect protocol has no authentication between the End-User and Authorization Server, Token Endpoint can't authenticate Client, while Client can authenticate Token Endpoint, because the signature is used in the Token Response from Token Endpoint to Client, while no security mechanisms are used in message transmission between other parties. To ensure the authentication, some security mechanisms like signature should be added to the message transmission.

References

1. Sakimura N, Bradley J, Jones M, et al. OpenID Connect Core 1.0. (2014-11-08). http://openid.net/specs/openid-connect-core-1_0.html
2. Hardt D. The OAuth 2.0 Authorization Framework. IETF. RFC 6749. http://tools.ietf.org/html/rfc6749
3. Siriwardena P. Advanced API Security: Securing APIs with OAuth 2.0, OpenID Connect, JWS, and JWE. Apress Media, 2014:181-200
4. Mladenov V, Mainka C, Krautwald J, et al. On the security of modern Single Sign-On Protocols – OpenID Connect 1.0. eprint arXiv,2015
5. Xue R, Lei X F. Security Protocol: The soul of information security – The latest research and development of the analysis of security protocol. Proceedings of the Chinese academy of sciences, 2011, 26(3):287-296
6. Alireza P S, Joao P S. Authentication, authorization and auditing for ubiquitous computing: a survey and vision. DOI: 10.1504/IJSSC.2011.039107:59-67
7. Xiaoping C, Stephane M. Formally testing the protocol performances. DOI: 10.1504/IJSSC.2015.069197:76-88
8. Xi S, ZhengTao J, MeiRong Z, Yumin W. Versatile identity-based signatures for authentication in multi-user settings. http://dx.doi.org/10.1504/IJGUC.2014.062905
9. Yan R, Hongbin W, Jiali D, Liqiong M. Code-based authentication with designated verifier. http://dx.doi.org/10.1504/IJGUC.2016.073779
10. Danai C, Christopher M. Security challenges in the internet of things. DOI: 10.1504/IJSSC.2015.070945:141-149
11. Fett D, Küsters R, Schmitz G. A Comprehensive Formal Security Analysis of OAuth 2.0, eprint arXiv, 2016
12. Li W P, Mitchell C J. Addressing threats to real-world identity management systems, ISSE 2015:251-259
13. Li W P, Mitchell C J. Analysing the Security of Google's implementation of OpenID Connect. eprint arXiv, 2015
14. Mainka C, Mladenov V, Schwenk J. On the security of modern Single Sign-On Protocols – Second-Order Vulnerabilities in OpenID Connect, eprint arXiv, 2016
15. Blanchet B. A Computationally Sound Mechanized Prover for Security Protocols. IEEE Symposium on Security&Privacy, Oakland, California, May 2006. 140-154

Jinli Zhang is a postgraduate at school of computer, South-Center University for Nationalities, China. Her current research interests include security protocols and formal methods.

Jintian Lu is now a postgraduate at school of computer, South-Center University for Nationalities, China. His current research interests include security protocols and formal methods.

Zhongyu Wan is a postgraduate at school of computer, South-Center University for Nationalities, China. Her current research interests include security protocols and formal methods.

Jing Li is now a postgraduate at school of computer, South-Center University for Nationalities, China. His current research interests include security protocols and formal methods.

Bo Meng received his M.S. degree in computer science and technology in 2000 and his Ph.D. degree in traffic information engineering and control from Wuhan University of Technology at Wuhan, China in 2003. From 2004 to 2006, he worked at Wuhan University as a postdoctoral researcher in information security. From 2014 to 2015, he worked at University of South Carolina as a Visiting Scholar. Currently, he is a full Professor at the school of computer, South-Center University for Nationalities, China. He has authored/coauthored over 50 papers in International/National journals and conferences. In addition, he has also published a book “secure remote voting protocol” in the science press in China. His current research interests include security protocols and formal methods.

Towards a Particular Prediction System to Evaluate Student's Success

David Baneres
IT, Multimedia and Telecommunications Department
Open University of Catalonia
Barcelona, Spain
dbaneres@uoc.edu

Abstract. In the last decades, different models have been developed in order to predict student's success. These models aimed to predict based on previous performance data whether new students will be able to pass the present course. These models have a large potential in adaptive systems where different learning paths can be offered to the learner based on her current progression through the course and comparing with the standard performance of students who have passed the course. In this paper, we describe the design of a prediction system and different models that can be used to help to predict student's success better. The presented models have particularly designed for an introductory course to learn digital systems design.

1 Introduction

Predicting students' performance has been extensively explored in the past decades. Predicting the success ratio of a course even prior to starting the instructional process has been attempted based on collected information of enrolled students. Prior-knowledge, student's profile or previously acquired skills are some examples of information that has been used as predicting factors.

Prediction on students' success (or in some cases classification of students) has a large potential. In addition to give additional information related to the enrolled students, this information can be used either by instructors and learners. Instructors may have an additional tool to know which students are susceptible to have problems to pass the course and personalized support can be additional given to help them. Learners may also have this information. Knowing their "chances" to pass the course is a double-edged sword. The positive effect is that the learner is aware in real time about its progression compared to other students and she can adapt her effort based on the prediction. The negative effect is that, in the case of predicting that the course will not be passed, the learner may feel depressed and she may decide to early dropout the

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_91

course. The learner should be aware all time that this is a predictive system and, therefore, there is some possible failure on the prediction.

However, predictive models may have other uses in learning. Predictive models can be used to provide additional services to learners. Adaptive systems based on the student's prior learning or current performance, recommender systems based on resource materials or activities or even course enrollment recommenders are some particular services that can be provided to students to personalize her learning experience.

In this paper, we describe a case study of a prediction system in the context of a first-year course in the Degree of Computer Science related to concepts of digital systems fundamentals. The aim of the approach is to help students and teachers during the instructional process with the final objective to successfully pass the course. Additionally, different models are presented to show the potential accuracy of the system based on the students' performance information during one semester. Good results are achieved even with a small sample. We expect that a larger sample on the future will return more accurate results.

The paper is organized as follows. Section 2 describes previous work related to models to predict student's success. Section 3 defines the characteristics of the course used for the case study. Section 4 specifies the design of the predictive model and the defined models are presented in Section 5. Finally; Section 6 summarizes accuracy results on presented models.

2 Related work

In this section, we describe some previous work related to predictive models in Computer Science. Initial prediction models were proposed in the seventies and eighties. The most challenging areas were initial programming courses and initial course fundamentals in Computer Science.

In Programming courses, models based on current academic success and the level of proficiency in some science related skills such as mathematics were developed [1][2]. These models tried to classify enrolled students based on prior knowledge. Other models incorporated student's demographic information [3][4][5] to enrich the models giving in some cases more refined models. Similar models were designed for Computer Fundamentals courses based on prior learning and demographics [6].

In the last decade, new opportunities arose. The evolution of the Virtual Learning Environments (VLE) and the possibility to (semi-)automatically collect data from students (i.e. performance and actions) enabled the possibility to create new prediction systems on dynamic data. Authors in [7] gathered information from the utilization of a programming IDE in a programming course. Data such as source code, compilation errors and correct executions of the code combined with the student's performance were merged to create the model. Similarly, quick questions within a face-to-face class were introduced in [8] to collect correct responses and time spent to answer the questions correctly. This information was used to analyze student's success and the most predictive questions. Collaboration among students was also

analyzed in [9]. Results showed that there is a correlation between final success and social connections among students.

Aforementioned, the potential of these prediction models is to provide addition services to learners and instructors. Grade prediction [10], automatic detection of students that may need assistance [11] or a recommender in an adaptive system [12][13] are some examples of the type of services that can be offered.

3 Case Study: Computer Fundamentals

In the course of Computer Fundamentals, a student has to acquire the skills of analysis and synthesis of small digital circuits and to understand the basic computer architecture in 150 hours. The synthesis process is presented at the logical level without describing all the inherent electronic problems.

The contents of the course are organized into 5 units: 1. Introduction to computer fundamentals, 2. Numeric representation and binary codification, 3. Combinational logic circuits and Boole algebra, 4. Sequential logic circuits and Moore's model, and 5. State machines and basic computer structure.

The assessment is divided into 3 continuous assessment activities (CAA) and one final project. The activities evaluate the contents of the second, third and fourth unit and the final project is related to the design and synthesis of a state machine explained in the last unit. There is also a final exam with problems related to the second, third, and fourth units (30% each one) and a question of the fifth unit (10%). The final mark (FM) of the course as we can observe on Equation 1 is obtained by combining the results of the continuous assessment activities (CAAs), the final project (FP) and a final exam (FE). The project and the exam are mandatory while the CAAs are optional, but they can improve the final mark.

$$FM = MAX(35\%FP + 35\%FE + 30\%CAAs,\ 50\%FP + 50\%FE)\ . \quad (1)$$

An intelligent tutoring system (ITS), called VerilUOC [14] is used to practice the design of digital circuits. The objective of the system is two-fold: In one hand, the student has an edition and simulation tool to design and test the digital circuit. On the other hand, there is a complementary module where exercises from an initial tutorial, the textbook, the CAAs and the exams of previous semesters can be solved. The module allows to validate automatically the correctness of the design performed by the student compared to the official solution provided by the instructors. In the case of mismatch, the tutoring system gives a counterexample as a personalized feedback to guide the student to the correct solution. The system currently has 152 exercises divided by the corresponding unit of the material (combinational circuits, sequential circuits, and state machines). The set of exercises proposed in the system should be enough to acquire the sufficient level of knowledge in order to be able to design simple circuits and to pass the course.

Computer Fundamentals is a first-year course on the Bachelors of Computer Science and Telecommunication Technologies at the Open University of Catalunya (UOC) and, consequently, there are a large number of students each semester (larger

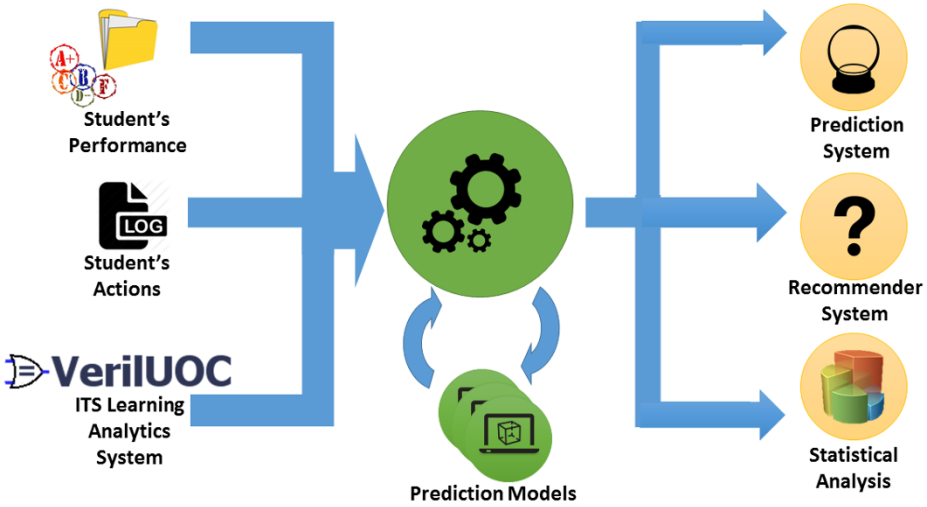


Fig. 1. Design of the prediction system

than 300 students). A recurrent problem in this course is the high dropout rate that reaches values nearly to 50% of enrolled students in some semesters.

After analyzing the satisfaction surveys, many students claim that there is too much material to learn and practice and they would be interested in a personalized guidance in order to pass the course.

This paper proposes the first phase to fulfill this request. A personalized guidance (or personalized path) can be automatically proposed based on the current performance on the course and the average performance and actions that previous students who have successfully passed the course have been performed. This evaluation of the average performance and actions has been done by applying machine learning techniques to data collected from previous semesters of the course. With this data, a classification model can be trained to predict the successful ending of the course.

4 Design of the prediction system

This section describes the design of the prediction system. The system architecture is very simple as it is illustrated in Fig. 1. The core system is composed of a machine learning engine. Some features are accepted as incoming data in order to train the models. Additionally, the generated models can be used to provide multiple services based on the knowledge acquired from the classification of the learners.

Currently, we propose several models as it is explained in next section. Each model has been particular designed depending on the number of available features. The type of the machine learning algorithms selected is a supervised learning method since we have data from to past students where we can learn. Specifically, the problem is a binomial classification problem (student passes the course or student fails the course).

We can increase the complexity of the problem by trying to predict the score of the student. However, this problem will be different and the algorithms used as well.

Many well-known algorithms can be used on binomial classification as we will observe in Section 5. These algorithms accept various types of features that describe several characteristics of the learners. Note that, the created models will not be static models since they can be dynamically improved by data obtained on each edition of the course. More data will generate in few editions fine-tuned models particularly applicable on this course.

The main challenge together with the selection of the best algorithm will be the selection of features. The model is characterized by the set of features that help to build and specify the model. Few features will generate a poor model with a high failure rate. A large number of features may generate an overfitted model or some features may be complementary or redundant. In this first design, we have identified three potential sources for our case study:

- Student's performance: This source of data provides the performance of students in terms of grades. A grade for each continuous assessment activity (CAA1, CAA2 and CAA3), the final continuous assessment (CA) and the final project (FP) can be retrieved. Note that, the grade of the final exam is excluded since there is no point to predict if a student will pass when all grades are available. UOC grading system provides a qualitative score (letter grade ranging from A to D) for all activities and an additionally quantitative score (numerical grade) for CA and FP.
- Student's actions: During the instructional process, the students perform some actions that can be recorded. For instance, in on face-to-face learning, a learner is highly participative during the class, she submits some non-assessment activities or she attend all laboratory/practice sessions. On an online learning, additional actions can be recorded, for instance, accessing to learning materials or interaction in the discussion forum. Online learning with a VLE is preferable since actions can be registered automatically.
- Learning analytics system: Aforementioned, Computer Fundamentals has the ITS VerilUOC to support the learning process. In this case, this ITS stores all interactions of the learners with the system. Currently, it stores the number of attempts and the correctness of any submission for each exercise. Note that, additional data could be useful such as the total time needed to complete each activity. However, this is not contemplated in this version of the ITS.

Different potential outputs have been identified focused either for learners and teachers:

- Prediction system: This service can be for learners and teachers. The objective is to predict whether the learner will pass the course based on her current performance. Note that, several predictions can also be enabled within the course, i.e. to pass a specific assessment activity. However, these intermediate models may suffer high failure rates since few features will be available.
- Recommender system: This service can be also used for learners and teachers. For the point of view of the learner, a recommender system should be able to specify which activity (or set of activities) should be done to "probably" pass the course. Also, this recommender system can be integrated into intelligent tutoring systems to create adaptive learning paths. More common exercises solved by previous

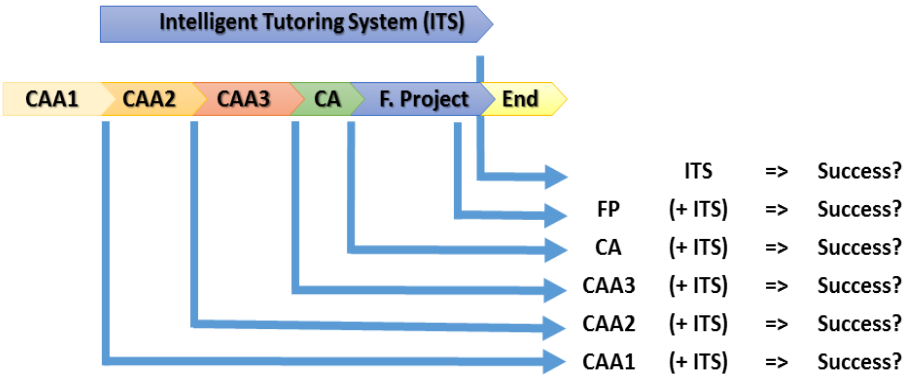


Fig. 2. Models and their correspondence to the activities of the course

learners that have passed the course can be suggested. For the point of view of the instructor, the system could be used to advise the learner during the course and know in real time the progression of the student. A personalized support is highly appreciated by learners. Note that, the recommendation information can be easily extracted from rule-based or tree-based machine learning algorithms where the list of rules or the binary decision tree to reach a prediction can be extracted from the model.

- Statistical analysis: This service will be only used by instructors and academic coordinators. Statistical analysis on predictions and correlation among similar courses can be very useful to improve the course design and sharing good practices.

In the next section, we will explore different models than can be applicable based on the available features.

5 Prediction Models

The models may be different depending on the course since they will be mostly influenced by the number of activities (assessment and non-assessment) and the grading system (the percentage of influence of each activity on the final mark). Therefore, the models should be particularly selected for each course.

Here, we have defined 6 basic models with different variations based on the distribution of assessment activities through the course. Each activity has been set at the end of each assessment activity since we have a grade at this point to be added as a feature. Fig. 2 illustrates the different models and their correspondence to the activities of the course.

First, three models (CAA1, CAA2, CAA3) have been created based on the assessment mark obtained at the end of each continuous assessment activity. These models are interesting for evaluating the progression of the student during the continuous assessment. Intuitively, the accuracy of the models will increase from CAA1 to CAA3. Moreover, some particular well-known cases within the course such

as "if the learner does not submit CAA1 will be difficult to pass the course" will be early identified.

Next, a model with the mark of the continuous assessment (CA) is defined. The score of the CA is obtained from a grading formula based on the grades of each continuous assessment. Some proactive actions of the learner such as participating in the discussion forum or helping other students are also taken into account to round the mark.

Finally, the score of the final project (FP) defines the last model. This model should be the most accurate since all the marks are available and the exam is the only assessment activity that remains.

As we can observe in Fig. 2, all models have an ITS label. This label specifies a variation of the model by taking into account the information of the ITS. This information should enrich the model by increasing the accuracy of the models. Our hypothesis is that if the learner uses the ITS, she would have more chances to pass the course. Additionally, we are interested to see the accuracy of a model where only the information of the ITS without any grade of assessment activities of the ITS is taken into account.

6 Experimental Results

In this section, we show the accuracy of each proposed model by using standard supervised machine learning algorithms. The objective is to show that even using a small sample of data as training set, good results can be obtained even in the initial stages of the course.

6.1 Experiment Setup

The experiment has been performed using WEKA [15]. Weka is a collection of machine learning algorithms for data mining tasks and there is also an API to use the algorithms from JAVA code. Here, the Experimenter module has been used to test a set of algorithms.

The features for the experiment were extracted from the 2015 Fall semester of the Computer Fundamentals course described in Section 3. The students' quantitative and qualitative scores for each assessment activity have been selected, except the final exam since the final mark can be automatically obtained when this score is available using Equation 1. Additionally, data from VerilUOC has been used. The number of tries and whether the submitted solution is correct or not have been extracted for each exercise. At this stage, information from the UOC VLE based on learning resources consumption or discussion forum participation is not available. However, we will study in the future how to add this information.

Two datasets have been created: the training (310 students) and the test (156 students) sets. These sets correspond to two independent classrooms (no interaction among students) in the same course where the same activities and resources with the only difference on the instructor and the language (Catalan for the training set and Spanish for the test set).

Table 1. Accuracy on Continuous Assessment Activities Models

CROSS VALIDATION	OneR		Decision Table		J48		Atribute Selection	
	Ac. (%)	RSME	Ac. (%)	RSME	Ac. (%)	RSME	Ac. (%)	RSME
CAA1	**70,32**	0,54	**70,32**	0,43	**70,32**	0,43	**70,32**	0,43
CAA1_ITS	**76,06**	0,48	73,87	0,43	71,65	0,48	73,55	0,44
CAA2	79,68	0,44	80,87	0,39	**80,13**	0,37	79,94	0,37
CAA2_ITS	**79,68**	0,44	78,65	0,38	74,90	0,44	77,52	0,40
CAA3	83,55	0,40	83,23	0,34	**83,58**	0,35	83,26	0,34
CAA3_ITS	**83,55**	0,40	82,97	0,34	80,03	0,41	82,29	0,36
TESTSET	**Ac. (%)**	**RSME**	**Ac. (%)**	**RSME**	**Ac. (%)**	**RSME**	**Ac. (%)**	**RSME**
CAA1	75,32	0,50	**75,32**	0,41	**75,32**	0,41	**75,32**	0,41
CAA1_ITS	74,05	0,51	72,15	0,44	75,95	0,45	**76,58**	0,40
CAA2	76,58	0,48	**80,38**	0,38	79,11	0,37	79,11	0,37
CAA2_ITS	74,05	0,51	72,15	0,44	**84,18**	0,39	81,65	0,38
CAA3	**86,08**	0,37	**86,08**	0,31	82,28	0,34	82,28	0,34
CAA3_ITS	86,08	0,37	85,44	0,32	**86,71**	0,36	86,08	0,32

Table 2. Accuracy on Continuous Assessment Activities Models

CROSS VALIDATION	OneR		Decision Table		J48		Atribute Selection	
	Ac. (%)	RSME	Ac. (%)	RSME	Ac. (%)	RSME	Ac. (%)	RSME
CA_QL	**85,94**	**0,37**	85,10	0,33	83,16	0,37	85,10	0,34
CA_QL_ITS	**85,94**	0,37	83,97	0,33	79,55	0,42	83,32	0,36
CA_QT	**85,94**	0,37	85,06	0,34	83,61	0,37	83,29	0,34
CA_QT_ITS	**85,94**	0,37	84,74	0,34	81,29	0,40	83,61	0,36
CA_QT_QL	**85,94**	0,37	85,03	0,33	82,58	0,37	83,29	0,34
CA_QT_QL_ITS	**85,94**	0,37	83,97	0,33	81,10	0,40	83,94	0,36
TESTSET	**Ac. (%)**	**RSME**	**Ac. (%)**	**RSME**	**Ac. (%)**	**RSME**	**Ac. (%)**	**RSME**
CA_QL	86,08	0,37	82,91	0,34	87,34	0,31	**91,77**	0,27
CA_QL_ITS	86,08	0,37	85,44	0,32	86,71	0,35	**89,24**	0,32
CA_QT	86,08	0,37	**86,08**	0,31	82,28	0,36	82,28	0,34
CA_QT_ITS	**86,08**	0,37	85,44	0,32	83,54	0,41	82,91	0,40
CA_QT_QL	89,24	0,33	**89,87**	0,31	82,28	0,40	84,81	0,36
CA_QT_QL_ITS	**89,24**	0,33	86,71	0,30	84,81	0,38	84,81	0,37

The selected models are the ones explained in Section 5. On those models where quantitative and qualitative scores are available (CA and FP), different models are explored based on the selected features: both (QL_QT), only quantitative (QT), only qualitative (QL).

Finally, the selected algorithm based on previous experimentation using WEKA Explorer are: 1) OneR (rule- based) and 2) Decision Table (rule-based); 3) J48 (tree-based), and 4) AttributeSelectedClassifier (meta-classifier). Note that, all algorithms were run with basic configuration. No optimization was performed.

Next section shows results on different tables. Tables compare for each model and algorithm the accuracy of the model and the RSME (root-mean-square error). Note that, models with a higher value in accuracy are better. In the case of models with similar accuracy, smaller RSME is better since they show smaller error between prediction and real value. For each model, two comparisons have been conducted. First, a 10-fold cross-validation on training set has been performed to observe how trained models fit on an independent data set. This test helps to analyze if the model

Table 3. Accuracy on Final Project Models

CROSS VALIDATION	OneR		Decision Table		J48		Atribute Selection	
	Ac. (%)	RSME	Ac. (%)	RSME	Ac. (%)	RSME	Ac. (%)	RSME
FP_QL	90,16	0,30	**90,81**	0,25	89,74	0,28	90,71	0,26
FP_QL_ITS	90,16	0,30	89,84	0,26	89,39	0,28	**91,00**	0,26
FP_QT	**92,19**	0,26	92,16	0,24	89,52	0,27	91,32	0,25
FP_QT_ITS	**92,19**	0,26	91,26	0,25	90,94	0,26	91,55	0,23
FP_QT_QL	**92,19**	0,26	92,23	0,24	90,55	0,26	91,32	0,25
FP_QT_QL_ITS	**92,19**	0,26	91,03	0,25	91,39	0,26	91,74	0,24
TESTSET	**Ac. (%)**	**RSME**	**Ac. (%)**	**RSME**	**Ac. (%)**	**RSME**	**Ac. (%)**	**RSME**
FP_QL	86,08	0,37	82,91	0,34	87,34	0,31	**91,77**	0,27
FP_QL_ITS	91,14	0,30	88,61	0,28	86,71	0,34	**93,04**	0,24
FP_QT	91,77	0,29	**91,77**	0,26	85,44	0,30	**91,77**	0,26
FP_QT_ITS	91,77	0,29	**91,77**	0,26	92,41	0,25	**91,77**	0,26
FP_QT_QL	91,77	0,29	**91,77**	0,26	89,87	0,30	**91,77**	0,26
FP_QT_QL_ITS	91,77	0,29	**91,77**	0,26	88,61	0,28	**91,77**	0,26

can be generalized. Next, a test is performed based on the training and test set to see how accurate are the models in the context of the course.

6.2 Accuracy Evaluation

As we can observe in Table 1, during the continuous assessment activity, the accuracy increases every new continuous assessment activity. On CAA1 (with no ITS information), with only 1 feature available, the model starts with an accuracy of 70%. Thus, it is difficult to make a good prediction at this stage. Subsequently with more activities, models accuracy increases by more than 10% (from CAA1 to CAA3). Models predict better on the test set by increasing accuracy and reducing RMSE. ITS information impacts differently in cross-validation and test set. However, in the test set produces better results. Thus, ITS information has a positive influence.

Table 2, when continuous assessment grade is available, shows a significant improvement in accuracy (larger than 85%). As we can observe, identical best results are obtained in cross-validation. However, results are better in the test set. It is worth to note that the model using only the continuous assessment qualitative mark predicts better that others. Here, by the contrary, the ITS information does not contribute to improve the accuracy.

In Table 3, when the final project mark is available, models reach their maximum accuracy rate ranging from 90% to 93%. Here, cross-validation can obtain its best result when the final project quantitative mark is used. However, the best result (93,04%) is obtained in the test set when only the FP qualitative mark is used with the ITS information.

Finally, we performed the last experiment only taking into account the ITS information. The best accuracy rates obtained by the OneR approach were 76,06% and 74,05% for cross-validation and test set respectively. The RSME was near 0,5 in both cases. This result implies that the information of grades is necessary to generate good prediction models.

After this experimentation, we can conclude that in most of the case the ITS information impacts positively in the models (in the CAAs and the FP). We explored

the features extracted from the ITS that impacts positively to pass the course. We observed that performing the first activity of the ITS tutorial and the first activity related to exams of previous semesters have a high impact to pass the course successfully. In this initial analysis, we conclude that starting to use the ITS system increases the chances significantly to pass the course. However, more experimentation is needed to support this hypothesis.

7 Conclusions and Future Work

In this paper, we have proposed the design of a system to predict whether a student will pass a course. This system can be used to offer different services to learners and instructors.

The system should be supported by a set of models. Here, we have proposed different model depending on the available data during the course. As we have observed, information related to the ITS system can be used on some models to improve the prediction and reduce the potential errors.

More experimentation should be performed in other to analyze the models and select the best-suited for this particular context. Additionally, an analysis to choose the best features should also be performed. This analysis can give insights of the information that has more impact in the prediction models and, therefore, in successfully passing the course. Finally, well-trained system should be used to create several services to support the instructional process of the course.

Acknowledgments. This work was funded by SOM Research group and the Spanish Government through the project: TIN2013-45303-P "ICT-FLAG: Enhancing ICT education through Formative assessment, Learning Analytics and Gamification".

References

1. Hostetler, T. R: Predicting student success in an introductory programming course. Proceedings of the fourteenth SIGCSE technical symposium on Computer science education (SIGCSE '83), 15 (3), 40-43 (1983).
2. Bergin, S.; Reilly, R.: Programming: factors that influence success. Proceedings of the 36th SIGCSE technical symposium on Computer science education (SIGCSE '05). ACM, New York, NY, USA, 411-415 (2005).
3. Sauter. V.L.: Predicting computer programming skill. Computers & Education. 10, 299-302 (1986).
4. Byrne, P., Lyons, G.: The effect of student attributes on success in programming. Proceedings of the 6th annual conference on Innovation and technology in computer science education (ITiCSE '01). ACM, New York, NY, USA, 49-52 (2001).
5. Lau, W.W.F., Yuen, A.H.K.: Modelling programming performance: beyond the influence of learner characteristics. Computers & Education. 57, 1202-1213, (2011).
6. Honour Werth, L.: Predicting student performance in a beginning computer science class. Proceedings of the seventeenth SIGCSE technical symposium on Computer science

education (SIGCSE '86), Joyce Currie Little and Lillian N. Cassel (Eds.). ACM, New York, NY, USA, 138-143 (1986).
7. Watson, C., Li, F. W. B., Godwin, J. L.: Predicting Performance in an Introductory Programming Course by Logging and Analyzing Student Programming Behavior. Proceedings of the IEEE 13th International Conference on Advanced Learning Technologies (ICALT '13). IEEE Computer Society, Washington, DC, USA, 319-323 (2013).
8. Porter, L., Zingaro, D., Lister, R.: Predicting student success using fine grain clicker data. Proceedings of the 10th annual conference on International computing education research (ICER '14). ACM, New York, NY, USA, 51-58 (2014).
9. Fire, M., Katz, G., Elovici, Y., Shapira, B., Rokach, L.: Predicting student exam's scores by analyzing social network data. Proceedings of the 8th international conference on Active Media Technology (AMT'12), Runhe Huang, Ali A. Ghorbani, Gabriella Pasi, Takahira Yamaguchi, and Neil Y. Yen (Eds.). Springer-Verlag, Berlin, Heidelberg, 584-595 (2012).
10. Sheehan, M., Park, Y.: pGPA: a personalized grade prediction tool to aid student success. Proceedings of the sixth ACM conference on Recommender systems (RecSys '12). ACM, New York, NY, USA, 309-310 (2012).
11. Ahadi, A., Lister, R., Haapala, H., Vihavainen, A.: Exploring Machine Learning Methods to Automatically Identify Students in Need of Assistance. Proceedings of the 11th annual International Conference on International Computing Education Research (ICER '15). ACM, New York, NY, USA, 121-130 (2015).
12. Eagle, M., Corbett, A., Stamper, J., McLaren, B. M., Baker, R., Wagner, A., MacLaren, B., Mitchell, A.: Predicting Individual Differences for Learner Modeling in Intelligent Tutors from Previous Learner Activities. Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization (UMAP '16). ACM, New York, NY, USA, 55-63 (2016).
13. Reddy, S., Labutov, I., Joachims, T.: Learning Student and Content Embeddings for Personalized Lesson Sequence Recommendation. Proceedings of the Third (2016) ACM Conference on Learning @ Scale (L@S '16). ACM, New York, NY, USA, 93-96, (2016).
14. Baneres, D., Clariso, R., Jorba, J., Serra, M.: Experiences in Digital Circuit Design Courses: A Self-Study Platform for Learning Support. IEEE Transactions on Learning Technologies (IEEE TTL), 7 (3), 1-15 (2014).
15. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann,, P., Witten, I. H.: The WEKA Data Mining Software: An Update, SIGKDD Explorations Newsletter, 11 (1), 10-18 (2009).

Evaluation of an eLearning Platform Featuring Learning Analytics and Gamification

David Gañán, Santi Caballé, Robert Clarisó, Jordi Conesa.
Department of IT, Multimedia, and Telecommunication, Open University of Catalonia, Barcelona, Spain
{dganan, scaballe, rclariso, jconesac}@uoc.edu

Abstract. This paper presents an initial evaluation of an innovative web-based eLearning platform, featuring Learning Analytics and Gamification called ICT-FLAG. Previous contributions presented the development and prototyping steps of the platform, and following these steps, this paper reports the evaluation on the first real experience of connecting and integrating the platform with a eLearning tool to support an online course by learning analytics services. Results of the evaluation serves as a first validation of the platform and guide the next iterative software development and evaluation process. The research reported in this paper is currently undertaken within the research project "Enhancing ICT education through Formative assessment, Learning Analytics and Gamification" (ICT-FLAG) funded by the Spanish Government.

1 Introduction

The main goal of this work is to report the implementation and evaluation phases of the construction of an eLearning platform aiming to provide learning management systems (LMS) with innovative services in terms of learning analytics and gamification. The development of this platform is the main technological goal of a research project called ICT-FLAG, which is currently undertaken in the context of university degrees in the area of Information and Communication Technologies (ICT). In this context, the ICT-FLAG platform aims to design and build eLearning tools and services that can benefit students, lecturers, managers and academic coordinators. This impact will be mainly achieved through three innovative pedagogical and technological axes:

- Formative Assessment Tools (FAT) that can provide immediate feedback by means of automatic assessment [1][2].
- Learning Analytics (LA) as a service that monitors the activity and progress of the on-line teaching and learning processes supported by e-Learning systems and applications, combining this information with other sources of academic and historical information.
- Gamification as an incentive scheme in order to motivate students to practice more frequently, more collaboratively and increase their engagement in the learning experience [3][4][5].

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_92

In addition to dealing with multiple components and services, the ICT-FLAG platform needs to support a variety of virtual campus platforms, given that universities have a wide range of alternatives, from closed-source in-house solutions to open-source platforms, such as Moodle.

Considering the above goals of the ICT-FLAG project, a modular architecture with the following features is required:

- Customization. Simple selection of which components are deployed in each installation.
- Distribution. Enable the distribution of components along different installations.
- Extensibility. Enable the easy extension of an installation with new components.

This modular architecture also needs to consider some additional architectural and application features, required to ensure extensibility and the correct performance of the system, such as [6]: dependency injection, pluggability, distribution of modules, events / messages between modules, web support (and especially REST [7]), testing support, easy deployment and configuration, and portability.

In the current status of the project, the eLearning analytics functionalities of the system were already designed. The aim of this paper is to present the implementation details and issues, and the evaluation of the constructed part of the system.

The remainder of the paper is structured as follows: Section 2 reviews some previously work on the development of learning analytics and gamification systems, as well as summarizes previous work in the context of this research project. The details of the general evaluation and validation methodology are presented in Section 3 while an initial evaluation results for the current development stage of the platform are shown and discussed in Section 4. Finally, Section 5 summarizes the main ideas and outlines the following steps in the evaluation of the ICT-FLAG platform.

2 Background

In this section, we discuss on the available tools and frameworks for applying LA and gamification in an educational context. Moreover, the state of the art in e-assessment tools is reviewed in relation to LA and gamification. Finally, we discuss how this previous work has influenced the requirements and the technical design of the ICT-FLAG platform.

2.1 Learning Analytics

Historically, Learning Analytics (LA) [8][9][10] appears from research about the processing, analysis and visualization of knowledge about the learning process. This knowledge may be gathered from large data sets of events at different levels of abstraction, such as the interactions of students with learning management systems, other students and instructors throughout the learning process [11].

In this sense, LA is built upon previous works [10] in which the interactive processing of learning data has been exhaustively analyzed [12][13]. There are many

results and applications in education, such as providing effective assessment and automatic feedback to students; supporting instructors by monitoring and predicting student performance; and modeling user profiles for content and process customization, according to the student needs, goals and individual skills.

Therefore, LA continues to evolve on interactive LA so as to apply them both in traditional eLearning systems and in new massive online course platforms (MOOCs) [13], where interaction data is generated at large or at very large scale. Large-scale data collection, processing and analysis using suitable models and reasoning can improve the use of existing eLearning systems, as well as improving the system themselves [14].

In this way, LA has addressed classical problems in eLearning, solving them at least partially [15]. However, there is a lack of integral solutions that can support all the previously mentioned aspects and incorporate them openly and transparently into current LMSs. Moreover, even though there are some common data models, their adoption for the e-Learning community is still scarce. As a consequence, each system tends to use its own model to represent student interactions, which hinders the construction of a LA model that manages information from different sources.

2.2. Gamification

"Gamification" is a term used to describe the use of game-based concepts and techniques outside recreational activities, with the goal of increasing the motivation of the participants and improving the results [16]. The benefits of gamification are an increase in motivation and engagement, which can be applied in education or work-related contexts [17]. In this sense, there are previous experiences where gamification techniques have been applied to both formal and informal learning scenarios [18][19].

On one hand, gamification offers an interesting collection of resources to increase learning engagement and motivation, in terms of achieved goals, such as reward mechanisms (scores, leaderboards, prices, achievements), progress logs ("to-do" and "done" tasks), social aspects (encouraging communication, reputation systems), feedback mechanisms and status notification (auditory or visual alerts) [18]. For example, a very simple gamification resource is an achievement system, a set of virtual badges that are awarded when a goal is achieved [20][21].

On the other hand, though promising gamification is still an emerging discipline without a strong theoretical foundation. For instance, there are no commonly accepted theoretical frameworks or general principles on how to apply gamification strategies to a specific problem and there is little empirical validation [17]. Hence it is necessary to consider the context of a particular problem in order to choose the best strategy for applying gamification strategies [22].

2.3 E-Assessment systems and tools based on LA and gamification

Most Formative Assessment Tools (FATs) offer some type of reporting mechanisms to visualize student activity and performance. However, FATs use custom mechanisms to capture, store and report information. This level of customization

allows for the definition of informative metrics that are well-suited to the learning process and the type of educational activities. On the other hand, it is hard to reuse these processes or to reuse the gathered information outside of the scope of a single FAT. For instance, it is hard to aggregate information from several tools along an academic program. In contrast, ICT-FLAG offers a generic solution that can be applied to a variety of tools with little implementation effort and can combine information from several sources.

LA can also consider additional information beyond that offered by a FAT, e.g. information provided by a LMS. Some well-known LMS, such as Moodle, have APIs that enable the interoperation with LA tools and the creation of plug-ins to embed LA functionalities [23]. LMS can also be used to incorporate gamification elements in the learning process [24], [25]. Finally, LA and gamification can be applied in the context of MOOCs [26][27][28]. However, existing LA tools are very restricted to the types of data that can be collected and analyzed, and usually the data visualization techniques are developed separately from the rest of the process of collection and analysis. Finally, the usability of these tools is limited, and because of its sophistication, is particularly complex for teachers and non-technical users to learn how to use and exploit them.

Regarding the inclusion of gamification in the current tools for e-assessment purposes, its use is still in its infancy and is based more on methodological aspects than technological ones. The inclusion of badges, for example, to reward the successful completion of an activity, does not correspond to a particular technological development [29].

2.4 Previous work

The ICT-FLAG system presented in this work is currently evolving. The decisions about the suitable technologies for the construction of the system were already discussed in [30]. The research concluded that the best technology for the construction of the system is Java EE (Java Enterprise Edition). Moreover, the analysis and design of the system was presented in [31] through the design methodology and the system's architectural view from different perspectives. Finally, [32] presented the implementation and prototyping steps. As part of the prototyping, testers integrated the ClientLib library into an e-assessment tool and sent data from tool to the platform through its APIs about the following academic indicators generated by the ICT-FLAG platform (see Table 1 for details):

- **Exercise:** considers the success rate, number of attempts and time spent on a given exercise.
- **Assignment:** aggregates the exercise information and includes information on the grades.
- **Course:** aggregates the exercise and assignment information at the course level.
- **Student**: tracks the performance of individual students along the course.

Table 1. Indicators of reports considered in the ICT-FLAG platform.

Exercise perspective
• Number of exercise submissions and results (correct / incorrect / failure). • Success rate (% of submissions with correct result). • Min / Average / Max number of attempts until getting a correct answer. • Min / Average / Max time between attempts. • Min / Average / Max duration of exercises.
Assignment perspective

• Grade distribution (alphabetic grade and min / avg / max numeric grade).
• Success rate (% of assignments with approved grade).
• Min / Average / Max duration of assignments.
• Aggregate results at the exercise level (average for all exercises).
• Show the N worst exercises in terms of:
  ○ Lowest success rate.
  ○ Longest time.
  ○ Highest number of attempts.

Course / Subject
• This report shows the same information as the assignment perspective, but showing data for all the assignments of the subject, and grouped by semester.
Students

• Show the N students with:
  ○ Highest failure rate
  ○ More attempted exercises.
  ○ Less correctly solved exercises.

Filters

• All the previous reports can be filtered by:
  ○ Program
  ○ Subject
  ○ Course
  ○ Activity
  ○ Semester
  ○ Course group
  ○ Data
  ○ Number of attempts
  ○ Tool
  ○ Exercise
• All the previous reports can be grouped by:
  ○ Tool
  ○ Program
  ○ Subject
  ○ Course
  ○ Course group
  ○ Activity
  ○ Exercise
  ○ Semester

3 Evaluation and Validation Methodology

In this section, we present a complete methodology to be used for the evaluation and validation process of our platform. Following the APA guidelines to report empirical studies (http://www.apastyle.org/manual/), information about the participants, procedure of the pilots and the apparatus used for validation are provided in this subsection:

- **Step 1:** *Goal*: To evaluate and validate the effects in the e-learning process and outcomes of ICT-FLAG. *Hypothesis*: Learners who use the ICT-FLAG platform will achieve better cognitive skills while improving drop-out rate, knowledge gain and overall satisfaction than those without using ICT-FLAG.
- **Step 2:** *Method.* Participants. General sample: (i) For formal learning pilots, the sample will consist of 60-80 university students distributed into experimental and control groups of balanced participants, also in terms of gender; (ii) informal learning pilots' sample may scale up to hundreds of participants, though the experimental group will be rather small; (iii) 1-3 lecturers and/or learning designers will evaluate the academic and technical features of the platform. The learners in all pilot sites will be supervised by tutors, assigned to the classrooms as the official instructor of the course. Apparatus and Stimuli: Learning environments: The pilots will be supported by specific e-assessment tools connected to the ICT-FLAG platform within the UOC learning management systems named Virtual Campus. Training for the appropriate use of the learning analytic and gamification services and the specific tools to visualize the resulting information will be provided at the beginning of the course. After the course, learners will be required to fill in a questionnaire, which will include test-based questions to evaluate the following ICT-FLAG aspects: (i) a valuable educational resource; (ii) technical performance (iii) system's usability (iv) emotional state when using the system and (v) overall satisfaction with the system. Moreover, open questions about the knowledge acquired during the course will be collected to evaluate the learners' cognitive progress when using the ICT-FLAG platform. Finally, all sections of the questionnaire will have a final field to express suggestions and further comments about aspects not represented in the questions. These open questions will complete the questionnaire by asking students for giving final hints for potential improvement of ICT-FLAG. Procedure: The procedure of the pilots will be driven by showcases in terms of the eLearning activities supported by the e-assessment tools.
- **Step 3:** *Evaluation Results.* To evaluate usability of the ICT-FLAG tools, the System Usability Scale (SUS) [34] will be used while for evaluating the emotional state we will use the Computer Emotion Scale (CES) [35]. Finally, for evaluating the level of worthiness of the ICT-FLAG tools, the 5 point Likert scale will be used plus open questions. For qualitative statistical analysis: Mean, Standard Deviation, Median and Chi-square.
- **Step 4:** *Validation Results.* The questionnaire will include a test asking the about the worthiness of the ICT-FLAG platform as educational tool. This part of the questionnaire will be assessed by tutors and learners using 10-

point scale to score. Statistical analysis will include Mean, Standard Deviation and Median.

- **Step 5:** *Conclusions.* The most important results with respect to the predefined research questions and goals will be discussed.

4 Evaluation and Validation Methodology

In this section, we present the first results of piloting our ICT-FLAG platform. Following our previous evaluation methodology, we propose the next steps for the experimentation (note we use a subset of the methodology showed in Section 3).

4.1 Research goals and hypotheses

The goal of this pilot is to evaluate and validate the effects of the ICT-FLAG platform integrated with an e-assessment tool named VerilUOC [34] within the real online learning context of the Open University of Catalonia. From this research goal, we formulate the following hypothesis: The VerilUOC tool with the integrated ICT-FLAG platform will provide better knowledge about what happened in the learning activities with the tool and will serve to improve the design of these learning activities than without ICT-FLAG.

4.2 Method

Following the reporting study presented in Section 3, information about the participants, the apparatus used for experimentation and the procedure of the experiment are provided in this subsection.

Participants. The sample consists of 355 university students, and 1 lecturer who evaluated the academic and technical features of the platform. The involved lecturer is the only active participant providing the experimental data while the students are passive participants involved in the experiments though without being aware of it.

Apparatus and Stimuli. The lecturer involved in the experiment used the reports of the ICT-FLAG platform connected to the VerilUOC e-assessment tool after the students finished the learning activity supported by the tool. Figure 1 shows a report created by the lecturer from using the functionality of the ICT-FLAG platform, which informed the lecturer of several indicators of the learning activity supported by the e-assessment tool (see Section 3).

After obtaining the reports, the lecturer was required to fill in a questionnaire, which included 5 test-based questions so as to evaluate the following ICT-FLAG aspects: (i) technical performance, (ii) system's usability, (iii) emotional state when using the system, (iv) the platform as a valuable educational resource and (v) overall satisfaction with the system. Moreover, open questions were addressed to the lecturer

about the knowledge acquired from using the ICT-FLAG platform after the learning activity in terms of learners' performance. Finally, all sections of the questionnaire had a final field to express suggestions from the lecturer for potential improvement of ICT-FLAG.

	Right	Wrong	Timeout	Error	Success rate	Attempts (min)	Attempts (avg)	Attempts (max)
Assignment: 146 Exercise: m3exG1	4	19	0	0	17.39	1	6.35	16
Assignment: 138 Exercise: m3exG3a	7	3	0	1	63.64	1	1.64	4
Assignment: 181 Exercise: m3exG8R	5	5	0	0	50.00	1	2.10	5
Assignment: 257 Exercise: 20131_PAC2_2A_	9	3	0	0	75.00	1	1.17	2
Assignment: 167 Exercise: m3exG9	5	15	0	0	25.00	1	5.05	13
Assignment: PAC1 Exercise: EX1	3	1	0	0	75.00	1	1.75	3
Assignment: 183								

Fig. 1. Exercise report.

For the Section ii (usability of the ICT-FLAG platform), we used the System Usability Scale (SUS) developed by [34] which contains 10 items and a 5 point Likert scale to state the level of agreement or disagreement. SUS is generally used after the respondent had an opportunity to use the system being evaluated. SUS scores have a range of 0 to 100 with an average score of 68, obtained from 500 studies. A score above 68 would be considered above average and anything below 68 is below average. A score above 80.3 is considered an A (the top 10% of scores). Scoring the mean (68) awards a C and anything below 51 is an F (putting one at the bottom 15%).

Finally, to investigate what emotional state of the lecturer, section (iii) of the questionnaire was concerned about the "emotional state" of the lecturer when using the functionality of the ICT-FLAG platform, which included 12 items of the Computer Emotion Scale (CES) [35]. CES scale is used to measure emotions related to learning new computer software by means of 12 items describing four emotions:

- Happiness ("When I used the tool, I felt satisfied/excited/curious.")
- Sadness ("When I used the tool, I felt disheartened/dispirited.")
- Anxiety ("When I used the tool, I felt anxious/insecure/helpless/nervous.")
- Anger ("When I used the tool, I felt irritable/frustrated/angry")

The answer categories in this section of the questionnaire are "None of the time", "Some of the time", "Most of the time" or "All of the time".

Procedure. The procedure of the pilots was driven in terms of reports on the performance and results of the eLearning activities supported by VerilUOC connected to the ICT-FLAG platform. The participating lecturer used the platform to generate the appropriate reports on specific learning activities of his courses at the UOC. From the reports obtained, the lecturer evaluated both the information received and the platform itself.

4.3 Evaluation results

This section shows the results collected from the aforementioned experiment where we evaluated the lecturer's satisfaction with the ICT-FLAG platform, considering the level of worthiness of the platform as well as the usability and emotional aspects while creating and using the reports generated by the platform (see Section 3).

Usability assessment. To evaluate the lecturer's satisfaction with the ICT-FLAG platform as regards efficient and user-friendly management, we collected data from students' ratings and open comments on the usability/functionality/integration of the tool. After calculating the SUS score for the participating lecturer, we obtained a score of *87.50* thus much above the SUS mean (class A). In particular, the lecturer thought the ICT-FLAG platform was easy to use (strongly agree). He did not find the tool unnecessarily complex (strongly disagree). In addition, he stated that he did not need the support of a technical person to use the platform (strongly disagree); he thought that most people would learn to use the platform very quickly (strongly agree) and it was easy to use (strongly agree). Finally, the lecturer did not think there was too much inconsistency in the platform (strongly disagree) and he did not find it very cumbersome to use (strongly disagree).

Emotional assessment. Regarding the lecturer' emotions during his work with the ICT-FLAG platform, the results from a 4-point rating scale are as follows: Happiness (Some of the time); Sadness (None of the time); Anxiety (Some of the time) and Anger (None of the time). In sum, the lecturer felt more often happiness than sadness, or anger when using the new platform. The most noticeable result is that the values of Sadness and Anger emotions were inappreciable (None of the time).

4.4 Validation results

In this section, we evaluate the level of worthiness of the ICT-FLAG platform as an educational tool. To this end, we collected quantitative and qualitative data of the questionnaire that included six open questions addressed to the lecturer participating in the experiment. The rating scales for the quantitative questions were based on the usual 0-10 point scale. The rating scale went from the lowest (0) to the highest (10) considering a "good" assessment 5 and above, and a "bad" assessment below 5.

1. *What did you like and what you did not like from the ICT-FLAG platform (score 0-10 and comments).*
   Score: 7. Comments: "ICT-FLAG is easy to use and give a global vision of the course performance. I would like to see more fine-grained information about the course performance. Global information can be easily obtained in different ways".

2. *Do you think the ICT-FLAG platform has helped you acquire more knowledge about the learning activities supported by VerilUOC tool? (score 0-10 and comments)*
   Score: 7. Comments: "ICT-FLAG will be very useful when a tool does not have a frontend or a learning analytics (LA) module. In the the VerilUOC case, the tool has a LA tool where more fine-grained information is given. In VerilUOC case, no new knowledge is provided".

3. *Do you think the ICT-FLAG platform has helped you to identify problems and improve the learning activities supported by VerilUOC tool? (score 0-10 and comments)*
   Score: 0. Comments: "No, no new problems have been identified. All problems related to the number of tries or performance are currently given by the Learning Analytics module in VerilUOC".

4. *Express your opinion about the assessment reports generated by the ICT-FLAG platform in terms of efficiency and performance (score 0-10 and comments)*
   Score: 7. Comments: "The efficiency is sufficient since each page provided different repots at the same time".

5. *Let us know your opinion about the potential of the ICT-FLAG platform to assess students and the learning process when connected to an e-assessment tool (such as VerilUOC) (score 0-10 and comments).*
Score: 9. Comments: "The tool has a large potential mostly in real time. More reports can be provided in the future to give more detailed information related to the performance of the course. Moreover, comparison among different editions could be also valuable information".

6. *Compare the advantages and disadvantages of the ICT-FLAG platform to those of other e-learning systems offering similar e-assessment reports (e.g., Moodle reports, etc). Score 0 – 10 and comments.*
Score: 2. Comments: "I don't know Moodle. Therefore, I cannot compare with them. Compared to some MOOC platforms (ie. edx Insights), the reports of ICTFLAG platform are currently very limited. However, the objectives are slightly different".

Overall the questions obtained the positive statistics, as follows: Mean: 5.3; Standard deviation: 3.5; Median: 7. As a general comment, the lack of compatibility with different web browsers was seen as an issue to consider.

4.5 Summary

In order to sum up the previous results, we conclude as follows. The usability test of the ICT-FLAG platform was judged very satisfactory in line with the general SUS score obtained. Regarding the emotional test, we consider it as a good result taking into account that the lecturer faced a complex type of reports that was new for him and he had to learn the platform' functionality and how to use it for his own benefit. This result is in line with the result concerning usability presented above. Finally, the platform was validated from the perspective as a valuable educational tool though the open comments shown improvements to be considered in terms of limited reports and their low repercussion when used in conjunction with FAT tools with LA features.

5. Conclusions and Future Work

This paper presented the results of a naïve test for evaluating a Learning Analytics and gamification platform for eLearning called ICT-FLAG. Even though the evaluation and validation results are positive in terms of functional and non functional aspects, from the comments and suggestions collected in the experiments, certain negative issues were mentioned that will be addressed in next development iterations of the platform. Main improvement is to extend the variety of assessment reports currently offered by the platform and include gamification services. In addition, the test did not evaluate important aspects, such as scalability, modularity and security. These and other features will be evaluated in next experiments at larger scale of participants both students and tutors and integrating different FAT tools.

Acknowledgements. This research was partially funded by the Spanish Government through the project TIN2013-45303-P "ICT-FLAG" (Enhancing ICT education through Formative assessment, Learning Analytics and Gamification).

References

1. Whitelock, D. Computer Assisted Formative Assessment : Supporting Students to Become More Re ective Learners. In 8th International Conference on Computer Based Learning in Science, CBLIS, pp. 492 – 503 (2007).
2. Nicol, D., & Macfarlane-Dick, D. Formative Assessment and Self-Regulated Learning: A Model and Seven Principles of Good Feedback Practice. Studies in Higher Education, 31 (2), 199 - 218 (2006).
3. Deterding, S., Dixon, D., Khaled, R., & Nacke, L. From game design elements to gamefulness: defining gamification. In Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments, pp. 9-15 (2011).
4. Domínguez, A., Saenz-de-Navarrete, J., De-Marcos, L., Fernández-Sanz, L., Pagés, C., & Martínez-Herráiz, J. J. Gamifying learning experiences: Practical implications and outcomes. Computers & Education, 63, 380-392 (2013).
5. Amriani, A., Aji, A. F., Utomo, A. Y., & Junus, K. M. An empirical study of gamification impact on e-Learning environment. In Computer Science and Network Technology (ICCSNT), 2013 3rd International Conference on, pp. 265-269 (2013).
6. Yener, M., & Theedom, A. Professional Java EE Design Patterns. John Wiley & Sons (2014).
7. Mehta, B. RESTful Java Patterns and Best Practices. Packt Publishing Ltd (2014).
8. Siemens, G., & d Baker, R. S. Learning analytics and educational data mining: towards communication and collaboration. In Proceedings of the 2nd international conference on learning analytics and knowledge, pp. 252-254 (2012).
9. Ferguson, R. Learning analytics: drivers, developments and challenges. International Journal of Technology Enhanced Learning, 4(5-6), 304-317 (2012).
10. Romero, C., & Ventura, S. Educational data mining: a review of the state of the art. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 40(6), 601-618 (2010).
11. Romero, C., & Ventura, S. Educational data mining: A survey from 1995 to 2005. Expert systems with applications, 33(1), 135-146 (2007).
12. Fahy, P. J. Online and face-to-face group interaction processes compared using Bales' interaction process analysis (IPA). European Journal of Open, Distance and E-Learning, 9(1) (2006).
13. Siemens, G. Massive open online courses: Innovation in education. Open educational resources: Innovation, research and practice, 5 (2013).
14. Caballé, S., & Xhafa, F. Distributed-based massive processing of activity logs for efficient user modeling in a Virtual Campus. Cluster computing, 16(4), 829-844 (2013).
15. Isabel Guitart, Jordi Conesa. Creating University Analytical Information Systems: A Grand Challenge for Information Systems Research. In Formative Assessment Learning Data Analytics and Gamification (pp. 167-186). Elsevier. ISBN: 978-0-12-803637-2 (2016).
16. Salen, K., & Zimmerman, E. Rules of play: Game design fundamentals. Cambridge, MA: MIT Press (2004).
17. Hamari, J., Koivisto, J., & Sarsa, H. Does gamification work?--a literature review of empirical studies on gamification. In System Sciences (HICSS), 2014 47th Hawaii International Conference on, pp. 3025-3034 (2014).

18. Lee, J. J., & Hammer, J. Gamification in education: What, how, why bother?. Academic exchange quarterly, 15(2), 146 (2011).
19. Borrás Gené, O., Martínez Núñez, M., y Fidalgo Blanco, Á. Gamification in MOOC: challenges, opportunities and proposals for advancing MOOC model. In F. J. García-Peñalvo (Ed.), Proceedings of the Second International Conference on Technological Ecosystems for Enhancing Multiculturality, pp. 215-220 (2014).
20. Denny, P. The effect of virtual achievements on student engagement. In Proceedings of the SIGCHI conference on human factors in computing systems, pp. 763-772 (2013).
21. Hakulinen, L., Auvinen, T., & Korhonen, A. Empirical study on the effect of achievement badges in TRAKLA2 online learning environment. In Learning and Teaching in Computing and Engineering (LaTiCE), pp. 47-54 (2013).
22. Groh, F. Gamification: State of the art definition and utilization. Institute of Media Informatics Ulm University, 39 (2012).
23. Dimopoulos, I., Petropoulou, O., Boloudakis, M., & Retalis, S. Using Learning Analytics in Moodle for assessing students' performance (2013).
24. HENRICK, G. Gamification-What is it and What it is in Moodle (2010)
25. Dicheva, D., Dichev, C., Agre, G., & Angelova, G. Gamification in education: a systematic mapping study. Educational Technology & Society, 18(3), 1-14 (2015).
26. Greller, W., & Drachsler, H. Translating Learning into Numbers: A Generic Framework for Learning Analytics. Educational technology & society, 15(3), 42-57 (2012).
27. Chatti, M. A., Dyckhoff, A. L., Schroeder, U., & Thüs, H. A reference model for learning analytics. International Journal of Technology Enhanced Learning, 4(5-6), 318-331 (2012).
28. Gené, O. B., Núñez, M. M., & Blanco, Á. F. Gamification in MOOC: Challenges, opportunities and proposals for advancing MOOC model. In Proceedings of the Second International Conference on Technological Ecosystems for Enhancing Multiculturality, pp. 215-220 (2014).
29. Gibson, D., Ostashewski, N., Flintoff, K., Grant, S., & Knight, E. Digital badges in education. Education and Information Technologies, pp. 1-8. Sringer (2013).
30. Gañan, D., Caballé, S., Clarisó, R., Conesa, C.(2016). Towards the Effective Software Development of an eLearning Platform Featuring Learning Analytics and Gamification. 4th International In proceedings of the IEEE 30th International Conference on Advanced Information Networking and Applications Workshops, pp. 177-182 (2016).
31. Gañan, D., Caballé, S., Clarisó, R., Conesa, J. Analysis and Design of an eLearning Platform Featuring Learning Analytics and Gamification. 10th International Conference on Complex, Intelligent, and Software Intensive Systems, pp. 87-94.
32. Gañan, D., Caballé, S., Clarisó, R., Conesa, J. Implementation and Prototyping of an eLearning Platform Featuring Learning Analytics and Gamification. In proceedings of the 10th International Conference on Complex, Intelligent, and Software Intensive Systems, pp. 87-94 (2016).
33. Baneres, D., Clariso, R., Jorba, J., Serra, M., Experiences in Digital Circuit Design Courses: A Self-Study Platform for Learning Support. Learning Technologies, IEEE Transactions on, vol.7, no.3, pp.1-15, (2014)
34. Brooke, J. SUS: A "quick and dirty" usability scale. Usability evaluation in industry. London: Taylor & Francis, (1996).
35. R.H. Kay and S. Loverock, Assessing emotions related to learning new software: The computer emotion scale. Computers in Human Behavior, 24, pp. 1605-1623 (2008).

MobilePeerDroid: A Platform for Sharing, Controlling and Coordination in Mobile Android Teams

Fatos Xhafa, Daniel Palou, Santi Caballé, Keita Matsuo and Leonard Barolli

Abstract The fast development in mobile technologies and the continuous improvements in smart mobile devices, propelled by wide adoption of Mobile Cloud computing, have prompted for the development of a new generation of mobile applications aimed at supporting team work. Such applications provide great flexibility to workforce, especially groups of people developing teamwork both geographically distributed and on the move. The ability of being able to work according to the A3 paradigm, namely, Anytime, Anywhere, Awareness, significantly improves the teamwork performance. Relevant examples of such applications are found in businesses–especially Small and Medium size Enterprises–, online collaborative learning in Virtual Campuses, Social Networking, disaster management scenarios, etc. However, designing and developing such mobile collaborative applications requires overcoming many issues and challenges. Among these, there are three main pillars of such applications, namely, sharing, controlling and coordination. In this paper, we present MobilePeerDroid a platform that support sharing, controlling and coordination in mobile android team. The platform is based on P2P architecture to achieve flexibility, efficiency and scalability. The platform has been instantiated for teams of members using mobile devices under Android. In addition, it enables extensions (via separation of concerns and modularity) with new features for supporting new application development.

Fatos Xhafa
Universitat Politècnica de Catalunya, BarcelonaTech, Barcelona, Spain, e-mail: fatos@cs.upc.edu

Daniel Palou
Universitat Politècnica de Catalunya, BarcelonaTech, Barcelona, Spain, e-mail: dpalou@cs.upc.edu

Santi Caballé
Universitat Oberta de Catalunya, Barcelona, Spain, e-mail: scaballe@uoc.edu

Keita Matsuo
Fukuoka Institute of Technology, Fukuoka, Japan, e-mail: kt-matsuo@fit.ac.jp

Leonard Barolli
Fukuoka Institute of Technology, Fukuoka, Japan, e-mail: barolli@fit.ac.jp

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_93

1 Introduction

With the fast development in mobile technologies and Cloud computing, we are seeing a shift from web-based applications to mobile cloud applications. Indeed, with the ever increasing use of mobile devices, their significant improvements in computational and storage capacities, web-based applications are leaving the room to mobile applications. This shift is expected to be accelerated each time more due to the combination of mobile and cloud computing. An important family of web-based applications that is seeing such shift is that of collaborative applications, which were design in early 90' with web technology to cover needs of teamwork in business and computer-supported collaborative learning. A widely used web-application is that of BSCW System [1], which supports cooperative and collaborative work according to various online learning and working paradigms. The emergence and unprecedented speed of social networks and Web 2.0 leaded to the new generation of collaborative applications, namely Web 2.0 applications according to the shift from CSCW to Web 2.0 [20].

Currently, with mobile computing increasingly penetrating to all application domains, we are witnessing a new shift of collaborative applications to mobile collaborative applications. Such applications besides offering the services of web-based applications accessed from wired devices commented to Internet, enable to overcome limitations of mobility [35, 34], namely, by using mobile devices (such as PDAs, smartphones, etc.) collaborative work can be carried out according to the A3 paradigm (*Anytime*, *Anywhere*, *Awareness*) [29, 19, 21]. This paradigm states that a team of people geographically distributed can collaborate at anytime due to continuous access to Internet and the collaborative application, can collaborate from anywhere and can be aware of events happening in the shared collaborative work. Altogether, makes the collaboration much more effective and significantly increases the performance of the teamwork as compared with 'traditional' web-based collaborative applications. Essentially, mobile collaborative applications shorten drastically the time to know, the time to act and the response time.

The design and development of collaborative applications in old and new forms, have always required overcoming many issues and challenges. Among these, there are three main pillars of such applications, namely,the *sharing*, *controlling* and *coordination*, which were addressed in early CSCL applications, web-based application [7, 11, 23, 24, 9]. These issues are further addressed for other computing paradigms that are used for supporting collaborative teamwork such as P2P and Grid computing [10, 22] and more recently in Mobile and Cloud Computing. In this later direction however there are still few works that exploit the advances in mobile cloud computing, especially the combination with P2P technologies, offered by recent libraries such as PeerDroid for the Android system [18, 12].

In this paper, we present the MobilePeerDroid, a platform for sharing, controlling and coordination in mobile Android teams. We analyse the main requirements, the design and implementation of the platform. Our main objective is to support smart devices, not only smartphones and tablets but also other devices such as robots to be able to join the platform as useful means for teamwork, sharing, controlling and coordination. The paper is based and refers to previous publications by the authors about the topic [32, 33].

The rest of the paper is organized as follows. We present the main requirements in Section 2. The design and architecture of the application is presented in Section 3. The implementation, user interface and some preliminary evaluation is presented in Section 4. We end the paper in Section 5 with some conclusions and remarks for future work.

2 Requirements

There are several requirements that a mobile application should satisfy to efficiently support teamwork. We briefly analyse them next.

2.1 Sharing

It is also fundamental for a teamwork to share information and the project state, which could be in various forms such as documents and files [26], workplan, messages, discussions, and collaboration agenda. Among them, the collaboration agenda is a key information to collaborative activity as it guides the members to accomplish pending tasks, organize online meetings, online discussions, voting, etc. Members of the team can have their own agenda while the team has the common agenda to all members. Both agendas should be synchronized and affected members informed.

2.2 Collaboration

Unlike many mobile applications where only individual actions are supported, teamwork requires collaboration as a fundamental requirement. Indeed, teamwork has a goal such as accomplishment of a common project, which therefore require the participation of team members. Different members are assumed to be in charge and accomplish specific tasks within the common project. The teamwork goal is thus achieved through team members collaboration.

2.3 Emotion awareness

Technology holds enormous potential to measuring emotions in working, learning and performance [4]. Specifically, affective computing [5] comprises an emerging set of innovations that allow systems to recognize, interpret and simulate human emotions. While current applications mainly focus on capturing and analysing emotional reactions of individual users to improve the efficiency and effectiveness of product or media testing, this technology holds great promise for developing social awareness, communication, collaboration and emotional skills, such as greater empathy, improved self-awareness and stronger relationships. In addition computers, smartphones, and global communications have shaped and educated the current generation of learners and workers as they are active and often enthusiastic participants in the creation of online communities since early childhood, indicating that we are moving from an individualistic knowledge acquisition culture towards a collaborative knowledge creation culture of work [6].

Nonetheless, the digital revolution has not transformed working environments, which are still struggling with incorporating emotional-aware tools despite major advancements in fields such as artificial intelligence, human-computer interaction or sensorial technologies resulting, in many cases, where, for instance, informal learning is much more engaging and effective than formal traditional learning [4].

2.4 Awareness

This is another important requirement to support teamwork. Indeed, members of a team need to be informed on the action carried out by other team members and re-act accordingly. Awareness can take various forms such as availability awareness (availability of resources, availability of members, ...), context awareness (where the actions/changes took place), group process awareness (changes with regard to project workflow), etc. [29, 22]. Implementation of these awareness forms enables a timely and effective collaboration and coordination of teams.

2.5 Coordination

The Android system has a calendar system, the Google Calendar, that allows teamwork and event management. Therefore, mobile applications that use mobile devices under Android can incorporate Google calendar. The drawback however is that this forces the members of the team to use Google accounts, which may not be desirable. Besides, by not using the Google calendar, we can be more flexible in implementing a proper calendar with functions and options more focused on teamwork, indeed, Google calendar does not properly support team formation, projects and coordination.

Among desired features, our software will allow users to create and modify their own calendars that contain events and tasks. Users can share certain tasks and events with other users of the team, they can participate and appear in several calendars (typically there would be a calendar per project), or directly share a whole calendar either with the whole team or just a sub-team. If a user has sufficient permissions, he can also modify an event or a task, and these changes will be propagated to all users participating in the event or calendar. Thus, implementing the proper teamwork calendar can be seen as corner-stone for the coordination of members of the team.

2.6 Synchronization

Synchronization is a key element in teamwork with common tasks and goals. Generally speaking, data synchronization aims to ensure exactly the same data in two or more different locations. In the case of our mobile application for teamwork, we want the data to be the same at different mobile phones (members of the team), that is, when a user makes a change this should be reflected on all devices that contain the same data.

There are many aspects that influence the synchronization, such as data storage strategy. By choosing to have the data spread across different devices, when a user makes a change to the data in its device, the change must be communicated directly to the other affected devices; or, the data can be stored centrally on a server and all the synchronization is made through it. In the former case, synchronization is more challenging due to the fact that the mobile devices can go on and off and data consistency is a real issue as it takes more time for the system to reach a consistent state when changes occur. In the later case, using a server, it is easier and more reliable to manage the synchronization and achieve desired degree of consistency.

There are various strategies for synchronizing data. The most used ones are full and incremental synchronization.

2.6.1 Full synchronization

The full synchronization requires always all data to be synchronized. That is, if a device wants to be synchronized with server information, the server will always send all the data to the device, and it is the device that should contrast that data received from the server against its data and update it accordingly. This synchronization can be seen as *safe mode synchronization* in terms of synchronization errors, because by receiving all the information of the server side we can be sure that the device will contain exactly the information of the server. Obviously, this synchronization method implies a high cost of temporary data traffic, and it should be noted that part of this data traffic is redundant as the device has already part of it (usually small changes are done on the data). However, this kind of synchronization is necessary

when a user connects for the very first time to the system or when it connects after a long period of time. In other cases, an incremental synchronization can used (see next).

2.6.2 Incremental synchronization

The incremental synchronization requires knowing only the data that has changed since the last synchronization. Thus, a device receives only the changes occurred, not all information. In this case it is more likely to have a synchronization error or a higher degree of inconsistency. For example, if for some reason the synchronization fails, but through the process we update just the time/date of the last synchronization, the synchronization would be considered completed correctly. Clearly, this synchronization is more error-prone in case of mobile devices but on the other hand it is much more efficient both in space and in traffic since it only sends the necessary information, not redundant data. Due to possible synchronization errors can happen, it has been decided to maintain full synchronization option when prompted by the user, as in this way can easily recover from this type of errors. Thus, if needed, and sporadically, the user can request a full synchronization with the server.

2.6.3 Conflicts

During the synchronization process conflicts can happen when two users may be modifying the same data. In these cases, it should be decided what data version should be considered and accepted as correct. There are many strategies to resolve conflicts. In our case, and due to the importance of correct conflict resolution to the collaboration of the teamwork, we decided to leave the conflict resolution to the users choice. When a conflict occurs, the affected users will be prompted to decide what values they'd like to be saved, and a decision is taken by a simple majority vote. In case of a tie, the voting will be repeated.

2.7 P2P communication

As explored in many other research works [19, 16, 31], P2P architecture provide the basis for efficient, scalable and flexible communication and performance. However, in most works so far in the literature the P2P paradigm has not much been explored for Mobile platforms, which is attracting each time more interest due to the significant increase in processing, connection and storage capacities of smart mobile devices.

2.8 Controlling services

The MobilePeerDroid is conceived to be able to support connecting not only mobile phones and tablets but also other smart devices that can be useful for collaborative work. In this regard, the platform enables to join other IoT devices and control them from mobile phones (peers). Such IoT devices comprise SmartBox, Robots and other smart devices [17, 14, 2].

3 Architecture and Technologies

In order to achieve the main requirements listed in Sect. 2, the architecture should be flexible, adaptive and inter-operable. Previous work have pointed out architectural concerns in distributed and mobile collaborative systems [8, 25]. Among them, achieving integration of process- and workspace management with Peer-to-Peer (P2P) Middleware, Publish-Subscribe, and Community and User Management are identified as key requirements. Therefore, the architecture of the software system supporting mobile teamwork should account for web servers, services, super-peer implementation and mobile peers. Some issues in developing such architectures have been addressed in [27, 10, 13].

The architecture pattern used is that of Model View ViewModel (see Fig. 1 a) and the P2P super-peer architecture (see Fig. 1 b).

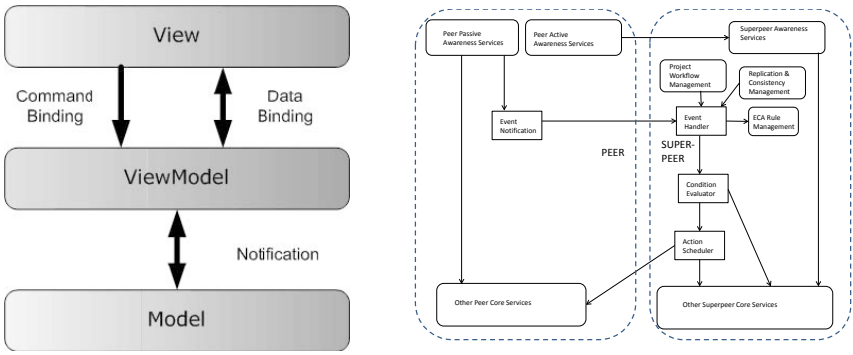


Fig. 1 a) Model View ViewModel Pattern and b) P2P super-peer architecture, the communication between a peer and a super-peer [19].

3.1 Implementation and technologies

Several core technologies have been employed in developing and implementing the system.

- **Java**: The system is designed to run on Android, whose programming language is Java.
- **SQLite**: We use SQLite as a local database because the resources are quite limited. Storing data in files can also be considered, but it was decided to use the database as a more effective solution.
- **Android**: Because we want our application to use the maximum number of available devices, it was decided that the application will be implemented to be compatible with this version. Nevertheless, the system can be fully implemented using other operating systems.
- **Android-Binding**: We use this external open source library to achieve a greater independence between the view and our system using the MVVM pattern.
- **SugarCRM**: SugarCRM is commonplace in applications for SMEs businesses. By choosing this application server, we simulate the case of a real project development in enterprises and make our application potentially useful for real setting.
- **PHP**: The server functionality must be extended in order to adapt it to our system. Because SugarCRM is implemented in PHP, it is then the programming language used to expand it.
- **Memcached**: Memcached is a distributed system that is used for storing cached data or objects, thus reducing the need to access an external data source.
- **JXTA**: JXTA (Juxtapose) is a P2P open source platform created by Sun Microsystems and defined as a set of XML-based protocols. These protocols allow devices connected to a network to exchange messages with each other regardless of the network topology. The definition of JXTA is abstract and based on a set of open protocols so that it can be brought to any modern programming languages. There are several implementations, the most advanced of which is JXSE the Java version. The version is called Java Micro Edition JXME. One important feature of JXTA is that it allows P2P communication even when peers are behind NATs or firewalls.
- **PeerDroid**: PeerDroid is an open source library that implements protocols JXME for the Android platform. It allows to create applications for the Android platform using JXTA properties, creating a network of mobile and other traditional peers (computers, for example).
- **Collaborative Complex Learning Resource (CC-LR)**: CC-LR [3] is an educational technology that enhances collaborative and social learning by virtualising the knowledge elicited during the live collaborative sessions in order to alleviate the isolation of online team members whilst providing them with social identity, empowerment and effective interactivity. CC-LR supports complex types of interactions, such as emotional and cognitive, providing relevant knowledge from all possible sources, increasing the interaction behaviour, and eventually improving the quantitative and qualitative learning analytics processes and outcomes.

As a result, online learners are provided with challenging and attractive resources that engage and self-motivate them to collaborate during the learning process.

4 Graphical User Interface and Evaluation

The system has been deployed at the RDLab infrastructure (http://rdlab.cs.upc.edu/), using SugarCRM server. It should be noted however that SugarCRM does not implement transactions nor semaphores. The later have been implemented through an external cache and used it to share variables among all calls. Mobile users were implemented under Android system (see Figs. 2 and 3 for snapshots of login, user notification preferences, calendar creation, task creation and task filter).

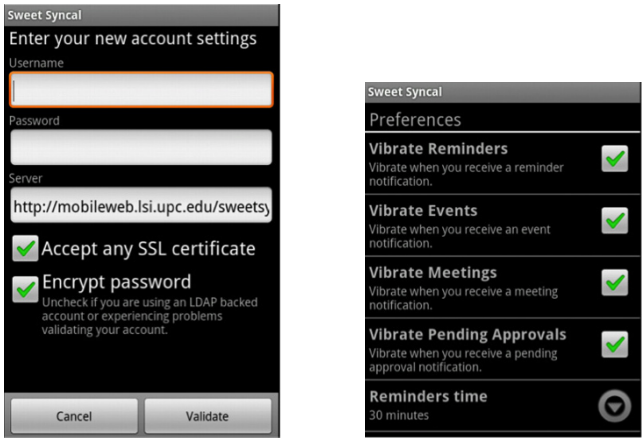


Fig. 2 a) Login snapshot and b) User notification preferences snapshot.

The system has been evaluated according to different usability and performance criteria. We have also done other evaluations such as the evaluation of the P2P synchronization *vs.* server-based synchronization. In the former, any changes occurred at a peer are propagated directly to all affected peers, while in the later, the synchronization is done by the server mediation. In this case, we measure the time needed to synchronize the mobile devices and the generated traffic (amount of messages and their size. Several replication scenarios using real peers and emulated peers were generated.

5 Conclusions and future work

In this paper we have presented MobilePeerDroid, a platform that support sharing, controlling and coordination in mobile android team. The platform is based on P2P

Fig. 3 a) Calendar creation snapshot; b)Task creation snapshot and c) Calendar / task filter view snapshot.

architecture to achieve flexibility, efficiency and scalability. The platform has been instantiated for teams of members using mobile devices under Android however it enables extensions with new features for supporting new application development as well as new types of smart devices such as robots, smartboxes, cameras and smart sensors. We have briefly analysed the main requirements, design and implementation. A prototype is developed using a combination of architectures and technologies (web, services, P2P, Android, PeerDroid, etc.) to support collaborative teamwork through calendars, task creation, awareness, event notification services, etc. We have also considered the precedences among tasks as a initial approach towards workflow implementation to support complex scenarios beyond independent or loosely coupled tasks. Several issues were identified during the application development, such as the synchronization, for which the P2P *vs.* server-mediated synchronizations were evaluated. A preliminary evaluation of the platform was done to get insights from both end user perspective and computational performance perspective.

In our future work we would like to fully evaluate the system by deploying it in a real Virtual Campus to support collaborative learning paradigms such as Project-Based Learning.

Acknowledgements This work has been supported by the Spanish ICT-FLAG Research Project TIN2013-45303-P (Enhancing ICT education through Formative assessment, Learning Analytics and Gamification) and Spanish Ministry for Economy and Competitiveness (MINECO) and the European Union (FEDER funds) under grant COMMAS (ref. TIN2013-46181-C2-1-R)

References

1. Appelt, W. WWW Based Collaboration with the BSCW System. In Proceedings of the 26th Conference on Current Trends in Theory and Practice of Informatics on Theory and Practice of Informatics (SOFSEM '99), Jan Pavelka, Gerard Tel, and Miroslav Bartosek (Eds.). Springer-Verlag, London, UK, 66-78 (1999)
2. Barolli, L., Spaho, E., Xhafa, F. and Younas, M. Application of SmartBox end-device for medical care using JXTA-Overlay P2P system. Computing 95(10-11): 1039-1051 (2013)
3. Caballé, S., Mora, N., Feidakis, M., Ganán, D., Conesa, J., Daradoumis, T., Prieto, J. CC-LR: Providing Interactive, Challenging and Attractive Collaborative Complex Learning Resources. Journal of Computer Assisted Learning, 30(1), 5167 (2014)
4. Calvo R. and D'Mello, S. Frontiers of Affect-Aware Learning Technologies. IEEE Intelligent Systems. 27(6), 86-89 (2012)
5. Calvo, R., D'Mello, S., Gratch, J., Kappas, A. The Oxford Handbook of Affective Computing. New York: Oxford University Press (2015)
6. Lutchyn, Y., Johns, P., Roseway, A. and Czerwinski, M. MoodTracker: monitoring collective emotions in the workplace. Proceedings of International Conference on Affective Computing and Intelligent Interaction (ACII), 295 - 301 (2015)
7. Dourish, P. and Bellotti, V. Awareness and coordination in shared workspaces. Proceedings of the 1992 ACM conference on Computer-supported cooperative work. ACM Press New York, NY, USA, 107-114 (1992)
8. Dustdar, S. and Gall, H. Architectural concerns in distributed and mobile collaborative systems, Parallel, Distributed and Network-Based Processing, 2003. *Proceedings of the Eleventh Euromicro Conference*, 475-483 (2003)
9. Franco, D. F., Neyem, A., Ochoa, S., Navarro, R. Supporting Mobile Virtual Team's Coordination with SOA-Based Active Entities. Establishing the Foundation of Collaborative Networks. *The International Federation for Information Processing* Volume **243**, 557-564 (2007)
10. Ghezzi, C, Cugola, G. and Picco, G. P. PeerWare: A Peer-to-Peer Middleware for Mobile TeamWork. *ERCIM News* No. **54** (2003)
11. Greenberg, S. and Marwood, D. Real time groupware as a distributed system: concurrency control and its effect on the interface. Proceedings of the 1994 ACM conference on Computer supported cooperative work. ACM Press New York, NY, USA, 207217 (1994)
12. Jabbar, W.A., M. Ismail, M. and Nordin, R. Peer-to-Peer Communication on Android-Based Mobile Devices: Middleware and Protocols. 2013 5th International Conference on Modeling, Simulation and Applied Optimization (ICMSAO), 1-6, IEEE CPS (2013)
13. Kirda, E., Fenkam, P., Reif G., and Gall, H. A service architecture for mobile teamwork. In Proceedings of the *14th international conference on Software engineering and knowledge engineering (SEKE '02)*. ACM, New York, NY, USA, 513-518 (2002)
14. Kolici, V., Spaho, E., Matsuo, K., Caball, S., Barolli, L and Xhafa, F. Implementation of a Medical Support System Considering P2P and IoT Technologies. CISIS 2014: 101-106, IEEE CPS (2014)
15. Leonardi, L., Mamei, M. and Zambonelli, F. A Physically Grounded Approach to Coordinate Movements in a Team. In *Proceedings of the 1st International Workshop on Mobile Teamwork* (2002)
16. Liu, Y., Sakamoto, Sh., Matsuo, K., Ikeda, M., Barolli, B. and Xhafa, F. A comparison study for two fuzzy-based systems: improving reliability and security of JXTA-overlay P2P platform. Soft Comput. 20(7): 2677-2687 (2016)
17. Ogata, Y., Matsuo, K., Spaho, E., Barolli, L. and Xhafa, F. Implementation of SmartBox End-Device for a P2P System and Its Evaluation for E-Learning and Medical Applications. BWCCA 2010: 794-799, IEEE CPS (2010)
18. PeerDroid: http://code.google.com/p/peerdroid/
19. Poulovassilis, A., Xhafa, F. and O'Hagan, T. Event-Based Awareness Services for P2P Groupware Systems. Informatica, Lith. Acad. Sci. 26(1): 135-157 (2015)

20. Randall, D. and Salembier, P. (Eds.) From CSCW to Web 2.0: European Developments in Collaborative Design, Selected Papers from COOP08, Springer (2010)
21. Rivadulla, X., Xhafa, F. and Caball, S. An Event-Based Approach to Support A3: Anytime, Anywhere, Awareness in Online Learning Teams. CISIS 2012: 947-952, IEEE CPS (2012)
22. Sapateiro, C., N. Baloian, P. Antunes and G. Zurita (2009) Developing Collaborative Peer-to-Peer Applications on Mobile Devices. Proceedings of the *13th International Conference on Computer Supported Cooperative Work in Design (CSCWD 2009)* Santiago, Chile (2009)
23. Shen, H.H. and Dewan, P. Access control for collaborative environments. Proceedings of the 1992 ACM conference on Computer-supported cooperative work. ACM Press New York, NY, USA. 51-58 (1992)
24. Sohlenkamp, M. and Chwelos, G. Integrating communication, cooperation, and awareness: the DIVA virtual office environment. Proceedings of the 1994 ACM conference on Computer supported cooperative work. ACM Press New York, NY, USA, 331343 (1994)
25. Su, X., Prabhu, B. S., Chi-Cheng Chu, Gadh, R. Middleware for Multimedia Mobile Collaborative System. I*EEE ComSoc 3rd Annual Wireless Telecommunication Symposium (WTS 2004)*, CalPoly Pomona, Pomona, California, USA (2004)
26. Xhafa, F., Barolli, L., Fernandez, R., Daradoumis, Th., Caball, S., Kolici, V. Extending JXTA Protocols for P2P File Sharing Systems. *CISIS 2008*: 35-40, IEEE CPS (2008)
27. Xhafa, F., Barolli, L., Caballé S., Fernandez, R. Efficient peerGroup management in JXTA-Overlay P2P system for developing groupware tools. *The Journal of Supercomputing* **53**(1): 45-65 (2010)
28. Xhafa, F., Barolli, L., Caballé, S., Fernandez, R. Supporting Scenario-Based Online Learning with P2P Group-Based Systems. *NBiS 2010*:173-180, IEEE CPS (2010)
29. Xhafa, F. and Poulovassilis, A. Requirements for Distributed Event-Based Awareness in P2P Groupware Systems. *AINA Workshops 2010*: 220-225, IEEE (2010)
30. Xhafa, F., Palou, D., Barolli, L, Spaho, S. and Takizawa, M. Coordination in Android Mobile Teams. NBiS 2013: 482-487, IEEE CPS (2013)
31. Xhafa, F., Potlog, A., Spaho, E., Pop, F., Cristea, V. and Barolli, L. Evaluation of intra-group optimistic data replication in P2P groupware systems. Concurrency and Computation: Practice and Experience 27(4): 870-881 (2015)
32. Xhafa, F., Palou, D. Caballé, S. and Barolli, L. On Sharing and Synchronizing Groupware Calendars Under Android Platform. 2016 10th International Conference on Complex, Intelligent, and Software Intensive Systems (CISIS 2016), 122-129, IEEE CPS (2016)
33. Xhafa, F., Palou, D., Barolli, L., Takizawa, M. Energy-aware Analysis of Synchronizing a Groupware Calendar. 2016 10th International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS 2016), 215-22, IEEE CPS (2016)
34. Yang, Y., Supporting online Web-based teamwork in offline mobile mode too. *Proceedings of the First International Conference on Web Information Systems Engineering*, Vol. **1**, 486-490 (2000)
35. Zachariadis, S. and Mascolo, C. and Emmerich, W. Adaptable mobile applications: Exploiting logical mobility in mobile computing. In: Horlait, E., Magedanz, T. and Glitho, RH, (eds.) Mobile Agents for Telecommunication Applications *Lecture Notes in Computer Science*, Vol. 2881, 170-179, Springer (2003)

Author Index

A
Abad, Ángel Flores, 885
Alatawi, Asma, 567
Alberti, Antonio Marcos, 293
Alfaraj, Sara, 567
Almanza, N., 903
Alomari, Amirah, 243
Amato, Alba, 325
Amato, Flora, 305, 315, 415, 679, 689
Amelino, Domenico, 747
Arion, Aurel-Florin, 173
Arora, Deepali, 853
Aubin, Mateus, 293
Awny, Mohameed Mamdouh, 813
Azuara, A., 903

B
Balzano, Walter, 737
Baneres, David, 935
Barbareschi, Mario, 747
Barolli, Leonard, 961
Benedusi, Paolo, 699
Bila, Jiri, 885
Blocho, Miroslaw, 145
Bosmans, Stig, 789

C
Caballé, Santi, 947, 961
Caggianese, Giuseppe, 105
Cai, Jin-yan, 69
Cai, Li-Jun, 435
Cervantes, S., 903
Chen, Bo-Wei, 425
Chen, Yuehui, 185
Chianese, Angelo, 699
Chilamkurti, Naveen, 425
Chun, Shan, 459
Ciampi, Mario, 35, 349
Ciccia, Simone, 221
Cilardo, Alessandro, 3
Clarisó, Robert, 947
Colace, Francesco, 305
Conesa, Jordi, 947
Corazza, Anna, 717
Coronato, Antonio, 325
Cortés, Julio Zenón García, 913
Cozzolino, Giovanni, 415, 679, 689
Cuomo, Salvatore, 707

D
d'Acierno, Antonio, 305
da Costa, Cristiano André, 293
da Rosa Righi, Rodrigo, 293
Danyang, LI, 525
De Bock, Yorick, 759
de Jesús Nandayapa Alfaro, Manuel, 885
de la Luz Carrillo González, Ma., 913
De Meulenaere, Paul, 779
De Michele, Pasquale, 707
De Pietro, Giuseppe, 163, 337
De Rosa, Francesco, 361
Del Sorbo, Maria Rosaria, 737
Deng, Shiqi, 259
Deng, Zi-Yun, 435
Di, Hongxi, 57
Ding, Jie, 373
Ding, Ye, 447
Donnarumma, Ciro, 3

E
Esposito, Angelo, 35, 349
Esposito, Massimo, 163

F
Fujihashi, Takuya, 599
Furukawa, Keita, 579

© Springer International Publishing AG 2017

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7

G
Gagliardi, Mirko, 3
Gallo, Luigi, 105
Gañán, David, 947
Gang, Dong, 517
García, Francisco Javier Carrillo, 913
Ge, Aijun, 115, 123
Giordanengo, Giorgio, 221
González-Deleito, Nicolás, 769
Gotoh, Yusuke, 589
Graziani, Alessio, 337
Guo, Yuanbo, 467

H
Hai, Tao, 825
He, Ting-Qin, 435
Hellinckx, Peter, 759, 779
Herrera, José Elías Cancino, 885
Hirota, Yusuke, 599
Hristoskova, Anna, 769
Hu, Yan-Jing, 639
Hu, Zhipeng, 197
Husheng, Wu, 555
Hussain, Farookh Khadeer, 233
Hussain, Omar Khadeer, 233
Hussain, Walayat, 233
Huybrechts, Thomas, 759

I
Ibrahim, Asmaa Abdelrehim Selim, 813
Isgrò, Francesco, 717
Ishi, Yoshimasa, 619
Ishida, Tetta, 599
Ishida, Tomoyuki, 667
Iyobe, Misaki, 667

J
Jiang, Kun, 93
Jiang, Wanyuan, 391
Jie, Zhang, 877
Jiménez, M.A., 903
Jinyan, Cai, 525
Jorge, Ricardo Rodríguez, 885
Ju, Ankang, 467

K
Kawakami, Tomoya, 619
Kito, Takahito, 599
Kurdi, Heba, 243, 567

L
Latif, Seemab, 895
Lei, Min, 259
Li, Chao, 835
Li, Dan-yang, 69
Li, Haoxuan, 759, 779
Li, Hongxia, 57
Li, Jing, 209, 925
Li, Kin Fun, 853, 863
Li, Li, 269
LI, Peng, 373, 383
Li, Peng, 391
Li, Peng, 447
Li, Wei, 373
Li, Zhe, 25, 497
Liang, Yu-ying, 69
Lili, , 801
Liu, Dong-sheng, 507
Liu, Qing, 49
Liu, Wei, 391
Long, Chen Xiao, 459
Long, Man-sheng, 49
Longobardo, Luca, 717
Lu, Jintian, 209, 925
Luo, Wen-lang, 49, 507
Lv, Jing-xiang, 507

M
Ma, Chuangui, 115, 123
Ma, Jialin, 157, 657
Ma, Jun, 467
Ma, Kun, 185
Maiorano, Francesco, 707
Marcellino, Livia, 707
Maricaux, Glenn, 789
Marrufo, Ángel Israel Soto, 885
Marulli, Fiammetta, 727
Matsumoto, Satoru, 619
Matsuo, Keita, 961
Mazzeo, Antonino, 679, 747
Mazzocca, Nicola, 315, 415
Medina, Antonia Mireles, 913
Mei, Li, 481
Meng, Bo, 209, 925
Meng, Tao, 435
Mexicano, A., 903
Minutolo, Aniello, 163
Moriwaki, Yurika, 869
Morosan, Adrian-Gabriel, 173
Moscato, Francesco, 315, 689
Moscato, Vincenzo, 305, 689
Mossucca, Lorenzo, 221
Mu, Shengdong, 269
Mungiello, Innocenzo, 361
Murano, Aniello, 737

N
Nalepa, Jakub, 145

Nardaggio, Alessandro, 727
Nunome, Toshiro, 579

O
Ogiela, Lidia, 253
Ogiela, Marek R., 253
Olguín, José Angel Montes, 913
Otomo, Iori, 599
Oyanagi, Tatsuya, 667

P
Palou, Daniel, 961
Pan, Gang, 69
Paragliola, Giovanni, 325
Peijun, Wang, 517
Penta, Antonio, 305
Pérez, J., 903
Picariello, Antonio, 305, 689
Piccialli, Francesco, 699
Pilosu, Luca, 221
Ponce, Israel U., 885
Pop, Florin, 173
Pratola, Roberto, 337
Prevete, R., 717

Q
Qi, Wei, 533, 543
QIAN, Cong, 383
Qing, Xue, 631

R
Racioppi, Adriano, 727
Ratan, Vivek, 863
Raziq, Nosheela, 895
REN, Baoxiang, 25, 497
Ren, Wei, 259
Renga, Flavio, 221
Renjun, Zhan, 555
Rho, Seungmin, 425
Rodríguez, R., 903
Romano, Sara, 679
Ruiu, Pietro, 221

S
Sánchez, Vianey Guadalupe Cruz, 885
Sannino, Giovanna, 337
Scionti, Alberto, 221
Shang, Chao-xuan, 69
SHEN, Weiwei, 383
Sheng-dong, Mu, 801
SHI, Kai, 497
Shibata, Yoshitaka, 667
Shihao, WU, 525
Sicuranza, Mario, 35, 349
Sodre, Arismar Cerqueira, 293
Song, Xingshen, 93
Stranieri, Silvia, 737
Sugita, Kaoru, 651, 667

T
Takagi, Yuki, 589
Takono, Kosuke, 869
TAN, Ruilian, 15, 25, 487, 497
Teranishi, Yuuichi, 619
Terzo, Olivier, 221
Tian, Yixiang, 269
Tsiporkova, Elena, 769

U
Uchida, Noriki, 667

V
Vallifuoco, Luca, 727
Van der Schueren, Filip, 789
Vecchi, Giuseppe, 221
Villegas, Osslan Osiris Vergara, 885

W
Wan, Zhongyu, 209, 925
Wang, Baocang, 405
Wang, Haixing, 825, 835
Wang, Jujian, 835
WANG, Ruchuan, 447
Wang, Wei, 281
Wang, Xu An, 57, 81, 197, 269, 281, 435, 801
Wang, Xu'an, 123
WANG, Xu-an, 49
Wang, Zhijian, 157
Watanabe, Takashi, 599
Watanabe, Yousuke, 609
Wei, Fushan, 115
Wu, GuangMing, 281

X
Xhafa, Fatos, 281, 961
Xiao, Guiping, 825
Xiao, Siyu, 115, 123
Xiaopan, LIU, 525
Xu, He, 373, 383, 391, 447

Y
Yamamoto, Noriyasu, 845
Yanbin, Ma, 555
Yang, Su, 133

Yang, Yuexiang, 93
Yi-xiang, Tian, 801
Yixiao, Sun, 555
Yoshihisa, Tomoki, 619
Yu, Ziqiang, 185

Z
Zeng, Lingling, 259
ZENG, Xiao-hui, 49, 507
Zhang, Jie, 123
Zhang, Jindan, 405
Zhang, Jing, 197
Zhang, Jinli, 209, 925
Zhang, Lin, 657
Zhang, Liqiong, 533
Zhang, Yongjun, 657
Zhao, Fadong, 543
Zhen, Nie, 517
Zhexin, Han, 555
Zhu, Taiming, 467
Zhuang, Hongwei, 533, 543
Zubaliy, Andriy, 769