Nabendu Chaki
Agostino Cortesi (Eds.)

Communications in Computer and Information Science 245

Computer Information Systems – Analysis and Technologies

10th International Conference, CISIM 2011
Kolkata, India, December 2011
Proceedings

Springer

Communications
in Computer and Information Science 245

Nabendu Chaki Agostino Cortesi (Eds.)

Computer Information Systems - Analysis and Technologies

10th International Conference, CISIM 2011
Kolkata, India, December 14-16, 2011
Proceedings

Springer

Volume Editors

Nabendu Chaki
University of Calcutta
Department of Computer Science and Engineering
92 APC Road, Kolkata 700009, India
E-mail: nabendu@ieee.org

Agostino Cortesi
Università Ca' Foscari di Venezia
Dipartimento di Informatica
via Torino 155, 30172 Mestre, Italy
E-mail: cortesi@unive.it

ISSN 1865-0929 e-ISSN 1865-0937
ISBN 978-3-642-27244-8 e-ISBN 978-3-642-27245-5
DOI 10.1007/978-3-642-27245-5
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2011943116

CR Subject Classification (1998): H.4, C.2, H.3, I.2, D.2, C.2.4

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

This volume contains the revised version of the papers presented at the 10th International Conference on Computer Information Systems and Industrial Management Applications (CISIM 2011), held in Kolkata, India, during December 14–16, 2011.

CISIM centers on state-of-the-art research relevant to computer information systems and their applications, with a focus on agent-based computing. The goal of this annual meeting is to bring together researchers and practitioners from academia and industry to focus on advanced concepts in secure and agent-based computing and industrial management.

Topics covered by CISIM include (1) agent-based computing: agents and service science, agent-based system development, WWW and Semantic Web agents, agent-based simulations, agent communications; (2) security and trust: authentication and access control, intrusion detection, prevention and response, key management and cryptographic protocols, network security, Web security; (3) biometric security and multimedia analysis: multimedia content representation and reasoning, classification and indexing of multimedia information objects, multimedia security and protection, multimedia in mobile and ambient systems, security and reliability assessment for biometric systems, biometric standards and interoperability, derivation of cryptographic keys from biometrics, biometric performance management.

The Program Committee selected 30 papers out of 67 submissions based on anonymous reviews and discussions in an electronic Program Committee meeting. The principal selection criteria were relevance and quality. Every paper was reviewed by two to four reviewers, and the articles published in this volume were improved based on reviewers' comments. The papers present not only the research ideas but also substantial evaluations that made their arguments stronger.

CISIM has a tradition of inviting distinguished speakers to give talks. This time the program included six keynote addresses and plenary lectures by:

- Young Im Cho, University of Suwon, South Korea
- Dipankar Dasgupta, University of Memphis, USA
- Aditya K. Ghose, University of Wollongong, Australia
- Sushil Jajodia, George Mason University Fairfax, USA
- Dong Hwa Kim, Hanbat National University, Korea
- Heinrich Voss, Hamburg University of Technology, Germany

We would like to thank these renowned professors from Europe, North America, Australia and Asia for delivering these invited talks.

We take this opportunity to thank all the members of the International Program Committee and the external reviewers for their dedicated effort in the paper selection process that had been so crucial in ensuring the quality of the conference.

We thank the University of Calcutta and the Honorable Vice Chancellor, University of Calcutta, as well as the Chief Patron of CISIM 2011, Suranjan Das, for hosting the 10th International Conference on Computer Information Systems and Industrial Management Applications (CISIM 2011) in Kolkata.

We would also like to thank Ryszard Tadeusiewicz and Subhansu Bandyopadhyay, the Honorary Chairs for CISIM 2011, and the Conference General Chairs, Khalid Saeed, Ajith Abraham, and Vaclav Snasel. We thank the Steering Committee Chair Aditya Bagchi, and other members of the committee, and last but not least the two Co-chairs of Organizing Committee, Sankhayan Choudhury and Rituparna Chaki, along with all the members of the Organizing Committee.

Finally from the desk of Program Chairs for CISIM 2011, we thank all the authors who contributed to the success of the conference. We also sincerely hope that all attendees are benefited academically from the conference and wish them every success in their research.

December 2011

Nabendu Chaki
Ryszard S. Choraś
Agostino Cortesi
Slawomir T. Wierzchoń

Organization

The 10th International Conference on Computer Information Systems and Industrial Management Applications was organized by the University of Calcutta.

Co-operating Institutions

Muroran Institute of Technology, Japan
University of Silesia, Poland
Nagaoka University of Technology, Japan
VSB-Technical University of Ostrava, Czech Republic
Silesian University of Technology, Poland
IPI PAN Polish Academy of Sciences, Warsaw, Poland
Machine Intelligence Research-MIR Labs, USA
University of Technology and Life Sciences, Poland
Suwon University, Korea
Hanbat National University, South Korea
Bialystok University of Technology, Poland
University of Finance and Management in Bialystok, Poland

Honorary Chairs

Ryszard Tadeusiewicz, Poland
Subhansu Bandyopadhyay, India

Conference General Chairs

Khalid Saeed, Poland
Ajith Abraham, USA
Vaclav Snasel, Czech Republic

Steering Committee

Aditya Bagchi, India, Chair
Rituparna Chaki, India
Samiran Chattopadhyay, India
Tetsuo Hattori, Japan
Andrzej Mitas, Germany
Pavel Moravec, Czech Republic
Romuald Mosdorf, Poland
Gerald Schaefer, UK
Debadutta Sinha, India

Publicity Chairs

Amlan Chakrabarti, India
Piotr Porwik, Poland
Katarzyna Wegrzyn, France
Sanjit K. Setua, India

Special Events Chair

Samiran Chattopadhyay, India

Local Organizing Committee

Sankhayan Choudhury, University of Calcutta, Co-chair
Rituparna Chaki, India, Co-chair
Nabendu Chaki, University of Calcutta
Amlan Chakrabarti, University of Calcutta
Moitrayee Dasgupta, GGSIP University
Soumava Das, Ixiacom Ltd.
Shilbhadra Dasgupta, Tata Consultancy Services
Kashinath Dey, University of Calcutta
Sourav Ghosh, University of Calcutta
Rakesh Kumar Mishra, Feroze Gandhi Institute of Engineering and Technology
Sudhindu Bikash Mandal, University of Calcutta
Suparno Moitra, NASSCOM
Soumya Sen, University of Calcutta
Sanjit K. Setua, University of Calcutta
Soharab Hossain Shaikh, University of Calcutta

Program Committee Co-chairs

Nabendu Chaki, India
Ryszard S. Choraś, Poland
Agostino Cortesi, Italy
Slawomir Wierzchoń, Poland

Program Committee

Aditya Bagchi, India
Ajith Abraham, USA
Ajith P. Madurapperuma, Sri Lanka
Akira Imada, Belarus
Akira Shionoya, Japan
Ambjorn Naeve, Sweden
Amlan Chakrabarti, India
Andrzej Mitas, Poland
Anirban Sarkar, India
Anna Bartkowiak, Poland
Athula Sumathipala, UK
Bogdan Smolka, Poland
De Silva C. Liyanage, Brunei
Dong Hwa Kim, Korea

Dusan Husek, Czech Republic
Ewa Pietka, Poland
Franciszek Seredyski, Poland
Gerald Schaefer, UK
Gisella Facchinetti, Italy
Halina Kwaśnicka, Poland
Handri Santoso, Japan
Huneong Sun Hwang, Korea
Hyun Chang Lee, Korea
Ihor Katernyak, Ukraine
Jay Rajasekera, Japan
Jerzy Rutkowski, Poland
Ji Hyung Lim, Korea
Jin Zhou, China
Juliusz L. Kulikowski, Poland
Katsuko T. Nakahira, Japan
Kenichi Itakura, Japan
Kevin Jia Jiancheng, USA
Khalid Saeed, Poland
Ki Sang Song, Korea
Koichi Nakayama, Japan
Kurosh Madani, France
Leon Bobrowski, Poland
Leonid Kompanets, Ukraine
Li Jinping, China
Lin Zhang, China
M. Sami Ashhab, Jordan
Makoto Fukumoto, Japan
Marek Kurzynski, Poland
Marian Srebrny, Poland
Marina Gavrilova, Canada
Marwan Al-Akaidi, UK
Masayosi Kamijo, Japan
Minetada Osano, Japan
Miroslaw Owoc, Poland
Muneyuki Unehara, Japan
Nabanita Das, India
Nadia Nedjah, Brazil
Nam-Gyu Kang, Japan
Nguyen Van Hop, Thailand
Nilse Enlund, Sweden
Nobuyuki Nishiuchi, Japan
Petri Helo, Finland
Pietro Ferrara, Switzerland
Piotr Augustyniak, Poland
Piotr Porwik, Poland
Raid Al-Taher, Trinidad and Tobago
Rajat De, India
Rajib Das, India
Rauf Kh. Sadykhov, Belarus
Renzo Orsini, Italy
Rikhsi Isaev, Uzbekistan
Rituparna Chaki, India
Roger Allan French, USA
Romuald Mosdorf, Poland
Ryszard Tadeusiewicz, Poland
Shan-Shan Ju, Taiwan
Soon Kon Kim, Korea
Sugata Sanyal, India
Sukriti Bhattacharya, Portugal
Swietlana Januszkiewicz, Canada
Tetsuo Hattori, Japan
Toru Yamaguchi, Japan
Uma Bhattacharya, India
Urszula Kaczmarska, Poland
Vaclav Snasel, Czech Republic
Waleed Abdulla, New Zealand
Witold Malina, Poland
Wladyslaw Skarbek, Poland
Yoshifumi Okada, Japan
Young Im Cho, Korea
Young Jae Lee, Korea
Yuehui Chen, China

Sponsors

Department of Science & Technology, Government of India, New Delhi
Department of Information Technology, Government of India, New Delhi
University of Calcutta, Kolkata
SkyTECH Solutions Pvt. Ltd.
Ixia

Table of Contents

Keynote and Plenary Lectures

Networking and Its Applications

Agent-Based Systems

Biometric Applications

Pattern Recognition and Image Processing

Industrial Applications

Algorithmic Applications and Data Management

Information and Network Security

The Future of Intelligent Ubiquitous Computing Systems

Young Im Cho

Department of Computer Science,
The University of Suwon, Korea
ycho@suwon.ac.kr

Abstract. Many ubiquitous systems and technologies have been developed as of now. As known the goal of ubiquitous computing is to achieve the well-being life. There are four different aspects for achieving ubiquitous computing, namely, they are within the computing paradigm, technical principle, application domain and application space. Nowadays, ubiquitous city (U-city) is the most applicable domain in the world. Therefore, the talk will mainly introduce and discuss an overview of the U-City idea and the known ubiquitous computing systems as well as new trends in this field.

Keywords: Ubiquitous computing, intelligent system, smart computing, ubiquitous city.

1 Introduction

Ubiquitous computing was started in 1984 by Mark Weiser in USA [1]. Ubiquitous computing is roughly the opposite of virtual reality. Usually, virtual reality puts people inside a computer-generated world, but ubiquitous computing forces the computer to live out here in the world with people. Virtual reality is primarily a horse power problem; ubiquitous computing is a very difficult integration of human factors, computer science, engineering, and social sciences. Ubiquitous computing is a post-desktop model of human-computer interaction in which information processing has been thoroughly integrated into everyday objects and activities.

The core technology of ubiquitous computing is an autonomic collaboration model for ubiquitous fusion services in ubiquitous computing environment.To do some action in ubiquitous computing, many elements coordinate to complete such action. Ubiquitous networking in ubiquitous computing is completed by the collaboration of many types of networks, city technologies, city sociology and governance.

Therefore, in my talk I will present and the main technologies and applied services in ubiquitous computing, particularly within the developing aspect of ubiquitous cities. I will also give more details and explain in examples more about the new trends of ubiquitous computing at the present time and what is expected to take place in this area of research in the near future.

2 Ubiquitous Computing Paradigm

Actually, the goal of ubiquitous computing is to make the safety, security and peace of mind society. The key technology to be a ubiquitous society is ubiquitous sensor network, in short USN. There are many application areas that use USN, such as home

N. Chaki and A. Cortesi (Eds.): CISIM 2011, CCIS 245, pp. 1–3, 2011.

control, health medical care, welfare, hospital, school or education, traffic, culture, sports, environments, tracking of produce, anti-disaster, city security, etc.

Human will, in an ambient intelligent environment, be surrounded by intelligent interfaces supported by computing and networking technology that is embedded in everyday objects such as furniture, clothes, vehicles, roads and smart materials-even particles of decorative substances like paint. Ambient intelligent is the vision of the world in which different classes of smart players are proactively assisted by the environment [2].The environment is able to aware of the smart player's current situation and his interact within his environment and of its own current state. And then the interpretation of those occurrences into smart player's goals and accordingly into possible reactions is activated. The translation of the interpreted goal into strategies to adapt itself to the smart player's needs is performed in ubiquitous computing. Ubiquitous Space [2] is the space that objects collaborate seamlessly and automatically to achieve user's or community's goal for supporting u-Society and u-Life.

To be in intelligent ubiquitous system, there are three technical principles of ubiquitous space: infrastructure with situation-awareness, system with autonomic management, service with self-growing intelligence. Already, the mega trend goes with the ubiquitous scenario according to the suggestions given by Mark Weiser 20 years ago [1].

There are 4 types of viewpoints in ubiquitous computing. In the aspect of computing paradigm, the technology is changed from static community to automatic community. In the aspect of technical principle, the technology is changed from situation aware to self growing. However, in the aspect of application domain, the technology is changed from healthcare to environment preservation. The fourth aspect concerns the application space, the technology is changed from home or building to society.

3 Ubiquitous City and Future Research Themes

Among many application areas, ubiquitous city, in short U-City, is the most applicable area. U-City is the constructed city by ubiquitous technologies and paradigm. To be the U-City, there are many ubiquitous services and hybrid technologies using RFID, USN, IPv6, sensing devices, monitoring, auto-control, real-time management etc. Usually, U-City is the same as smart city because smart has intelligent concept.

There are some examples of U-City like the media hub city and cool town in Singapore, cyber port in Hong Kong, Finland as well as Korea [3].

The U-City application service models are still being developed. I would suggest some matrix style service model according to the classification by personal life as well as the classification by land and facilities. The details will be given in my talk at the conference,

In the future we still need to develop what has already been done so far. The first theme is the research about the infra structure of U-City such as platform, security,

pattern, service scenario etc. The second theme is the research about the paradigm of `U-city such as role play between government and local government to perform U-City etc. The third theme is the research about the consulting of U-City such as the best service model according to many types of organs, and business model, and so on.

As a result, ubiquitous society is a combination of safety and security for the sake of peaceful society. Now, many services of IT mobile devices are provided with personalized manners. For completion of real ubiquitous space through U-City, the standard model and some platforms should be a prerequisite condition for success.

References

1. Cho, Y.I.: U-Society and U-Convergence Paradigm. 2008 Asia Pacific Woman Leaders Forum for Science & Technology, 183–204 (September 3, 2008)
2. Roman, M., et al.: A Middleware Infrastructure for Active Spaces. IEEE Pervasive Computing (2002)
3. Cho, Y.I.: Practical Approaches for Ubiquitous Convergence System. In: Joint 4th International Conference on Soft Computing and Intelligent Systems and 9th International Symposium on Advanced Intelligent Systems, September 17-21 (2008)

Who Is Responsible for Security and Privacy in the Cloud?

Dipankar Dasgupta

Director, Center for Information Assurance
Professor, Department of Computer Science
University of Memphis, Memphis, TN 38152
dasgupta@memphis.edu

Abstract. Cloud computing is becoming increasingly attractive both in commercial and government sectors because of the significant cost reduction in their IT operations. The technical benefits of cloud environment (and its service-oriented capabilities) are the availability of computing resources those can quickly be engaged for service execution and released when no longer needed. As the cloud services moving to the mainstream computing, the issues of ownership and the chain of custody of customer data are becoming very critical. In this talk, I will discuss various cloud security issues, and the role and responsibilities of cloud service providers since the secure cloud environment is essential for providing uninterrupted services to customers. This talk will introduce various metrics for cyber security coverage, and demonstrate a tool, called MEGHNAD for estimating security coverage for cloud services. This tool can serve as a specialized Cloud Doctor in prescribing the right combination of security tools for different cloud services and according to the level of security assurance required. It will useful to a cloud provider who wants to incorporate security insurance as part of their Service Level Agreements (SLA).

Keywords: Cloud computing, Security, cloud doctor, services, SLA.

References

1. Bhattarai, A., Dasgupta, D.: A Self-Supervised Approach to Comment Spam Detection Based on Content Analysis. International Journal of Information Security and Privacy (IJISP) 5(1), 14–32 (2011)
2. Dasgupta, D., Rahman, M.: A Framework for Estimating Security Coverage for Cloud Service Insurance. In: ACM ICPS Proceedings of Cyber Security and Information Intelligence Research Workshop (CSIIRW-7), Oak Ridge, USA, October 12-14 (2011)
3. Ferebee, D., Dasgupta, D., Wu, Q., Schmidt, M.: Security Visualization: Cyber Security Storm Map and Event Correlation. In: The Proceedings of IEEE Symposium on Computational Intelligence in Cyber Security (CICS) Under IEEE Symposium Series in Computational Intelligence, Paris, France (April 2011)
4. Yu, S., Dasgupta, D.: An Effective Network-based Intrusion Detection Using Conserved Self Pattern Recognition Algorithm Augmented with Near-deterministic Detector Generation. In: The Proceedings of IEEE Symposium on Computational Intelligence in Cyber Security (CICS) Organized in IEEE Symposium Series in Computational Intelligence, Paris, France (April 2011)
5. Carvalho, M., Dasgupta, D., Grimaila, M.: Mission Resilience in Cloud Computing: A Biologically Inspired Approach. In: 6th International Conference on Information Warfare and Security, Washington, DC, USA (March 2011)

N. Chaki and A. Cortesi (Eds.): CISIM 2011, CCIS 245, p. 4, 2011.

The Optimizing Web: Leveraging Agent Technology for Sustainability

Aditya Ghose

Decision Systems Laboratory
School of Computer Science and Software Engineering
University of Wollongong, NSW 2522 Australia
aditya@uow.edu.au

Abstract. This paper describes how agent technology might be leveraged to deliver a critical solution to the carbon mitigation challenge - in the context of the Optimizing Web project.

1 Introduction

There is widespread recognition of the climate change crisis, and the need to develop scientific, technological and managerial responses. Current thinking on climate change responses emphasizes the development of alternative energy sources, the development of smart automotive technology and the introduction of macroeconomic levers (e.g., carbon taxes, emission trading schemes) to alter energy consumption behaviour at the level of both enterprises and individuals. While these are laudable objectives, these initiatives might be somewhat premature. We do not yet have the ability to ensure *efficient utilization* of existing infrastructure - it stands to reason that this must be a necessary first step before we make massive investments in novel technological bases for energy generation or transportation infrastructure.

The notion of *efficient resource utilization* is inextricably tied to the notion of *optimization* - in particular, the ability to optimize energy use - yet this has been largely ignored in the current discourse. The connection between optimization and carbon mitigation is compelling: optimization enables efficient resource utilization, thus lowering energy consumption and the carbon footprint. The global industrial/technological infrastructure, including transportation systems, manufacturing plants, human habitat and so on, is typically operated in an ad-hoc and significantly sub-optimal fashion. This remains the case despite the availability of sophisticated optimization technology for almost the past seven decades (present day operations research techniques trace their roots to the pioneering work of George Dantzig in the early 1940s that resulted in the original optimization algorithm - linear programming).

Most applications of optimization technology have been in piecemeal, monolithic optimization systems. Yet the climate change crisis requires optimization on a large-scale, and in a manner that permits entities in a massive planetary supply chain (it can be instructive to view the global network of human activity

N. Chaki and A. Cortesi (Eds.): CISIM 2011, CCIS 245, pp. 5–8, 2011.

as such) to collaborate to achieve the commonly agreed upon carbon mitigation objective. Traditional stand-alone "batch" optimization technology cannot be deployed in this setting for a variety of reasons. It is impractical to conceive of a centralized "global optimizer"'. It is equally impractical to expect business entities to reveal what if often highly sensitive information about their supply chains to central optimizer. Finally, the scale of the problem is too large to be feasibly addressed. The problem, then, is to support decentralized, distributed, collaborative optimization on a global scale. The nearest point of departure for such an enterprise is the literature on agent-based, distributed optimization.

The climate change crisis has presented the community of researchers interested in agent technology with a historic opportunity. For the first time ever, we have a globally agreed-upon objective function: the carbon footprint minimization objective. This opens up the possibility for devising large-scale, agent-based, *collaborative optimization architectures*, where large numbers of agent-based optimizers solve local optiimization problems, while collaborating to improve the cumulative system performance relative to a shared objective function. The Optimizing Web project (see www.optimizing-web.org) seeks to design and validate the conceptual underpinnings of an infrastructure that would support very large scale collaborative optimization across a potentially global collection of optimizers. The Optimizing Web project grew out of the University of Wollongong Carbon-Centric Computing Initiative (see www.ccci.uow.edu.au) which has the broader agenda of exploring ways in which the full spectrum of computing technologies might contribute to solutions to the climate change crisis. The Optimizing Web vision is to provide ubiquitous collaborative optimization services, at the level of individual devices, vehicles within transportation systems, units within organizations or manufacturing plants - as well aggregations of all of these. The Optimizing Web would be a systems of systems, and would provide a protocol (or a set of protocols) for local optimizers to inter-operate to optimize the global carbon footprint minimization objective, while making appropriate trade-offs in relation to their local objectives. While the modelling and solution of "local" optimization has been the focus of attention for the operations research (OR) community for several decades, this project addresses the question of how large collections of optimization problems (with associated solvers), with possibly intersecting signatures (sets of common variables), might be made to inter-operate to optimize a shared function (the carbon footprint minimization objective).

There are three specific challenges for the agents community: the design of agent-based *optimization architectures*, the development of the next generation of agent-based distributed optimization protocols and the integration of optimization with distributed agent-based planning. We address the first of these in some detail below.

2 Agent-Based Optimization Architectures

Fundamental to the optimizing web is the notion of an optimization architecture, i.e., a collection of appropriately configured inter-operating optimizers. It

specifies the constituent optimizers, their signatures (the decision variables whose values are determined by the optimizer in question), their parameters (the variables whose values constrain the optimizer), and the nature of the inter-agent coordination messages exchanged. The architecture is agnostic to the internals of individual optimizers. We might design an optimization architecture from scratch, or we might engineer one around pre-existing, legacy optimizers. Both approaches present challenges. Key to understanding optimization architectures is an appreciation of the competing pulls of *local* objectives and system-wide (societal or global) objectives, and the implications of resolving them in different ways. Consider an agent-based traffic planning setting (Srivastav, 2011), where individual road users get routing advice from decision-support agents executing on hand-held devices (smartphones, PDAs etc.). Our empirical studies confirm the intuition that locally sub-optimal decisions at the level of individual road users can contribute to improving the system-wide objective (of reducing the cumulative carbon footprint, for example). Sometimes, an individual road-user on a short-hop journey will need to be incentivized to take a longer route, in the interests of reducing the cumulative carbon footprint of road users on longer journeys who would have been obliged to take significantly longer routes to avoid the congestion that the short-hop users might have contributed to. Similarly, our empirical work on designing optimal resource allocation mechanisms in clinical settings [1] suggests that making patients incur a small wait-time (within clinically acceptable limits) achieves far better system-wide efficiencies than a "first-come-first-served"' logic. Foremost amongst these is the notion of objective alignment (or consistency). An objective function determines the optimization behaviour of an agent, i.e., the choices it makes amongst possible alternative solutions. Objective alignment helps ensure that optimizers use objectives that are aligned with the global carbon footprint minimization objective. Given a set of objective functions, we need to be able to determine if these jointly determine a consistent set of optimization behaviours. Consider the pair of objective functions minimize x and minimize -x. If the set of feasible solutions (i.e., solutions that satisfy all of the applicable constraints) is non-singleton, then an agent will not be able to satisfy both objectives (since they, in effect, "pull in opposite directions"). If there is exactly one feasible solution, however, the pair of objectives is in fact aligned. Similarly, the objectives minimize x and minimize x2 are not aligned in general, but may be aligned if x is restricted to be positive. Definitions of objective alignment did not exist in the literature, until our preliminary work in [2], where we view an objective function as a preference relation defined over the set of feasible solutions. Alignment then reduces to the absence of contradictory preferences. While this approach provides the conceptual framework for understanding objective alignment, it does not immediately lead to practical tools for checking alignment. A major challenge is the fact that alignment cannot be determined on the basis of the objectives alone, but is also contingent on the set of applicable constraints, and hence the set of feasible solutions (as the examples above illustrate). Additionally, exhaustively enumerating the preference relation on the set of feasible solutions is impractical, yet there are no easy

approaches to performing alignment checking analytically. A compromise is to perform approximate checking using heuristic techniques, with no guarantee of completeness. The methodological bases for designing optimization architectures need to be defined. Ensuring that the objectives within an optimization architecture are aligned with the global carbon footprint minimization objective by design also requires the ability to decompose objectives (for instance, how do we start with a set of high-level organizational objectives and decompose these into the objectives of the constituent business units, while maintaining consistency with a global objective?). Finally, we need to understand how to measure (or monetize) the trade-offs between the local objectives of an optimizer and the global (carbon mitigation) objective. In other words, we need to devise mechanisms to incentivize an agent to adopt behaviour that is potentially sub-optimal relative to its own objectives, in the interests of the global objective.

There are several other interesting challenges. We need to be able devise means for agents to *discover footprints*, i.e., answer the question: which agents does a given agents share constrainst with? Sometimes the answer to this question is relatively static, but in other settings (such as traffic planning) the answer can be highly dynamic, and some modicum of predictive reasoning is required. The maintenance of optimization architectures in highly dynamic settings is another major challenge. The design of social mechanisms (such as carbon credits) to incentivize agents to adopt locally sub-optimal behaviour poses challenges. Existing agent-based optimization protocols need to extended, as do agent communication standards to enable the kinds of messaging/negotiation necessary in this context. Finally, we need to be able to *disocver constraints and objectives* by mining the "big data"' repositories that our current technologies have made possible.

References

[1] Billiau, G., Ghose, A.: Agent-based optimization in healthcare. Decision Systems Lab., Univ. of Wollongong Technical Report 2011-TR-02 (2011)
[2] Dasgupta, A., Ghose, A.K.: Implementing reactive BDI agents with user-given constraints and objectives. Int'l Journal of Agent-Oriented Software Engineering (IJAOSE) 4(2), 141–154 (2010)
[3] Srivastav, B., Billiau, G., Lim, M., Lee, T., Ghose, A.: Optimal traffic planning. Decision Systems Lab., Univ. of Wollongong Technical Report 2011-TR-01 (2011)

Scalable Detection of Cyber Attacks*

Massimiliano Albanese[1], Sushil Jajodia[1], Andrea Pugliese[2], and V.S. Subrahmanian[3]

[1] George Mason University, Fairfax, VA 22030, USA
{malbanes,jajodia}@gmu.edu
[2] University of Calabria, 87036 Rende (CS), Italy
apugliese@deis.unical.it
[3] University of Maryland, College Park, MD 20742, USA
vs@umiacs.umd.edu

Abstract. Attackers can exploit vulnerabilities to incrementally penetrate a network and compromise critical systems. The enormous amount of raw security data available to analysts and the complex interdependencies among vulnerabilities make manual analysis extremely labor-intensive and error-prone. To address this important problem, we build on previous work on topological vulnerability analysis, and propose an automated framework to manage very large attack graphs and monitor high volumes of incoming alerts for the occurrence of known attack patterns in real-time. Specifically, we propose (i) a data structure that merges multiple attack graphs and enables concurrent monitoring of multiple types of attacks; (ii) an index structure that can effectively index millions of time-stamped alerts; (iii) a real-time algorithm that can process a continuous stream of alerts, update the index, and detect attack occurrences. We show that the proposed solution significantly improves the state of the art in cyber attack detection, enabling real-time attack detection.

Keywords: Attack graphs, attack detection, scalability.

1 Introduction

An ever increasing number of critical applications and services rely today on Information Technology infrastructures, exposing companies and organizations to an elevated risk of becoming the target of cyber attacks. Attackers can exploit network configurations and vulnerabilities to incrementally penetrate a network and compromise critical systems. Most of the elementary steps of an attack are intercepted by intrusion detection systems, which generate alerts accordingly. However, such systems typically miss some events and also generate a large number of false alarms. More importantly, they cannot derive attack scenarios from individual alerts.

* This material is based upon work supported by the Army Research Office under MURI grant W911NF-09-1-0525 and DURIP grant W911NF-11-1-0340.

N. Chaki and A. Cortesi (Eds.): CISIM 2011, CCIS 245, pp. 9–18, 2011.

The enormous amount of raw security data available to analysts, and the complex interdependencies among vulnerabilities make manual analysis extremely labor-intensive and error-prone. Although significant progress has been made in several areas of computer and network security, powerful tools capable of making sense of this ocean of data and providing analysts with the "big picture" of the cyber situation in a scalable and effective fashion are still to be devised. To address this important problem, we build on previous work on topological vulnerability analysis [7,14], and propose an automated framework to manage very large attack graphs and monitor high volumes of incoming alerts for the occurrence of known attack patterns in real-time.

Specifically, we propose (i) a data structure that merges multiple attack graphs and enables concurrent monitoring of multiple types of attacks; (ii) an index structure that can effectively index millions of time-stamped alerts; (iii) a real-time algorithm that can process a continuous stream of alerts and update the index. We show that the proposed solution significantly improves the state of the art in real-time attack detection. Previous efforts have in fact indicated that it is possible to process alerts fast, under certain circumstances, but have not considered the impact of very large attack graphs. Finally, we report on preliminary experiments, and show that the proposed solution scales well for large graphs and large amounts of security alerts.

The paper is organized as follows. Section 2 discusses related work. Section 3 introduces the notion of *temporal attack graph*, whereas Section 4 presents the proposed data structures and the algorithm to process and index security alerts. Finally, Section 5 reports the results of experiments, and Section 6 provides some concluding remarks.

2 Related Work

To reconstruct attack scenarios from isolated alerts, some correlation techniques employ prior knowledge about attack strategies [3] or alert dependencies [9]. Some techniques aggregate alerts with similar attributes [13] or statistical patterns [12]. Hybrid approaches combine different techniques for better results [9]. To the best of our knowledge, the limitation of the nested-loop approach, especially for correlating intensive alerts in high-speed networks, has not been addressed yet. Network vulnerability analysis enumerates potential attack sequences between fixed initial conditions and attack goals [7]. In [10], Noel *et al.* adopt a vulnerability-centric approach to alert correlation, because it can effectively filter out bogus alerts irrelevant to the network. However, the nested loop procedure is still used in [10]. Real-Time detection of isolated alerts is studied in [11]. Designed for a different purpose, the RUSSEL language is similar to our approach in that the analysis of data only requires one-pass of processing [5].

Hidden Markov Models (HMMs) and their variants have been used extensively for plan recognition. Luhr et al. [8] use Hierarchical HMMs to learn probabilities of sequences of activities. [4] introduces the Switching Hidden Semi-Markov Model (SHSMM), a two-layered extension of the Hidden Semi-Markov Model

(HSMM). The bottom layer represents atomic activities and their duration using HSMMs, whereas the top layer represents a sequence of high-level activities defined in terms of atomic activities. In [6], Hamid et al. assume that the structure of relevant activities is not fully known a priori, and provide minimally supervised methods to learn activities. They make the simplifying assumption that activities do not overlap. The problem of recognizing multiple interleaved activities has been studied in [2], where the authors propose a symbolic plan recognition approach, relying on hierarchical plan-based representation and a set of algorithms that can answer a variety of recognition queries.

To the best of our knowledge, there has been virtually no work on efficient indexing to support scalable and real-time attack detection. Our work differs from previous work by providing a mechanism to index alerts as they occur in a data structure that also unifies a set of known attack graphs. The index we propose in this paper extends the index proposed by Albanese et al. [1]. Differently from [1], the index we propose in this paper can efficiently manage both large attack graphs and large sequences of alerts, whereas the authors of [1] do not address scale issues with respect to the size of the graphs.

3 Temporal Attack Graphs

In this paper, we extend the attack graph model of [14] with temporal constraints on the unfolding of attacks. We assume that each step of an attack sequence is taken within a certain temporal window after the previous step has been taken. Without loss of generality, we assume an arbitrary but fixed time granularity. We use $\mathcal{T}$ to denote the set of all time points. The definition of *temporal attack graph* is given below.

Definition 1 (Temporal Attack Graph). *Given an attack graph* $G = (V \cup C, R_r \cup R_i)$ *a* temporal attack graph *built on* G *is a labeled directed acyclic graph* $A = (V, E, \delta, \gamma)$ *where:*

- V *is the finite set of vulnerability exploits in the attack graph;*
- $E = R_i \circ R_r$ *is the* prepare for *relationship between exploits;*
- $V^s = \{v \in V \mid \nexists v' \in V \text{ s.t. } (v', v) \in E\} \neq \emptyset$*, i.e., there exists at least one* start node *in* V*;*
- $V^e = \{v \in V \mid \nexists v' \in V \text{ s.t. } (v, v') \in E\} \neq \emptyset$*, i.e., there exists at least one* end node *in* V*;*
- $\delta : E \to \mathcal{T} \times \mathcal{T}$ *is a function that associates a pair* (t_{min}, t_{max}) *with each edge in the graph;*
- γ *is a function that associates with each vulnerability* $v_i \in V \setminus V^s$ *the condition*

$$\gamma(v_i) = \bigwedge_{c_j \in C \text{ s.t. } (c_j, v_i) \in R_r} \left(\bigvee_{v_k \in V \text{ s.t. } (v_k, c_j) \in R_i} \nu_{j,i} \right),$$

where each $\nu_{j,i}$ *denotes that* v_i *must be execute after* v_j*, within the temporal window defined by* $\delta(v_j, v_i)$*.* □

Intuitively, an edge $e = (v_j, v_i)$ in a temporal attack graph denotes the fact that exploit v_j *prepares for* exploit v_i. The pair $\delta(e) = (t_{min}, t_{max})$ labeling the edge indicates that the time elapsed between the two exploits must be in the interval $[t_{min}, t_{max})$. The condition $\gamma(v_i)$ labeling a node v_i encodes the dependencies between v_i and all the exploits preparing for it. In the following, we often abuse notation and use v_j instead of $\nu_{j,i}$ in condition $\gamma(v_i)$, when v_i is clear from the context. We say that a set V^* of exploits satisfies condition $\gamma(v_i)$ if exploits in V^* imply all the security conditions *required* by v_i.

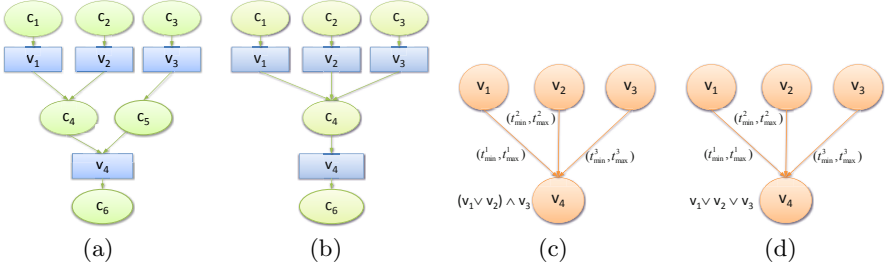


Fig. 1. Examples of attack graphs and corresponding temporal attack graphs

Example 1. Consider the two attack graphs shown in Figures 1(a) and 1(b). In the one of Figure 1(a), both conditions c_4 and c_5 are required to execute exploit v_4. Condition c_4 can be achieved by either exploit v_1 or exploit v_2, whereas condition c_5 can only be achieved by exploit v_3. Thus v_4 must be preceded by v_3 and one of v_1, v_2. In the graph of Figure 1(b), only condition c_4 is necessary for exploit v_4, and it can be achieved by any of v_1, v_2, or v_3. The corresponding temporal attack graphs are shown in Figure 1(c) and 1(d) respectively.

An attack may be executed in multiple different ways, which we refer to as *instances* of the attack. Informally, an instance of a temporal attack graph A is a tree $T = (V_T, E_T)$ over A, rooted at an end node of A. The root of the tree is an exploit implying the attack's target condition, whereas the leaves represent exploits depending on initial security conditions. Figure 2 shows the two possible instances of the temporal attack graph of Figure 1(c). Note that the execution of v_3 is always required to prepare for v_4, whereas only one of v_1, v_2 is required.

4 Attack Detection and Indexing

Intrusion alerts reported by IDS sensors typically contain several attributes. For the purpose of our analysis, we assume that each alert o is a tuple ($type$, ts, $host_{src}$, $host_{dest}$), where $type$ denotes the event type, $ts \in \mathcal{T}$ is a timestamp, and $host_{src}$, $host_{dest}$ are the source and destination host respectively. We refer to a sequence of such alerts as the *observation sequence*. Finally, we assume the

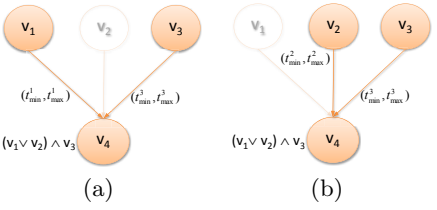


Fig. 2. Examples of temporal attack instances

existence of a function $\phi : L \rightarrow V$ that maps alerts to exploits (alerts that cannot be mapped to any known vulnerability are ignored). Given a temporal attack graph A and an observation sequence O, we are interested in identifying sequences of alerts corresponding to instances of A. We refer to such sequences as *occurrences* of A.

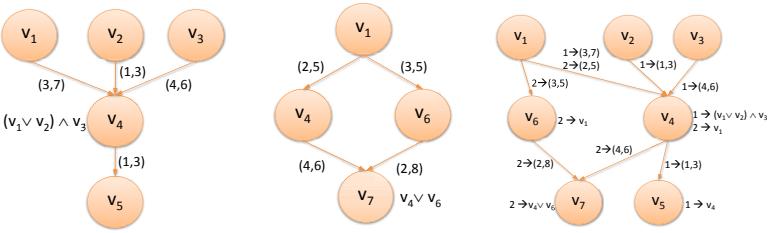
Definition 2 (Temporal Attack Occurrence). *Given an observation sequence* $O = \langle o_1, o_2, \ldots, o_n \rangle$ *and a temporal attack graph* $A = (V, E, \delta, \gamma)$*, an occurrence of* A *in* O *is a sequence* $O^* = \langle o_1^*, \ldots, o_k^* \rangle \subseteq O$ *such that:*

- $o_1^*.ts \leq o_2^*.ts \leq \ldots \leq o_k^*.ts$*;*
- $\exists$ *an instance* $T = (V_T, E_T)$ *of* A *such that* $V_T = \{v \in V \mid \exists o^* \in O^*$ *s.t.* $\phi(o^*) = v\}$*, i.e.,* O^* *includes alerts corresponding to all the exploits in* T*;*
- $(\forall e = (v', v'') \in E_T)\ \exists o_{i'}, o_{i''} \in O^*$ *s.t.* $\phi(o_{i'}) = v' \wedge \phi(o_{i''}) = v'' \wedge t_{min} \leq o_{i''}.ts - o_{i'}.ts \leq t_{max}$*, where* $(t_{min}, t_{max}) = \delta(e)$*.*

The span *of* O^* *is the time interval* $span(\langle o_1^*, \ldots, o_k^* \rangle) = [o_{i_1}.ts, o_{i_k}.ts]$. □

Note that multiple concurrent attacks generate interleaved alerts in the observation sequence. In order to concurrently monitor incoming alerts for occurrences of multiple types of attacks, we first merge all temporal attack graphs from $\mathcal{A} = \{A_1, \ldots, A_k\}$ into a single graph. We use $id(A)$ to denote a unique identifier for attack graph A and $I_{\mathcal{A}}$ to denote the set $\{id(A_1), \ldots, id(A_k)\}$. Informally, a *Temporal Multi-Attack Graph* is a graph $G = (V_G, I_{\mathcal{A}}, \delta_G, \gamma_G)$, where V_G is the set of all vulnerability exploits. A temporal multi-attack graph can be graphically represented by labeling nodes with vulnerabilities and edges with the id's of attack graphs containing them, along with the corresponding temporal windows. Note that the temporal multi-attack graph needs to be computed only once before building the index. Figure 3 shows two temporal attack graphs A_1 and A_2 and the corresponding temporal multi-attack graph.

Definition 3 (Temporal Multi-Attack Graph Index). *Let* $\mathcal{A} = \{A_1, \ldots, A_k\}$ *be a set of temporal attack graphs, where* $A_i = (V_i, E_i, \delta_i, \gamma_i)$*, and let* $G = (V_G, I_{\mathcal{A}}, \delta_G, \gamma_G)$ *be the temporal multi-attack graph built over* $\mathcal{A}$*. A* Temporal Multi-Attack Graph Index *is a 5-tuple* $I_G = (G, start_G, end_G, tables_G, completed_G)$*, where:*



(a) Attack graph A_1 (b) Attack graph A_2 (c) Multi-attack graph

Fig. 3. Example of temporal multi-attack graph

- *$start_G : V_G \rightarrow 2^{I_{\mathcal{A}}}$ is a function that associates with each node $v \in V_G$, the set of attack graph id's for which v is a start node;*
- *$end_G : V_G \rightarrow 2^{I_{\mathcal{A}}}$ is a function that associates with each node $v \in V_G$, the set of attack graph id's for which v is an end node;*
- *For each $v \in V_G$, $tables_G(v)$ is a set of records of the form (current, attackID, ts_0, previous, next), where current is a reference to an observation tuple (alert), attackID $\in I_{\mathcal{A}}$ is an attack graph id, $ts_0 \in \mathcal{T}$ is a timestamp, previous and next are sets of references to records in $tables_G$;*
- *$completed_G : I_{\mathcal{A}} \rightarrow 2^{\mathcal{P}}$, where $\mathcal{P}$ is the set of references to records in $tables_G$, is a function that associates with each attack identifier $id(A)$ a set of references to records in $tables_G$ corresponding to completed attack occurrences.*□

Note that G, $start_G$, end_G can be computed a-priori, based on the set $\mathcal{A}$ of given attack graphs. All the index tables $tables_G$ are initially empty. As new alerts are received, index tables are updated accordingly, as described in Section 4.1. The index tracks information about which nodes are start and/or end nodes for the original attack graphs. For each node v, the index maintains a table $tables_G(v)$ that tracks partially completed attack occurrences, with each record pointing to an alert, as well as to previous and successor records. In addition, each record stores the time at which the partial occurrence began (ts_0).

4.1 Index Insertion Algorithm

This section describes an algorithm (Algorithm 1) to update the index when a new alert is received. The algorithm takes as input a temporal multi-attack graph index I_G, a new alert o_{new} to be added to the index, and a boolean flag f_{TF} indicating whether the Time Frame (TF) pruning strategy must be applied (we will explain the TF pruning strategy in Section 4.2).

Line 1 maps the newly received alert o_{new} to a known vulnerability exploit v_{new}. Lines 3–7 handle the case when v_{new} is the start node of an attack graph. A new record is added to $tables_G(v_{new})$ for every graph in $start_G(v_{new})$ (Lines 4-6), denoting the fact that o_{new} may represent the start of a new attack sequence.

Algorithm 1. $insert(o_{new}, I_G, f_{TF})$

Input: New alert to be processed o_{new}, temporal multi-attack graph index I_G, boolean flag f_{TF} indicating whether the Time Frame Pruning strategy must be applied.
Output: Updated temporal multi-attack graph index I_G.

```
1: v_new ← φ(o_new) // Map the new alert to a known vulnerability exploit
2: // Look at start nodes
3: if start_G(v_new) ≠ ∅ then
4:    for all id ∈ start_G(v_new) do
5:       add (o_new^↑, id, o_new.ts, ∅, ∅) to tables_G(v_new)
6:    end for
7: end if
8: // Look at intermediate nodes
9: for all node v ∈ V_G s.t. ∃id ∈ I_A, δ_G(v, v_new, id) ≠ null do
10:    if TF then
11:       r_first ← min{r ∈ tables_G(v) | o_new.ts − r.current.ts ≤
          max_{id∈I_A | δ_G(v,v_new,id)≠null} δ_G(v, v_new, id).t_max}
12:    else
13:       r_first ← tables_G(v).first
14:    end if
15:    for all record r ∈ tables_G(v) s.t. r ≥ r_first do
16:       id ← r.attackID
17:       if δ_G(v, v_new, id) ≠ ∅ then
18:          (t_min, t_max) ← δ_G(v, v_new, id)
19:          if (t_min ≤ t_new.ts − r.current.ts ≤ t_max) ∧ γ(v_new) then
20:             r_n ← (o_new^↑, id, r.ts_0, {r^↑}, ∅)
21:             add r_n to tables_G(v_new)
22:             r.next ← r.next ∪ {r_n^↑}
23:             // Look at end nodes
24:             if id ∈ end_G(v_new) then
25:                add r_n^↑ to completed_G(id)
26:             end if
27:          end if
28:       end if
29:    end for
30: end for
```

Lines 9–30 look at the tables associated with the nodes that precede v_{new} in the temporal multi-attack graph and check whether the new alert can be correlated to existing partially completed occurrences. For each predecessor v of v_{new}, Lines 10-14 determine where the algorithm should start scanning $tables_G(v)$. Note that records are added to the index as new alerts are generated. Therefore, records r in each index table $tables_G(v)$ are ordered by $r.current.ts$, i.e., the time at which the corresponding alert was generated. Given two records r_1, $r_2 \in tables_G(v)$, we use $r_1 \leq r_2$ to denote the fact that r_1 precedes r_2 in $tables_G(v)$, i.e., $r_1.current.ts \leq r_2.current.ts$. In the unrestricted case, the whole table is scanned, $tables_G(v).first$ being the first record in $tables_G(v)$. If TF pruning is being applied, only the "most recent" records in $tables_G(v)$ are considered. On Lines 18–19, time intervals labeling the edges of temporal attack graphs are used to determine whether the new alert can be linked to record r for the attack graph in $\mathcal{A}$ identified by $id = r.attackID$, given the amount of time elapsed between $r.current.ts$ and $o_{new}.ts$. Additionally, we verify whether the condition $\gamma(v_{new})$ labeling node v_{new} is satisfied. If all these conditions are met, a new record r_n is added to $tables_G(v_{new})$ and $r.next$ is updated to point to r_n (Lines 20–22). Note that r_n inherits ts_0 from its predecessor; this ensures that the starting and ending times can be quickly retrieved by looking directly

at the last record for a completed occurrence. Finally, lines 24–26 check whether v_{new} is an end node for some attack graph. If yes, a pointer to r_n is added to $completed_G$, telling that a new occurrence has been completed (Line 25).

Algorithm *insert*, can be used iteratively for loading an entire observation sequence at once (we refer to this variant as *bulk-insert*).

4.2 Performance

In its unrestricted version, algorithm *bulk-insert*, has quadratic time complexity w.r.t. the size of the observation sequence, as typical of a *nested loop* procedure. However, we propose a pruning strategy – called *Time Frame* (TF) – that leverages the temporal constraints on the unfolding of attacks and lowers the complexity of bulk loading, while not altering the solutions. Algorithm *insert* is said to apply *Time Frame pruning* if, for each alert o_{new} and each predecessor v of $v_{new} = \phi(o_{new})$, it starts scanning $tables_G(v)$ at the first record r such that $o_{new}.ts - r.current.ts \leq \max_{id \in I_{\mathcal{A}} \mid \delta_G(v, v_{new}, id) \neq \texttt{null}} \delta_G(v, v_{new}, id).t_{max}$.

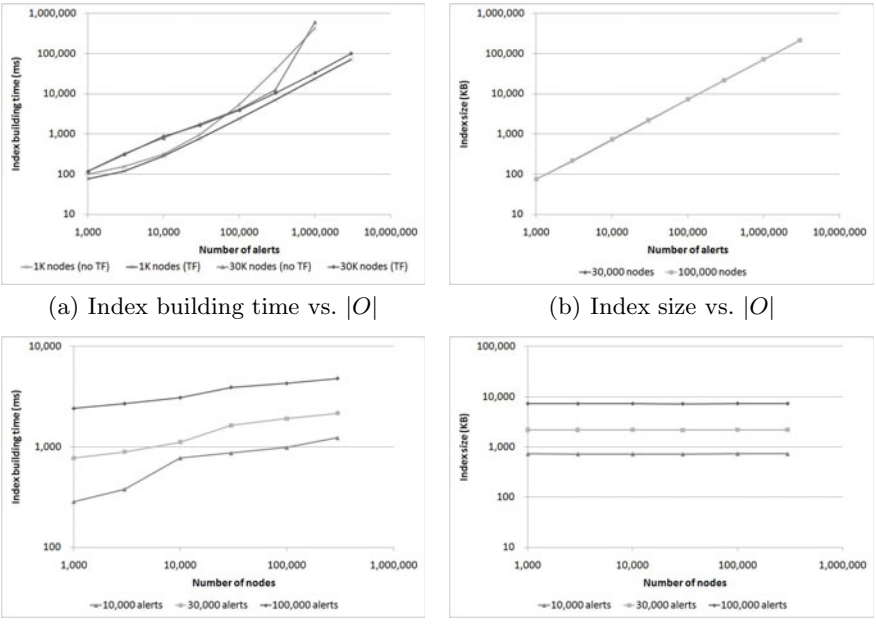
This strategy avoids scanning the entire predecessor table when most of the records in $tables_G(v)$ cannot be linked to records corresponding to v_{new}, because too long has passed since their corresponding alerts were generated. It can be proved that this strategy does not alter the search space, and the worst case complexity of the algorithm *insert* (resp. *bulk-insert*) when the TF pruning is applied is $O(k^{|V_G|} \cdot |\mathcal{A}|)$ (resp. $O(k^{|V_G|} \cdot |\mathcal{A}| \cdot |O|)$), where O is the observation sequence and k is the level of concurrency (i.e., maximum number of concurrent attacks). However, complexity is in practice lower than the worst case scenario and does not depend on the size of the graph, as experimentally shown in Section 5), since the number of index tables to be examined at each step is typically bounded and much smaller than $|V_G|$.

5 Preliminary Experiments

In this section, we report the results of the experiments we conducted to evaluate the time and memory performance of the proposed index.

We evaluated the system using synthetic datasets generated using two separate tools. The first tool generates random attack graphs by taking as input a set of vulnerabilities. The second tool generates alert sequences by simulating attacks described by the graphs generated using the first tool. We used the first tool to generate sets of attack graphs of varying size. We then used the second tool to generate sequences of 3 million alerts for each set of attack graphs.

Figure 4(a) shows how the time to build the entire index increases as the number of alerts increases. It is clear that, when TF is applied, the index building time is linear in the number of observations (note that both axes are on a log scale), and the bulk indexing algorithm can process between 25 and 30 thousands alerts per second. Also note that the size of the graphs does not significantly affects the index building time, as confirmed by Figure 4(c). In fact, when the size of the graphs changes by orders of magnitude, the processing time increases slightly, but remains in the same order of magnitude. This can be easily explained by considering that,



(a) Index building time vs. $|O|$

(b) Index size vs. $|O|$

(c) Index building time vs. $|V_G|$

(d) Index size vs. $|V_G|$

Fig. 4. Experimental results

for a given number of alerts, when the size of the graphs increases, the number of index tables (one for each $v \in V_G$) increases as well, but at the same time the average number of occurrences of each observation and the average size of each table decrease, keeping total memory occupancy (Figure 4(d)) and processing time (Figure 4(c)) almost constant. When TF is not applied, the index building time becomes quadratic, and starts to diverge. This effect is more evident for smaller graphs, as shown in Figure 4(a). This can also be explained by considering that, for a given number of alerts, when the size of the graphs decreases, there are fewer tables each containing a larger number of records, and the algorithm has to scan a constant number of such tables at each step. Finally, Figure 4(b) shows that memory occupancy is linear in the number of alerts, and confirms that it is independent from the size of the graphs.

In conclusions, we have shown that our framework can handle very large attack graphs, with hundreds of thousands of nodes, and can process incoming alerts at the rate of 25-30 thousands per second, which makes attack detection real-time for many real-world applications.

6 Conclusions and Future Work

In this paper, building on previous work on topological vulnerability analysis, we proposed an automated framework to manage very large attack graphs and

monitor high volumes of incoming alerts for the occurrence of known attack patterns in real-time. Specifically, we proposed (i) a data structure to concurrently monitor multiple types of attacks; (ii) an index structure to effectively index millions of time-stamped alerts; and (iii) a real-time algorithm to process a continuous stream of alerts and update the index. Experimental results confirmed that our framework enables real-time attack detection.

References

1. Albanese, M., Pugliese, A., Subrahmanian, V.S., Udrea, O.: MAGIC: A multi-activity graph index for activity detection. In: Proc. of the IEEE Intl. Conference on Information Reuse and Integration (IRI 2007), pp. 267–272 (August 2007)
2. Avrahami-Zilberbrand, D., Kaminka, G., Zarosim, H.: Fast and Complete Symbolic Plan Recognition: Allowing for Duration, Interleaved Execution, and Lossy Observations. In: Proc. of the AAAI Workshop on Modeling Others from Observations, MOO-2005 (2005)
3. Dain, O., Cunningham, R.K.: Fusing a heterogeneous alert stream into scenarios. In: Proc. of the 2001 Workshop on Data Mining for Sec. App., pp. 1–13 (2001)
4. Duong, T.V., Bui, H.H., Phung, D.Q., Venkatesh, S.: Activity Recognition and Abnormality Detection with the Switching Hidden Semi-Markov Model. In: Proc. of IEEE CVPR-2005, vol. 1, pp. 838–845 (2005)
5. Habra, N., Charlier, B., Mounji, A., Mathieu, I.: ASAX: Software Architecture and Rule-Based Language for Universal Audit Trail Analysis. In: Deswarte, Y., Quisquater, J.-J., Eizenberg, G. (eds.) ESORICS 1992. LNCS, vol. 648, pp. 435–450. Springer, Heidelberg (1992)
6. Hamid, R., Maddi, S., Johnson, A.Y., Bobick, A.F., Essa, I.A., Isbel Jr., C.L.: A novel sequence representation for unsupervised analysis of human activities. Artificial Intelligence 173(14), 1221–1244 (2009)
7. Jajodia, S., Noel, S.: Topological Vulnerability Analysis. In: Cyber Situational Awareness, pp. 139–154. Springer, Heidelberg (2010)
8. Lühr, S., Bui, H.H., Venkatesh, S., West, G.A.W.: Recognition of human activity through hierarchical stochastic learning. In: Proc. of the 1st IEEE Intl. Conf. on Pervasive Computing and Comm. (PerCom-2003), pp. 416–422 (2003)
9. Ning, P., Xu, D.: Learning attack strategies from intrusion alerts. In: Proc. of the 10th Conf. on Computer and Comm. Security (CCS 2003), pp. 200–209 (2003)
10. Noel, S., Robertson, E., Jajodia, S.: Correlating intrusion events and building attack scenarios through attack graph distances. In: Proc. of the 20th Annual Computer Security Applications Conference, ACSAC 2004, pp. 350–359 (2004)
11. Paxson, V.: Bro: a system for detecting network intruders in real-time. Computer Networks 31(23-24), 2435–2463 (1999)
12. Qin, X., Lee, W.: Statistical Causality Analysis of INFOSEC Alert Data. In: Vigna, G., Krügel, C., Jonsson, E. (eds.) RAID 2003. LNCS, vol. 2820, pp. 73–93. Springer, Heidelberg (2003)
13. Valdes, A., Skinner, K.: Probabilistic Alert Correlation. In: Lee, W., Mé, L., Wespi, A. (eds.) RAID 2001. LNCS, vol. 2212, pp. 54–68. Springer, Heidelberg (2001)
14. Wang, L., Liu, A., Jajodia, S.: Using attack graphs for correlating, hypothesizing, and predicting intrusion alerts. Computer Comm. 29(15), 2917–2933 (2006)

An Effective Research of Emotion

Dong Hwa Kim[1], Khalid Saeed[2], Cha Geun Jung[3]

[1] Dept. of Instrumentation and Control Eng., Hanbat National University
koreahucare@gmail.com
[2] Dept. of Physics and Applied Computer Science AGH University in Krakow
saeed@agh.edu.pl1
[3] Dept. of System and Control Eng., Hoseu University
cheong@hoseo.edu

Abstract. This paper proposes the importance of emotion technology for robot. Emotion technology is absolutely necessary for real human being in robot. However, it is very difficult to expression effectively emotion function in robot. Visual system (static and moving status), sound system, and tactile sense can express by physical sensor. However, there are some difficulties about how we can express feeling like human being's hart. Therefore, those who work in science have to cooperate with researcher in quite difference area such as, psychologist, philosopher, musician, material science, and etc.

1 Introduction

Many research center and University have been prospecting robot as economic driving force and future technology. However, robot technology should be developed more than recent status. Of course, there so many kinds of technologies are needed for real robot in various areas. Emotion technology is one of them. Because emotion absolutely seems to be a complex phenomenon we cannot express easily. Also, reasonable and unified theory lacks. Happiness, surprise, enthusiasm, anger, fear, and disgust are some of the emotion. A neutral emotion is also can defined with which emotions can occur concurrently and cause one to yield to the other thereby changing the behavior of emoting. Really expressing and modeling emotion may be tricky. Especially, psychologically expression might be more difficult. A weighting, normalizing and scaling mechanism should be used to express an intensity of the emotion from sensors on a robot.

2 Physiologically Derived Emotions

Physical function is happy, sad, cry, fear, etc. We can measure by the input sensor for these functions to application. Emotion generated by the physical functions can drive the shape or mood of body or face. That is, technical measuring function can take signal from the several emotions by Happiness, Fear and Anger, etc and resolves them to generate the internal resultant mood of the system for application such as, robot, intelligence control, and etc. The Multi-Hybrid Architecture forEmotion Express

N. Chaki and A. Cortesi (Eds.): CISIM 2011, CCIS 245, pp. 19–21, 2011.

The emotion architecture is composed of six systems (see Fig. 1). The input system (vision or signal) is responsible for identifying and tracking the goal. The action system is responsible for choosing higher level emotion (robot motions) to move the robot control system to a specified goal. In this system, these six systems must cooperate to achieve the overall task of reaching the goal control goal or position (robot). The systems are also competing—there are some trade-offs between them. For example, both the sad and the angry system compete for the action system. The sad system needs input for action, while the angry system needs sad level for target detection and tracking. For this cooperation and competition, each system generates bids for the information offered from the input and action system.

The information actually executed by each system depends on the winning bid at each point in time.

The emotion itself is implemented as a multiagent system. This system is composed of six agents with the emotion functions.

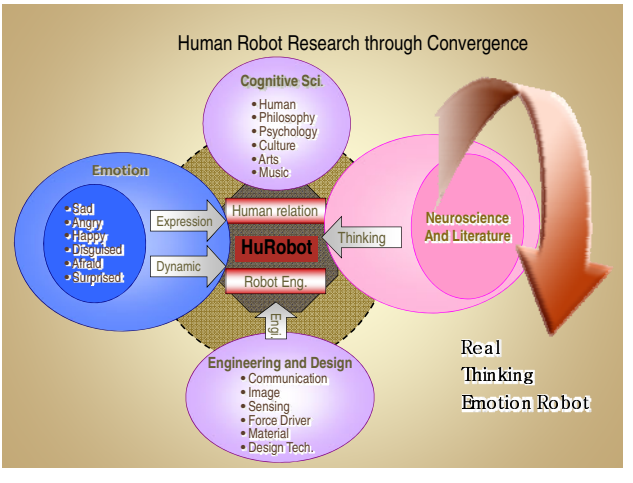


Fig. 1. Architecture of agent system for Human Robot

3 Psychological Phenomenon Derived Emotion

Adrenalin is known to be generated within the body during conditions of anguish, fear and anger. That is, these emotions are fuzzy system processed by a fuzzy mood generation module to provide the current mood of the system. This fuzzy function takes the several functions from different agents and resolves them to generate the internal resultant mood of human being. In order to know that emotions are based on a psychological aspect of feeling and a physical or physiological aspect of bodily reaction, we have to introduce psychology related areas into research topic. Because emotion reactions are triggered by feelings and external reactions, culture and psychology are also to be studied. Feeling that is, one of psychological components of emotion and the results can give an impact on the bodily reaction. Emotions are very complex

function. For example anger is a complex emotion comprising of a feeling of irritability and a bodily reaction of rapid heartbeats, reddened face, etc. It is necessary to clearly research related topics for complex emotions. Recent studies in consciousness have tried to understand what feeling really is and it is essential to identify the emotions and also the associated feeling and reaction components.

4 Results and Discuss

An emotional function sound, smell, and touch can be sensed from its physical sensor. However, happy, fear, and anger can be obtained from physical functions because those individually be generated based on a set of sensors with more sophisticated signal conditioning. Therefore a concept of emotional based on psychology resource have to be introduced for human being emotional function. We have used a fuzzy module to realize this emotional dynamic which are generated by a multi-agent system. This paper suggests how we express and generate emotional function by both methods physical and psychological approach. Therefore, we have to make a strong human network for several topics.

For this, tests may be carried out with real robots to understand the way in which emotional robots behave in an environment. Of course, this will require knowledge of the hardware on the robot side and communication software running on the computing end that in turn also runs the agents.

Acknowledgements. This work was supported by the Brainpool program 2010 of KOFST and thanks for supporting.

References

1. Lin, Chen, J.: Facial expressions classification with hierarchical radial basis function networks. In: Int. Conf. on Neural Information Processing (1994)
2. Yoshitomi, Y., Kim, S., Kawano, T., Kitazoe, T.: Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face. In: IEEE Int. Workshop on Robot and Human Interactive Communication (2000)
3. Nair, S.B., Kim, D.H.: Towards a Dynamic Emotional Model. In: IEEE-ISIE 2009, May, Seoul, Korea (2009)
4. Kim, D.H., Nair: KOFST Brain pool Report 2008 (2009)
5. Kim, D.H.: Novel dynamic express for robot. In: IEEE SAMI 2011, Slovakia (2011)
6. Kim, D.H., Baranyi, P.: Emotion Dynamic Express By Fuzzy Function For Emotion Robot. In: CogInform 2011, Budapest (2011)

Regularization of Large Scale Total Least Squares Problems

Heinrich Voss[1] and Jörg Lampe[2]

[1] Institute of Numerical Simulation, Hamburg University of Technology
D-21071 Hamburg, Germany
voss@tuhh.de
[2] Germanischer Lloyd SE, D-20457 Hamburg, Germany
joerg.lampe@gl-group.com

Abstract. The total least squares (TLS) method is an appropriate approach for linear systems when not only the right-hand side but also the system matrix is contaminated by some noise. For ill-posed problems regularization is necessary to stabilize the computed solutions. In this presentation we discuss two approaches for regularizing large scale TLS problems. One which is based on adding a quadratic constraint and a Tikhonov type regularization concept.

1 Introduction

Many problems in data estimation are governed by over-determined linear systems

$$Ax \approx b, \quad A \in \mathbb{R}^{m\times n},\ b \in \mathbb{R}^m, m \geq n. \tag{1}$$

In the classical least squares approach the system matrix A is assumed to be free from error, and all errors are confined to the observation vector b. However, in engineering application this assumption is often unrealistic. For example, if not only the right-hand side b but A as well are obtained by measurements, then both are contaminated by some noise.

An appropriate approach to this problem is the total least squares (TLS) method which determines perturbations $\Delta A \in \mathbb{R}^{m\times n}$ to the coefficient matrix and $\Delta b \in \mathbb{R}^m$ to the vector b such that

$$\|[\Delta A, \Delta b]\|_F^2 = \min! \quad \text{subject to } (A + \Delta A)x = b + \Delta b, \tag{2}$$

where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix. An overview of TLS methods and a comprehensive list of references is contained in [18].

When solving practical problems they are usually ill-conditioned, for example the discretization of ill-posed problems such as integral equations of the first kind. Then least squares or total least squares methods for solving (1) often yield physically meaningless solutions, and regularization is necessary to stabilize the computed solution.

N. Chaki and A. Cortesi (Eds.): CISIM 2011, CCIS 245, pp. 22–37, 2011.

A well established approach is to add a quadratic constraint to problem (2) yielding the regularized total least squares (RTLS) problem

$$\|[\Delta A, \Delta b]\|_F^2 = \min! \quad \text{subject to } (A + \Delta A)x = b + \Delta b,\ \|Lx\| \le \delta, \tag{3}$$

where $\|\cdot\|$ denotes the Euclidean norm, $\delta > 0$ is the quadratic constraint regularization parameter, and the regularization matrix $L \in \mathbb{R}^{p\times n}$, $p \le n$ defines a (semi-) norm on the solution through which the size of the solution is bounded or a certain degree of smoothness can be imposed [6,10,11,14,19,20].

If the minimization problem (3) attains its solution then it can be rewritten into the more tractable form

$$f(x) := \frac{\|Ax - b\|^2}{1 + \|x\|^2} = \min! \quad \text{subject to} \quad \|Lx\| \le \delta. \tag{4}$$

Closely related is an approach which adopts Tikhonov's regularization concept to stabilize the TLS solution [2,15]:

$$f(x) + \lambda\|Lx\|^2 = \min!. \tag{5}$$

By comparing the corresponding Lagrangian functions of problems (4) and (5) it is obvious that they are equivalent in the following sense. For each Tikhonov parameter $\lambda \ge 0$ there exists a corresponding value of the quadratic constraint δ such that that the solutions of (4) and (5) are identical.

In this paper we review numerical methods for large scale regularized total least squares problems of both types (4) and (5). In Section 2 the first order conditions of (4) are solved via a sequence of quadratic or linear eigenvalue problems; and in Section 3 the nonlinear first order condition of (5) is solved by Newton's method. To avoid the solution of large scale eigenproblems and nonlinear systems of equations we apply iterative projection methods which allow for an excessive reuse of information from previous iteration steps. The efficiency of these approaches is evaluated with a couple of numerical examples in Section 4. Conclusions can be found in Section 5.

2 Quadratically Constrained Total Least Squares Problems

We consider the quadratically constrained TLS problem (4). We assume that the solution x_{RTLS} is attained and that the inequality is active, i.e., $\delta = \|Lx_{RTLS}\|$ (otherwise no regularization would be necessary). Under this condition Golub, Hansen and O'Leary [6] derived the following first order necessary conditions: The solution x_{RTLS} of problem (4) is a solution of the problem

$$(A^TA + \lambda_I I_n + \lambda_L L^TL)x = A^Tb, \tag{6}$$

where the parameters λ_I and λ_L are given by

$$\lambda_I = -f(x), \quad \lambda_L = \frac{1}{\delta^2}\big(b^T(b - Ax) - f(x)\big). \tag{7}$$

This condition was used in the literature in two ways to solve problem (4): In [10,20] for a chosen parameter λ_I problem (6) is solved for (x, λ_L), which yields a convergent sequence of updates for λ_I. Conversely, in [6,11,19] λ_I is chosen as a free parameter; for fixed λ_L problem (6) is solved for (x, λ_I), and then λ_L is updated in a way that the whole process converges to the solution of (4).

In either case the first order conditions require to solve a sequence of quadratic and linear eigenvalue problems, respectively. Efficient implementations recycling a great deal of information from previous iteration steps are presented in [10] for the first approach and in [11] for the latter one.

2.1 RTLS via a Sequence of Quadratic Eigenvalue Problems

The first algorithm is based on keeping the parameter λ_I fixed for one iteration step and treating $\lambda := \lambda_L$ as a free parameter. The first order optimality conditions then reads

$$B(x^k)x + \lambda L^T L x = A^T b, \quad \|Lx\|^2 = \delta^2, \tag{8}$$

with

$$B(x^k) = A^T A - f(x^k)I, \;\; f(x^k) = -\lambda_I(x^k). \tag{9}$$

which suggests the following Algorithm 1.

Algorithm 1. RTLSQEP

Require: Initial vector x^1.

1: **for** $k = 1, 2, \ldots$ until convergence **do**

2: With $B_k := B(x^k)$ solve

$$B_k x^{k+1} + \lambda L^T L x^{k+1} = A^T b, \quad \|Lx^{k+1}\|^2 = \delta^2 \tag{10}$$

for (x^{k+1}, λ) corresponding to the largest $\lambda \in \mathbb{R}$

3: **end for**

It was shown in [12] that this algorithm converges in the following sense: Any limit point x^* of the sequence $\{x^k\}$ constructed by Algorithm 1 is a global minimizer of the minimization problem

$$f(x) = \frac{\|Ax - b\|^2}{1 + \|x\|^2} \quad \text{subject to } \|Lx\|^2 = \delta^2.$$

Sima, Van Huffel and Golub [20] proposed to solve (10) for fixed k via a quadratic eigenvalue problem. This motivates the name RTLSQEP of the algorithm.

If L is square and nonsingular, then with $z = Lx^{k+1}$ problem (10) is equivalent to

$$W_k z + \lambda z := L^{-T} B_k L^{-1} z + \lambda z = L^{-T} A^T b =: h, \quad z^T z = \delta^2. \tag{11}$$

Assuming that $W_k + \lambda I$ is positive definite, and denoting $u := (W_k + \lambda I)^{-2} h$, one gets $h^T u = z^T z = \delta^2$, and $h = \delta^{-2} h h^T u$ yields that $(W_k + \lambda I)^2 u = h$ is equivalent to the quadratic eigenvalue problem

$$Q_k(\lambda) u := (W_k + \lambda I)^2 u - \delta^{-2} h h^T u = 0. \tag{12}$$

If the regularization matrix L is not quadratic the problem (8) has to be reduced to the range of L to obtain the quadratic eigenproblem of the same type as (11), cf. [10,20].

In [10] $W_k + \hat{\lambda} I$ is shown to be positive semidefinite, with $\hat{\lambda}$ as the rightmost eigenvalue of (12). We are only considering the generic case of $W_k + \hat{\lambda} I$ being positive definite. In this case the solution of the original problem (10) is recovered from $z = (W_k + \hat{\lambda} I) u$, and $x^{k+1} = L^{-1} z$ where u is an eigenvector corresponding to $\hat{\lambda}$ which is scaled such that $h^T u = \delta^2$. The case that $W_k + \hat{\lambda} I \geq 0$ is singular means that the solution of (12) may not be unique, which is discussed by Gander, Golub and von Matt [5,9].

Remark 2.1. The transformation to the quadratic eigenproblem (12) seems to be very costly because of the inverse of L. Notice however, that typical regularization matrices are discrete versions of 1D first (or second) order derivatives like

$$L = \begin{bmatrix} 1 & -1 & & \\ & \ddots & \ddots & \\ & & 1 & -1 \end{bmatrix} \in \mathbb{R}^{(n-1)\times n}. \tag{13}$$

Smoothing properties of L are not deteriorated significantly if they are replaced by regular versions like (cf. [3])

$$\hat{L} := \begin{bmatrix} 1 & -1 & & \\ & \ddots & \ddots & \\ & & 1 & -1 \\ & & & \varepsilon \end{bmatrix} \quad \text{or} \quad \hat{L} := \begin{bmatrix} \varepsilon & & & \\ -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{bmatrix}$$

with some small diagonal element $\varepsilon > 0$. Which one of these modifications is chosen depends on the behavior of the solution of (4) close to the boundary. It is not necessary to calculate the inverse of L explicitly since the presented algorithms touch L^{-1} only by matrix-vector multiplications. And solving a system with $\hat{L}$ is cheap, i.e., an $\mathcal{O}(n)$–operation.

If it holds that $\operatorname{rank}(L) < n$ and one is not willing to perturb L, a basis of the range and kernel are needed. In [10,20] it is explained how to obtain a similar expression for W_k in the quadratic eigenvalue problem (12). For the L in (13) the spectral decomposition is explicitly known. For solutions on a 2D or 3D domain one can take advantage of Kronecker product representations of L and its decomposition, cf. [14]. □

An obvious approach for solving the quadratic eigenvalue problem (12) at the k-th iteration step of Algorithm 1 is linearization, i.e., solving the linear eigenproblem

$$\begin{bmatrix} -2W_k & -W_k^2+\delta^{-2}hh^T \\ I & 0 \end{bmatrix} \begin{bmatrix} \lambda u \\ u \end{bmatrix} = \lambda \begin{bmatrix} \lambda u \\ u \end{bmatrix}, \tag{14}$$

and choosing the maximal real eigenvalue, and the corresponding u–part of the eigenvector, which is an eigenvector of (12).

This approach is reasonable if the dimension n of problem (1) is small. For larger n it is not efficient to determine the entire spectrum of (12). In this case one could apply the implicitly restarted Arnoldi method implemented in ARPACK [16] (and included in MATLAB as function `eigs`) to determine the rightmost eigenvalue and corresponding eigenvector of (12). However, it is a drawback of linearization that symmetry properties of the quadratic problem are destroyed.

More efficient for large scale problems are the second order Arnoldi reduction (SOAR) introduced by Bai and Su [1] or a Krylov subspace projection method for quadratic eigenvalue problems proposed by Li and Ye [17]. The paper [14] contains implementation details for these methods and a couple of numerical examples demonstrating their efficiency.

In Algorithm 1 we have to solve a sequence of quadratic eigenproblems where due to the convergence of the sequence $\{f(x_k)\}$ the matrices

$$W_k = L^{-T}(A^TA - f(x_k)I)L^{-1} = L^{-T}A^TAL^{-1} - f(x_k)L^{-T}L^{-1} \tag{15}$$

converge as well. This suggests to reuse as much information from previous steps as possible when solving $Q_k(\lambda)u = 0$.

The only degree of freedom in the Krylov methods mentioned above is the initial vector. Thus, the only information that can be recycled from previous iterations is the eigenvector of the preceding step that can be used as initial vector. Much more information can be exploited in general iterative projection methods such as the nonlinear Arnoldi algorithm [21] which can be started with the entire search space of the previous eigenvalue problem. Notice, that the nonlinear Arnoldi method may be applied to more general nonlinear eigenvalue problem $T(\lambda)x = 0$.

Algorithm 2. Nonlinear Arnoldi

Require: Initial basis V, $V^TV = I$
1: Find rightmost eigenvalue λ of $V^TT(\lambda)V\tilde{u} = 0$ and corresponding eigenvector $\tilde{u}$
2: Determine preconditioner $P \approx T(\sigma)^{-1}$, σ close to wanted eigenvalue
3: Set $u = V\tilde{u}$, $r = Q_k(\lambda)u$
4: **while** $\|r\|/\|u\| > \epsilon_r$ **do**
5: $v = Pr$
6: $v = v - VV^Tv$
7: $\tilde{v} = v/\|v\|$, $V = [V, \tilde{v}]$
8: Find rightmost eigenvalue λ of $V^TT(\lambda)V\tilde{u} = 0$ and corr. eigenvector $\tilde{u}$
9: Set $u = V\tilde{u}$, $r = Q_k(\lambda)u$
10: **end while**

Some comments on an efficient implementation of RTLSQEP with the nonlinear Arnoldi solver are in order.

- A suitable initial basis V of Algorithm 2 for the first quadratic eigenvalue problem (12) can be determined by a small number of Lanczos steps applied to the linear eigenproblem $W_1 z = \lambda z$ with a random starting vector z_0 because this is cheaper than executing the nonlinear Arnoldi method.
- Since the dimensions of the projected problems are small they can be solved by linearization and a dense eigensolver like the QR algorithm.
- In our numerical examples it turned out that we obtained fast convergence without preconditioning, so we simply set $P = I$.
- The representation of $W_k = L^{-T}(A^T A - f(x_k)I)L^{-1}$ demonstrates that the projected eigenvalue problem

$$V^T Q_k(\lambda) V \tilde{u} = ((W_k + \lambda I)V)^T((W_k + \lambda I)V)\tilde{u} - \delta^{-2}(h^T V)^T(h^T V)\tilde{u} = 0$$

  can be determined efficiently if the matrices $L^{-T}A^T AL^{-1}V$, $L^{-T}L^{-1}V$ and $h^T V$ are known. These can be updated cheaply by appending in every iteration step of the nonlinear Arnoldi method one column and component to the current matrices and vector, respectively.

The considerations above demonstrate that due to the reuse of the entire search space it is rather inexpensive to provide $V^T Q_k(\lambda)V$ if $V^T Q_{k-1}(\lambda)V$ is known. This suggests early updates, i.e., to leave the inner loop of the nonlinear Arnoldi method for determining the rightmost eigenpair long before convergence. It turned out that while reducing the residual of the approximated rightmost eigenpair of a quadratic eigenproblem in step k by a factor 100 (instead of solving it to full accuracy), sufficient new information is added to the search space V. So the stopping criterion in Line 4 of Algorithm 2 is replaced by $\|r\|/\|r_0\| > 0.01$ with the initial residual r_0 calculated in Line 3. This approach leads to more outer iterations but overall to less inner iterations.

The early update variant reduces the overall computation time substantially when compared to the standard version. Implementing early update strategies in the Krylov-type algorithms destroyed the convergence of the overall process.

2.2 RTLS via a Sequence of Linear Eigenvalue Problems

The second algorithm is based on keeping the parameter λ_L fixed for one iteration step and letting $\lambda := -\lambda_I$ be a free parameter.

The following version of the first order optimality conditions was proven by Renaut and Guo in [19].

Theorem 1. *The solution x_{RTLS} of the RTLS problem (3) subject to the active constraint satisfies the augmented eigenvalue problem*

$$B(\lambda_L(x_{RTLS}))\begin{bmatrix} x_{RTLS} \\ -1 \end{bmatrix} = -\lambda_I(x_{RTLS})\begin{bmatrix} x_{RTLS} \\ -1 \end{bmatrix}, \tag{16}$$

with

$$B(\lambda_L) := M + \lambda_L N, \quad M := [A, b]^T[A, b], \quad N := diag\{L^T L, -\delta^2\}$$

and λ_L and λ_I as given in (7).

This condition suggested Algorithm 3 called RTLSEVP for obvious reasons.

Algorithm 3. RTLSEVP

Require: Initial guess $\lambda_L^0 > 0$ and $B_0 = B(\lambda_L^0)$

1: **for** $k = 1, 2, \ldots$ until convergence **do**

2: Solve

$$B_{k-1} y^k = \lambda y^k \tag{17}$$

for eigenpair (y^k, λ) corresponding to the smallest λ

3: Scale y^k such that $y^k = \begin{bmatrix} x^k \\ -1 \end{bmatrix}$

4: Update $\lambda_L^k = \lambda_L(x^k)$ and $B_k = B(\lambda_L^k)$

5: **end for**

The choice of the smallest eigenvalue is motivated by the fact that we are aiming at $\lambda = -\lambda_I$ (cf. (16)), and by the first order conditions of (4) it holds that

$$-\lambda_I = f(x) = \frac{\|Ax - b\|^2}{1 + \|x\|^2}$$

is the function to be minimized.

The straightforward idea in [7] to update λ_L in Line 4 with (7), i.e.,

$$\lambda_L^{k+1} = \frac{1}{\delta^2}\left(b^T(b - Ax^{k+1}) - \frac{\|Ax^{k+1} - b\|^2}{1 + \|x^{k+1}\|^2}\right) \tag{18}$$

does not lead in general to a convergent algorithm.

To enforce convergence Renaut and Guo [19] proposed to determine a value θ such that the eigenvector $(x_\theta^T, -1)^T$ of $B(\theta)$ corresponding to the smallest eigenvalue of $B(\theta)$ satisfies the constraint $\|Lx_\theta\|^2 = \delta^2$, i.e., find a non-negative root $\hat{\theta}$ of the real function

$$g(\theta) := \frac{\|Lx_\theta\|^2 - \delta^2}{1 + \|x_\theta\|^2}. \tag{19}$$

Then the corresponding eigenvector $(x_{\hat{\theta}}^T, -1)^T$ is a solution of (4).

Solving the first order conditions in this way assumes that the last component of the eigenvector can be scaled to be -1. But the last component of an eigenvector corresponding to the smallest eigenvalue of $B(\theta)$ does not need to be different from zero, and then $g(\theta)$ is not necessarily defined. To fill this gap the following generalization was introduced in [11]:

Definition 1. *Let $\mathcal{E}(\theta)$ denote the eigenspace of $B(\theta)$ corresponding to its smallest eigenvalue. Then*

$$g(\theta) := \min_{y \in \mathcal{E}(\theta)} \frac{y^T N y}{y^T y} = \min_{(x^T, x_{n+1})^T \in \mathcal{E}(\theta)} \frac{\|Lx\|^2 - \delta^2 x_{n+1}^2}{\|x\|^2 + x_{n+1}^2} \tag{20}$$

is the minimal eigenvalue of the projection of N onto $\mathcal{E}(\theta)$.

This extends the definition of g to the case of eigenvectors with zero last components. The modified function g has the following properties which were proven in [11].

Theorem 2. *The function $g : [0, \infty) \to \mathbb{R}$ has the following properties:*

(i) *If $\sigma_{\min}([A, b]) < \sigma_{\min}(A)$ then $g(0) > 0$*
(ii) *$\lim_{\theta \to \infty} g(\theta) = -\delta^2$*
(iii) *If the smallest eigenvalue of $B(\theta_0)$ is simple, then g is continuous at θ_0*
(iv) *g is monotonically not increasing on $[0, \infty)$*
(v) *Let $g(\hat{\theta}) = 0$ and let $y \in \mathcal{E}(\hat{\theta})$ such that $g(\hat{\theta}) = y^T N y / \|y\|^2$. Then the last component of y is different from 0.*
(vi) *g has at most one root.*

Theorem 2 demonstrates that if $\hat{\theta}$ is a positive root of g, then $x := -y(1 : n)/y_{n+1}$ solves the RTLS problem (4) where y denotes an eigenvector of $B(\hat{\theta})$ corresponding to its smallest eigenvalue.

REMARK 2.2. If the smallest singular value $\sigma_{n+1}(\tilde{\theta})$ of $B(\tilde{\theta})$ is simple, then it follows from the differentiability of $\sigma_{n+1}(\theta)$ and its corresponding right singular vector that

$$\frac{d\sigma_{n+1}(B(\theta))}{d\theta}\Big|_{\theta=\tilde{\theta}} = g(\tilde{\theta}). \tag{21}$$

Hence, searching the root of $g(\theta)$ can be interpreted as searching the maximum of the minimal singular values of $B(\theta)$ with respect to θ. □

REMARK 2.3. Notice that g is not necessarily continuous. If the multiplicity of the smallest eigenvalue of $B(\theta)$ is greater than 1 for some θ_0, then g may have a jump discontinuity at θ_0. It may even happen that g does not have a root, but it jumps below zero at some θ_0. This indicates a nonunique solution of problem (4). See [11] how to construct a solution in this case. Here we assume a unique solution, which is the generic case. □

To determine the minimum eigenvalue and corresponding eigenvector of

$$B(\theta_k)y = \big(M + \theta_k N\big)y = \lambda y \tag{22}$$

Renaut and Guo [19] proposed the Rayleigh quotient iteration initialized by the eigenvector found in the preceding iteration step. Hence, one uses information from the previous step, but an obvious drawback of this method is the fact that each iteration step requires $\mathcal{O}(n^3)$ operations providing the LU factorization of $B(\theta_k)$.

Similar to the approach in RTLSQEP the entire information gathered in previous iteration steps can be employed solving (22) via the nonlinear Arnoldi Algorithm 2 with thick starts applied to

$$T_k(\lambda)u = (M + \theta_k N - \lambda I)u = 0$$

This time in lines 1 and 8 we aim at the minimum eigenvalue of $T_k(\lambda)$.

The projected problem

$$V^T T_k(\lambda) V\tilde{u} = \big(([A,b]V)^T([A,b]V) + \theta_k V^T N V - \lambda I\big)\tilde{u} = 0 \tag{23}$$

can be updated efficiently in both cases, if the search space is expanded by a new vector and if the iteration counter k is increased (i.e., a new θ_k is chosen).

The explicit form of the matrices M and N is not needed. If a new vector v is added to the search space span$\{V\}$, the matrices $A_V := [A,b]V$ and $L_V := LV(1:n,:)$ are refreshed appending the new column $A_v := [A,b]v$ and $L_v := Lv(1:n)$, respectively, and the projected matrix $M_V := V^T M V$ has to be augmented by the new last column $c_v := (A_V^T A_v; A_v^T A_v)$ and last row c_v^T. Since the update of L_V is usually very cheap (cf. Remark 2.1) the main cost for determining the projected problem is essentially 1 matrix–vector multiplication.

For the preconditioner in Line 3 it is appropriate to chose $P \approx N^{-1}$ which according to Remark 2.1 usually can be implemented very cheaply and can be kept constant throughout the whole algorithm.

Assuming that g is continuous and strictly monotonically decreasing Renaut and Guo [19] derived the update

$$\theta_{k+1} = \theta_k + \frac{\theta_k}{\delta^2} g(\theta_k) \tag{24}$$

for solving $g(\theta) = 0$ where $g(\theta)$ is defined in (19), and at step k, $(x_{\theta_k}^T, -1)^T$ is the eigenvector of $B(\theta_k)$ corresponding to its minimal eigenvalue. Since this sequence usually does not converge, an additional backtracking was included, i.e., the update was modified to

$$\theta_{k+1} = \theta_k + \iota \frac{\theta_k}{\delta^2} g(\theta_k) \tag{25}$$

where $\iota \in (0,1]$ was bisected until the sign condition $g(\theta_k)g(\theta_{k+1}) \geq 0$ was satisfied. However, this safeguarding hampers the convergence of the method considerably.

We propose a root-finding algorithm taking into account the typical shape of $g(\theta)$. g has at most one root $\hat{\theta}$, and exactly one root in the generic case, which we assume here. Left of $\hat{\theta}$ the slope of g is often very steep, while right of $\hat{\theta}$ it is approaching its limit $-\delta^2$ quite quickly. This makes it difficult to determine $\hat{\theta}$ by Newton's method because for an initial value less than $\hat{\theta}$ convergence is extremely slow, and if the initial value exceeds $\hat{\theta}$ then usually Newton's method yields a negative iterate.

We approximate the root of g based on rational interpolation of g^{-1} (if it exists) which has a known pole at $\theta = -\delta^2$. Assume that we are given three pairs $(\theta_j, g(\theta_j))$, $j = 1,2,3$ with

$$\theta_1 < \theta_2 < \theta_3 \quad \text{and} \quad g(\theta_1) > 0 > g(\theta_3). \tag{26}$$

We determine the rational interpolation

$$h(\gamma) = \frac{p(\gamma)}{\gamma + \delta^2}, \text{ where } p \text{ is a polynomial of degree 2,}$$

and p is chosen such that $h(g(\theta_j)) = \theta_j$, $j = 1, 2, 3$, and we evaluate $\theta_4 = h(0)$. In exact arithmetic $\theta_4 \in (\theta_1, \theta_3)$, and we replace θ_1 or θ_3 by θ_4 such that the new triple satisfies (26).

The coefficients of p are obtained from a 3 by 3 linear system which may become very badly conditioned, especially close to the root. We therefore use Chebyshev polynomials transformed to the interval $[g(\theta_3), g(\theta_1)]$ as a basis for representing p.

If g is strictly monotonically decreasing in $[\theta_1, \theta_3]$ then h is a rational interpolation of $g^{-1} : [g(\theta_3), g(\theta_1)] \to \mathbb{R}$.

Due to nonexistence of the inverse g^{-1} on $[g(\theta_3), g(\theta_1)]$ or due to rounding errors very close to the root $\hat{\theta}$, it may happen that θ_4 is not contained in the interval (θ_1, θ_3). In this case we perform a bisection step such that the interval definitely still contains the root of g. If $g(\theta_2) > 0$, then we replace θ_1 by $\theta_1 = (\theta_2 + \theta_3)/2$, otherwise θ_3 is exchanged by $\theta_3 = (\theta_1 + \theta_2)/2$.

To initialize Algorithm 3 we determine three values θ_j such that not all $g(\theta_j)$ have the same sign. Given $\theta_1 > 0$ we multiply either by 0.01 or 100 depending on the sign of $g(\theta_1)$ and obtain after very few steps an interval that contains the root of g.

If a discontinuity at or close to the root is encountered, then a very small $\epsilon_\theta = \theta_3 - \theta_1$ appears with relatively large $g(\theta_1) - g(\theta_3)$. In this case we terminate the iteration and determine the solution as described in [11].

The evaluation of $g(\theta)$ can be performed efficiently by using the stored matrix $LV(1:n,:)$ to determine $\|Lu\|^2 = (LV(1:n,:)\tilde{u})^T(LV(1:n,:)\tilde{u})$ in much less than a matrix–vector multiplication.

3 Tikhonov Type Regularization

For the Tikhonov regularization problem (5) Beck and Ben–Tal [2] proposed an algorithm where in each iteration step a Cholesky decomposition has to be computed, which is prohibitive for large scale problems. We present a method which solves the first order conditions of (5) via an iterative projection method.

Different from the least squares problem the first order condition for (5) is a nonlinear system of equations. After some simplification one gets

$$q(x) := (A^T A + \mu L^T L - f(x)I)x - A^T b = 0, \quad \text{with } \mu := (1 + \|x\|^2)\lambda. \tag{27}$$

Newton's method then reads

$$x^{k+1} = x^k - J(x^k)^{-1} q(x^k)$$

with the Jacobian matrix

$$J(x) = A^T A + \mu L^T L - f(x)I - 2x\frac{x^T A^T A - b^T A - f(x)x^T}{1 + \|x\|^2}. \tag{28}$$

Taking advantage of the fact that J is a rank-one modification of

$$\hat{J}(x) := A^T A + \mu L^T L - f(x)I$$

and using the Sherman–Morrison formula, Newton's method obtains the following form:

$$x^{k+1} = \hat{J}_k^{-1}A^Tb - \frac{1}{1-(v^k)^T\hat{J}_k^{-1}u^k}\hat{J}_k^{-1}u^k(v^k)^T(x^k - \hat{J}_k^{-1}A^Tb), \qquad (29)$$

where

$$u^k := 2x^k/(1+\|x^k\|^2) \quad \text{and} \quad v^k := A^TAx^k - A^Tb - f(x^k)x^k.$$

Hence, in every iteration step we have to solve two linear systems with system matrix $\hat{J}(x^k)$, namely $\hat{J}_k z = A^Tb$ and $\hat{J}_k w = u^k$.

To avoid the solution of the large scale linear systems with varying matrices $\hat{J}_k$ we combine Newton's method with an iterative projection method similar to our approach for solving the eigenvalue problems in Section 2.

Assume that $\mathcal{V}$ is an ansatz space of small dimension k, and let the columns of $V \in \mathbb{R}^{n\times k}$ form an orthonormal basis of $\mathcal{V}$. We replace $z = \hat{J}_k^{-1}A^Tb$ with Vy_1^k where y_1^k solves the linear system $V^T\hat{J}_kVy_1^k = V^TA^Tb$, and $w = \hat{J}_k^{-1}u^k$ is replaced with Vy_2^k where y_2^k solves the projected problem $V^T\hat{J}_kVy_2^k = V^Tu^k$.

If

$$x_{k+1} = V_ky^k := V_ky_1^k - \frac{1}{1-(v^k)^TV_ky_2^k}V_ky_2^k(v^k)^T(x^k - V_ky_1^k)$$

does not satisfy a prescribed accuracy requirement, then $\mathcal{V}$ is expanded with the residual

$$q(x^{k+1}) = (A^TA + \mu L^TL - f(x^k)I)x^{k+1} - A^Tb$$

and the step is repeated until convergence.

Initializing the iterative projection method with a Krylov space $\mathcal{V} = \mathcal{K}_\ell(A^TA+\mu L^TL, A^Tb)$ the iterates x^k are contained in a Krylov space of $A^TA+\mu L^TL$. Due to the convergence properties of the Lanczos process the main contributions come from the first singular vectors of $[A; \sqrt{\mu}L]$, which for small μ are close to the first right singular vectors of A. It is common knowledge that these vectors are not always appropriate basis vectors for a regularized solution, and it may be advantageous to apply the regularization with a general regularization matrix L implicitly.

To this end we assume that L is nonsingular and we use the transformation $x := L^{-1}y$ of (5) (for general L we had to use the A-weighted generalized inverse $L_A^\dagger$, cf. [4]) which yields

$$\frac{\|AL^{-1}y - b\|^2}{1+\|L^{-1}y\|^2} + \lambda\|y\|^2 = \min!.$$

Transforming the first order conditions back and multiplying from the left with L^{-1} one gets

$$(L^TL)^{-1}(A^TAx + \mu L^TLx - f(x)x - A^Tb) = 0.$$

This equation suggests to precondition the expansion of the search space with L^TL or an approximation $P \approx L^TL$ which yields Algorithm 4.

Algorithm 4. Tikhonov TLS Method

Require: Initial basis V_0 with $V_0^T V_0 = I$, starting vector x^0
1: **for** $k = 0, 1, \ldots$ until convergence **do**
2: Compute $f(x^k) = \|Ax^k - b\|^2/(1 + \|x^k\|^2)$
3: Solve $V_k^T \hat{J}_k V_k y_1^k = V_k^T A^T b$ for y_1^k
4: Compute $u^k = 2x^k/(1 + \|x^k\|^2)$ and $v^k = A^T Ax^k - A^T b - f(x^k)x^k$
5: Solve $V_k^T \hat{J}_k V_k y_2^k = V_k^T u^k$ for y_2^k
6: Compute $x^{k+1} = V_k y_1^k - \frac{1}{1-(v^k)^T V_k y_2^k} V_k y_2^k (v^k)^T (x^k - V_k y_1^k)$
7: Compute $q^{k+1} = (A^T A + \mu L^T L - f(x^k)I)x^{k+1} - A^T b$
8: Compute $\tilde{r} = P^{-1} q^{k+1}$
9: Orthogonalize $\hat{r} = (I - V_k V_k^T)\tilde{r}$
10: Normalize $v_{\text{new}} = \hat{r}/\|\hat{r}\|$
11: Enlarge search space $V_{k+1} = [V_k, v_{\text{new}}]$
12: **end for**
13: Output: Approximate Tikhonov TLS solution x^{k+1}

Possible stopping criteria in Line 1 are the following ones (or a combination thereof):

- Stagnation of the sequence $\{f(x^k)\}$, i.e., the relative change of two consecutive values of $f(x^k)$ is small: $|f(x^{k+1}) - f(x^k)|/f(x^k)$ is smaller than a given tolerance.
- The relative change of two consecutive Ritz vectors x^k is small, i.e., $\|x^{k+1} - x^k\|/\|x^k\|$ is smaller than a given tolerance.
- The relative residual q^k from Line 7 is sufficiently small, i.e., $\|q^k\|/\|A^T b\|$ is smaller than a given tolerance.

Some remarks how to efficiently determine an approximate solution of the large scale Tikhonov TLS problem (5) with Algorithm 4 are in order. For large scale problems matrix valued operations are prohibitive, thus our aim is to carry out the algorithm with a computational complexity of $\mathcal{O}(mn)$, i.e., of the order of a matrix-vector product with a (general) dense matrix $A \in \mathbb{R}^{m \times n}$.

- Similar to Algorithms 1 and 3 the Tikhonov TLS methods allows for a massive reuse of information from previous iteration steps. Assume that the matrices V_k, $A^T AV_k$, $L^T LV_k$ are stored. Then, neglecting multiplications with L and L^T and solves with P (due to the structure of a typical regularization matrix L and a typical preconditioner P) the essential cost in every iteration step is only two matrix-vector products with dense matrices A and A^T for extending $A^T AV_k$. With these matrices $f(x^k)$ in Line 2 can be evaluated as

$$f(x^k) = \frac{1}{1 + \|y^k\|^2} \left((x^k)^T (A^T AV_k y^k) - 2(y^k)^T V_k^T (A^T b) + \|b\|^2 \right),$$

  and q^{k+1} in Line 7 can be determined according to

$$q^{k+1} = (A^T AV_k)y^{k+1} + \mu(L^T LV_k)y^{k+1} - f(x^k)x^{k+1} - A^T b.$$

- A suitable initial basis V_0 is an orthonormal basis of the Krylov space $\mathcal{K}_\ell\left(P^{-1}(A^TA+\mu L^TL), P^{-1}A^Tb\right)$ of small dimension, e.g. $\ell = 5$.
- Typically the initial vector $x^0 = 0$ is sufficiently close to the RTLS solution. In this case it holds that $f(x^0) = \|b\|^2$. If a general starting vector is used, e.g. the solution of the projected problem with initial space $\mathrm{span}\{V_0\}$ or some other reasonable approximation to x_{RTLS}, an additional matrix–vector product has to be spent for computing $f(x^0)$.
- It is enough to carry out one LDL^T-decomposition of the projected matrix $V_k^T\hat{J}_kV_k$, which then can be used twice to solve the linear systems in Lines 3 and 5.
- Since the the number of iteration steps until convergence is usually very small compared to the dimension n, the overall cost of Algorithm 4 is of the order $\mathcal{O}(mn)$.

4 Numerical Examples

To evaluate the performance of Algorithms 1, 3, and 4 we use large scale test examples from Hansen's *Regularization Tools* [8] which contains MATLAB routines providing square matrices $A_{\mathrm{true}} \in \mathbb{R}^{n\times n}$, right-hand sides b_{true} and true solutions x_{true}, with $A_{\mathrm{true}}x_{\mathrm{true}} = b_{\mathrm{true}}$. All examples originate from discretizations of Fredholm equations of the first kind and the matrices A_{true} and $[A_{\mathrm{true}}, b_{\mathrm{true}}]$ are ill-conditioned.

To adapt the problem to an overdetermined linear system of equations we stack two error-contaminated matrices and right-hand sides (with different noise realizations), i.e.,

$$A = \begin{bmatrix} \bar{A} \\ \bar{A} \end{bmatrix}, \qquad b = \begin{bmatrix} \bar{b} \\ \bar{b} \end{bmatrix},$$

with the resulting matrix $A \in \mathbb{R}^{2n\times n}$ and $b \in \mathbb{R}^{2n}$.

The regularization matrix L is chosen to be the nonsingular approximation to the scaled discrete first order derivative operator in one space-dimension, i.e., L in (13) with an additional last row $(0,\ldots,0,0.1)$. The numerical tests are carried out on an Intel Core 2 Duo T7200 computer with 2.3 GHz and 2 GB RAM under MATLAB R2009a (actually our numerical examples require less than 0.5 GB RAM).

Table 1 contains the results for six problems of dimension $n = 2000$ from the Regularization Toolbox for the three methods mentioned before. The underlying problems are all discrete versions of Fredholm's integral equations of the first kind. For problem *heat* the parameter κ controls the degree of ill-posedness.

Table 1 shows the relative residuals, number of iterations and of matrix-vector products, and the relative error averaged over 10 examples of the same type with different realizations of the error. σ denotes the noise level.

The examples demonstrate that the Tikhonov method outperforms the two methods based on eigenvalue solvers. For most of the examples the number of matrix-vector products of Algorithm 4 is about only $50-75\%$ of the ones required

Table 1. Problems from Regularization Tools

Problem	Method	$\frac{\|q(x^k)\|}{\|A^T b\|}$	Iters	MatVecs	$\frac{\|x-x_{true}\|}{\|x_{true}\|}$
phillips	TTLS	8.5e-16	8.0	25.0	8.9e-2
$\sigma = 1e-3$	RTLSQEP	5.7e-11	3.0	42.0	8.9e-2
	RTLSEVP	7.1e-13	4.0	47.6	8.9e-2
baart	TTLS	2.3e-15	10.1	29.2	1.5e-1
$\sigma = 1e-3$	RTLSQEP	1.0e-07	15.7	182.1	1.4e-1
	RTLSEVP	4.1e-10	7.8	45.6	1.5e-1
shaw	TTLS	9.6e-16	8.3	25.6	7.0e-2
$\sigma = 1e-3$	RTLSQEP	3.7e-09	4.1	76.1	7.0e-2
	RTLSEVP	2.6e-10	3.0	39.0	7.0e-2
deriv2	TTLS	1.2e-15	10.0	29.0	4.9e-2
$\sigma = 1e-3$	RTLSQEP	2.3e-09	3.1	52.3	4.9e-2
	RTLSEVP	2.6e-12	5.0	67.0	4.9e-2
heat(κ=1)	TTLS	8.4e-16	19.9	48.8	1.5e-1
$\kappa = 1e-2$	RTLSQEP	4.1e-08	3.8	89.6	1.5e-1
	RTLSEVP	3.2e-11	4.1	67.2	1.5e-1
heat(κ=5)	TTLS	1.4e-13	25.0	59.0	1.1e-1
$\sigma = 1e-3$	RTLSQEP	6.1e-07	4.6	105.2	1.1e-1
	RTLSEVP	9.8e-11	4.0	65.0	1.1e-1

for the RTLSQEP and RTLSEVP methods. This is reconfirmed by many more examples in [15]. The relative errors in the last column of Table 1 indicate that all three methods yield reliable numerical solutions.

5 Conclusions

Three methods for solving regularized total least squares problems are discussed two of which require the solution of a sequence of (linear or quadratic) eigenvalue problems and the third one needs the solution of a sequence of linear problems. In either case it is highly advantageous to combine the method with an iterative projection method and to reuse information gathered in previous iteration steps. Several examples demonstrate that all three methods require fairly small dimensions of the ansatz spaces and are qualified to solve large scale regularized total least squares problems efficiently.

In this paper we assumed that the regularization parameters λ, δ and μ, respectively are given. In practical problems this parameter has to be determined by some method like the L-curve criterion, the discrepancy principle, generalized cross validation, or information criteria. All of these approaches require to solve the RTLS problem repeatedly for many different values of the regularization parameter. Notice however, that the eigenvalue problems in Section 2 and the first order condition in Section 3 depend in a very simple way on the respective regularization parameter such that the matrices characterizing the projected problem for one parameter can be reused directly when changing the parameter.

For the quadratically constrained TLS problem and the L-curve criterion the details were worked out in [13] demonstrating the high efficiency of reusing search spaces when solving the sequence of eigenvalue problems.

References

1. Bai, Z., Su, Y.: SOAR: A second order Arnoldi method for the solution of the quadratic eigenvalue problem. SIAM J. Matrix Anal. Appl. 26, 640–659 (2005)
2. Beck, A., Ben-Tal, A.: On the solution of the Tikhonov regularization of the total least squares problem. SIAM J. Optim. 17, 98–118 (2006)
3. Calvetti, D., Reichel, L., Shuibi, A.: Invertible smoothing preconditioners for linear discrete ill-posed problems. Appl. Numer. Math. 54, 135–149 (2005)
4. Eldén, L.: A weighted pseudoinverse, generalized singular values, and constrained least squares problems. BIT 22, 487–502 (1982)
5. Gander, W., Golub, G., von Matt, U.: A constrained eigenvalue problem. Linear Algebra Appl. 114-115, 815–839 (1989)
6. Golub, G., Hansen, P., O'Leary, D.: Tikhonov regularization and total least squares. SIAM J. Matrix Anal. Appl. 21, 185–194 (1999)
7. Guo, H., Renaut, R.: A regularized total least squares algorithm. In: Van Huffel, S., Lemmerling, P. (eds.) Total Least Squares and Errors-in-Variable Modelling, pp. 57–66. Kluwer Academic Publisher, Dodrecht (2002)
8. Hansen, P.: Regularization tools version 4.0 for Matlab 7.3. Numer. Alg. 46, 189–194 (2007)
9. Lampe, J.: Solving regularized total least squares problems based on eigenproblems. Ph.D. thesis, Institute of Numerical Simulation, Hamburg University of Technology (2010)
10. Lampe, J., Voss, H.: On a quadratic eigenproblem occurring in regularized total least squares. Comput. Stat. Data Anal. 52/2, 1090–1102 (2007)
11. Lampe, J., Voss, H.: A fast algorithm for solving regularized total least squares problems. Electr. Trans. Numer. Anal. 31, 12–24 (2008)
12. Lampe, J., Voss, H.: Global convergence of RTLSQEP: a solver of regularized total least squares problems via quadratic eigenproblems. Math. Modelling Anal. 13, 55–66 (2008)
13. Lampe, J., Voss, H.: Efficient determination of the hyperparameter in regularized total least squares problems. Tech. Rep. 133, Institute of Numerical Simulation, Hamburg University of Technology (2009); To appear in Appl. Numer. Math., doi10.1016/j.apnum.2010.06.005
14. Lampe, J., Voss, H.: Solving regularized total least squares problems based on eigenproblems. Taiwanese J. Math. 14, 885–909 (2010)
15. Lampe, J., Voss, H.: Large-scale Tikhonov regularization of total least squares. Tech. Rep. 153, Institute of Numerical Simulation, Hamburg University of Technology (2011); Submitted to J. Comput. Appl. Math.
16. Lehoucq, R., Sorensen, D., Yang, C.: ARPACK Users' Guide. Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods. SIAM, Philadelphia (1998)

17. Li, R.C., Ye, Q.: A Krylov subspace method for quadratic matrix polynomials with application to constrained least squares problems. SIAM J. Matrix Anal. Appl. 25, 405–428 (2003)
18. Markovsky, I., Van Huffel, S.: Overview of total least squares methods. Signal Processing 87, 2283–2302 (2007)
19. Renaut, R., Guo, H.: Efficient algorithms for solution of regularized total least squares. SIAM J. Matrix Anal. Appl. 26, 457–476 (2005)
20. Sima, D., Van Huffel, S., Golub, G.: Regularized total least squares based on quadratic eigenvalue problem solvers. BIT Numerical Mathematics 44, 793–812 (2004)
21. Voss, H.: An Arnoldi method for nonlinear eigenvalue problems. BIT Numerical Mathematics 44, 387–401 (2004)

Reliability Estimation of Delay Tolerant QoS Mobile Agent System in MANET

Chandreyee Chowdhury and Sarmistha Neogy

Department of Computer Science and Engineering
Jadavpur University
Kolkata, India

Abstract. Mobile agents are recently used in wireless and mobile network applications because these save bandwidth and time. But reliability should be analyzed before mobile agent based systems (MAS) are used for broad range of applications in Mobile Adhoc Network (MANET). Here we propose an algorithm for estimating the task route reliability of MAS that is based on the conditions of the underlying MANET including QoS parameters.

The MAS consists of a no. of agent groups (having different QoS requirements). Each group corresponds to a particular application for which these agents are deployed. Thus different agent group has varying QoS requirements and hence perceive different connectivity pattern for the same MANET configuration. The complexity of MAS combined with the underlying dynamic topology of MANET drives us to estimate it using Monte Carlo simulation. The agents are so designed that they tolerate transient faults. The results achieved demonstrate the robustness of the proposed algorithm.

Keywords: Mobile Agent System, Mobile Ad hoc network, QoS, Monte-Carlo, Reliability, Mobility Model.

1 Introduction

A mobile agent is a combination of software program and data which migrates from a site to another site to perform tasks assigned by a user according to a static or dynamic route. It can be viewed as a distributed abstraction layer that provides the concepts and mechanisms for mobility and communication [1]. An agent consists of three components: the program which implements it, the execution state of the program and the data. It may migrate in two ways namely weak migration and strong migration [2]. The platform is the environment of execution. Typical benefits of using mobile agents include bandwidth conservation, reduced latency, load balancing etc. The route of the mobile agent can be decided by its owner or the agent can decide its next hop destination on the fly.

Here, we assume the underlying network to be a Mobile Ad Hoc Network (MANET) that typically undergoes constant topology changes, which disrupt the flow of information over the existing paths. Mobile agents are nowadays used in MANETs for various purposes like service discovery [3], network discovery, automatic network

N. Chaki and A. Cortesi (Eds.): CISIM 2011, CCIS 245, pp. 38–47, 2011.

reconfiguration, clustering etc. Due to motion and location independence [4] the reliability of underlying network becomes a factor that may affect the performance, availability, and strategy of mobile agent systems [5]. Features like scalability and reliability becomes critical in challenging environment like MANET [6]. However the scalability/reliability issue of agents has been highlighted in [7] but the work does not focus on MANET.

In this paper, we define a Mobile Agent-based System (MAS) to be a system consisting of different agent groups where each group accomplishes an independent task. Thus each group of agents has its own quality of service (QoS) requirements in terms of maximum delay. The ad hoc network has decentralized control, mobility of nodes and time-varying channels. All these parameters make it harder to control the resources in the network and maintain QoS for the agents while they accomplish their task and migrate in the network. If a mobile agent fails to move (that is, it tries to send its replica but no reply comes) at a particular time then it will wait and retry (sends another replica) up to a predefined number of times. The number of retries depends on the maximum delay that the application (which deployed the group of agents) can tolerate and the amount of work done till that point. Thus the ability of an agent to tolerate transient faults depends on its required QoS. We study the effect of QoS parameters on the behavior of MAS (having heterogeneous agent groups) and hence on MAS reliability in turn, with respect to the unique environmental issues in MANET.

In order to accomplish this goal we have considered a number of inherent environmental issues of MANET that affects MAS reliability, for example, transient faults (represented by non homogeneous Poison process), multipath propagation of signals and of course node mobility. Smooth Random mobility model (SRMM) is used to calculate node location at a given time. The connectivity between the nodes is calculated according to the two-ray propagation model [8] for signal propagation reflecting multipath propagation effect of radio signals. We estimate the reliability of a mobile agent based system using Monte Carlo (MC) simulation technique. This technique is used to avoid the typical computational complexity that may arise.

Section 2 discusses some contemporary work in this area followed by our work on reliability estimation in section 3.The simulation results are summarized in section 4. Finally section 5 concludes with an indication of our future endeavor in this area.

2 Related Works

Reliability analysis of MAS in adhoc network is a complicated problem for which little attention has been paid. Most of the work done in this area is related to distributed systems and distributed applications. But as pointed out in [6], features like scalability and reliability becomes critical in challenging environment with wireless networks. We did not see any work that considers transient environmental effects (apart from node mobility) and QoS into the reliability calculation for MANET. Due to the analytical complexity and computational cost of developing a closed-form solution, simulation methods, specifically MC simulation are often used

to analyze network reliability. In [9], analytical and MC-based methods are presented to determine the two-terminal reliability (2TR) for the adhoc scenario. The existence of links is considered in a probabilistic manner to account for the unique features of the MANET. However, there remains a gap in understanding the exact link between a probability and a specific mobility profile for a node. Later in [10] MC-based methods are presented to determine the 2TR for the adhoc scenario. It includes mobility models to allow mobility parameters to be varied and analyzed directly. The methods in this paper allow for the determination of reliability impacts under specific mobility considerations. The simplicity of this model often results in unrealistic conclusions. Little attention has been given to the reliability analysis of MAS. In [11], two algorithms have been proposed for estimating the task route reliability of MAS depending on the conditions of the underlying computer network. A third algorithm [12] is also stated based on random walk generation. However, in both the works the agents are assumed to be independent and the planning strategy seemed to be static. So this work does not address the scenario where agents can change their routes dynamically. In [13] a preliminary work has been done on estimating reliability of independent mobile agents roaming around in MANET. The protocol considers independent agents only. Node and link failure due to mobility or other factors is predicted according to NHPP. In [8] the MAS is assumed to be consisting of a no. of agent groups demanding for a minimum link capacity. The reliability calculation shows that even with large no. of heterogeneous agent groups with differing demands of link capacity, the MAS gradually reached a steady state.

3 Our Work

In this paper we assume that MAS (S) consists of independent agent groups and hence the tasks accomplished by each group do not depend on the movement or presence of any other agents of same or different group. However, agents of the same group tolerate same amount of delay for migration. The reliability (R) of (S) is defined as the conditional probability (p) that (S) is operational during a period of time [1] subject to the conditions of the underlying network. Consequently S is said to be fully operational if all its currently existing mobile agents are operational [2], whereas it is fully down if all its currently existing mobile agents are non-operational. Moreover, (S) can be partially operational if some of its currently existing mobile agents are operational. Here we propose a model for finding reliability. It is described in two parts - modeling agents on MANET and quantifying reliability of agents. Commonly used terms are summarized below.

R_s reliability of S; P_{link} link existence probability; avgDelay Average propagation delay; $\lambda_i(t)$ task route reliability of i^{th} agent in a step of simulation; r No. of retries; $\lambda(t)$ average reliability of all the agents; P_{node} the probability that $MN_i$is working correctly at time t that is the individual software reliability of MN_i; MN Mobile Node; Q no. of simulation steps; Δt time increments; D_{max} limit for delay tolerance;

3.1 Modeling Mobile Agent Based System considering QoS

We can think of a mobile agent as a token visiting one node to another in the network (if the nodes are connected) to accomplish its task [13], based on some strategy as needed by the underlying application. We start with a dynamic planning strategy where each agent is expected to visit N (<=no. of nodes in the network) nodes in the network to accomplish its task [13]. Each agent starts its journey from a given node which acts as its owner. Here we assume that agents belonging to the same owner share the same QoS requirement as the agents are expected to work collectively to achieve some goal set by the owner. Each mobile agent belonging to different groups may have different QoS (we have considered maximum tolerated delay of agent group in the present work) requirements. From the maximum tolerated delay (D_{max}) (specified for each group) we can calculate the maximum no. of retries r_{max} allowed for an agent using equation (1) as follows

$$r_{max} = \frac{D_{max}/N-1}{\text{avgDelay}} \tag{1}$$

If an agent is expected to visit N nodes then it will migrate (N-1) times. Thus the delay tolerance at one hop will be (D_{max}/N-1). Average propagation delay is the average time needed by the agent to traverse from current node to its next destination. Here D_{max} can be simulated using uniform distribution. Higher values of tolerance limit signify real time traffic (for example multimedia data) and lower values indicate computation intensive applications. So each agent having different QoS requirements will interpret the network differently, that is, in its own way. Thus a network may appear to be connected (and thus highly reliable) for computation intensive applications but for real time applications the same may appear to be unreliable. However, in equation 1 the progress of an agent is not considered. But in reality, r is decided by D_{max} along with its task progress.

So if there is a path between the current position and next destination of a mobile agent and both the nodes are operational, then only the agent will try to reach its next destination taking that path. But if received power is low then the agent will fail. Besides, if there is a transient error at the link, an agent may prefer to wait to overcome the effect (for example an obstacle in the path causing shadowing or may be frequency selective fading) according to the following equation

$$r< \left[\frac{\alpha*(D_{max}-\text{elapsed\_time})-(1-\alpha)*(N-visitednodes)*\text{avgDelay}}{avgDelay}\right] \tag{2}$$

In equation 2 α is weight that lies in the range 0<α<1. The first term of R.H.S of (2) represents the maximum remaining time for the agent before it retracts and reports back to its owner. Consequently the last term indicates an agent's progress of its task. The agent residing at a node at any point can randomly choose any other (preferably neighboring) node (from a list, if any is provided by the owner) to be its next destination. In this context R_s is calculated in the following subsection.

3.2 Modeling Reliability of MAS

System reliability R_s can be defined as

$$R_s = \{R_{MAS} | R_{MANET}\} \tag{3}$$

Reliability of MANET (R_{MANET}) is based on the probability that nodes are working (P_{node}) and connected according to a probability (P_{link}). P_{Node} can be shown to follow Weibull distribution [8]. Fault tolerance of the nodes can also be incorporated to further enhance the reliability to simulate node status [10]. P_{link} can be treated as a combination of the mobility model and link existence probability in terms of received power.

Now given such a MANET, we calculate (reliability of MAS) R_{MAS} as follows:

If an agent is in working condition (according to a Weibull distribution) and the agent can successfully visit M nodes out of N (desired) then it has accomplished M/N portion of its task. Thus reliability in this case will be M/N.

But if the application requires all N nodes to be visited in order to accomplish the task then reliability calculation will be modified as:

If an agent is in working condition (according to a Weibull distribution) and the agent can successfully visit all N nodes desired then reliability will be 1. Otherwise it will be 0.

Thus if agent reliability (M/N) <1, S is said to be partially operational. In this paper we assume that the tasks accomplished by each group do not depend on the movement or presence of any other agents of same or different group. But the only specialty is that agents of the same group share the same limit for tolerating delay in the network. So the probability that (S) is operational can be calculated as the mean of reliability of all its components (that is, agent groups in this system) given the current state of the nodes and links. The reliability of a group can again be calculated to be the mean of individual agent reliabilities. Thus

$$R_s = \frac{\sum\{(\sum Agent\,\mathrm{Re}liabilities / No.ofAgents)\}}{No.ofAgentGroups} \tag{4}$$

Here the individual agent reliabilities are calculated as stated earlier. To estimate R_s in equation 3 an algorithm is proposed in this paper.

3.3 Estimating Reliability of the Proposed Network

The initial position of the MNs is obtained first. Once the nodes are assigned a node reliability value, the connection between the operating MNs are calculated (as in [8]) thus forming the topology of the network. Network connectivity is determined from that structure at a given instant of time. Then using SRMM the movement is calculated for the next time instant. These steps are iterated to simulate each of the time increments. With each run, the indicator variable q is updated until a pre-specified number of runs (Q) have been exhausted (dictated by desired accuracy determined by variance). Finally, the performance metrics are calculated. Here follows the detailed steps.

1) Input Parameters. **k (no. of independent mobile agents in the system)**

2) Detailed Steps:

1. Initialize n (that is the no. of MNs visited by an agent) to 0

2. According to Weibul distribution we find individual software reliability of the agents r_i.
3. Nodes move according to SRMM and received signal power (P_r) (according to Two ray propagation model) and hence node connectivity is calculated as in [8].
4. The D_{max} for the different agent groups is simulated according to uniform distribution.
5. Breadth First Search (BFS) is used iteratively to find all the connected components unless all connected subgraphs are assigned a proper cluster id as in [13].
6. The agents will perform their job on this modified graph.
   - 6.1. If an agent has failed to migrate to a previously chosen destination then it either retries or goes to step 6.2 depending on equation 2.
     - 6.1.1. If migration attempt is unsuccessful then increment r by 1.
     - 6.1.2. If successful then increment n_i by 1 and reset r.
   - 6.2. An agent may randomly choose a yet unvisited node to be its next destination
     - 6.2.1. If that destination falls in the same cluster as it is now residing, then try to migrate to that node with a certain probability.
     - 6.2.2. If migration is successful then increment n_i by 1 and reset corresponding r.
7. Repeat steps 6 for all agents (k) in the system.
8. Repeat steps 2 to 7 until all nodes are visited or time out occurs.
9. Calculate

$$\lambda_i(t) = \frac{n}{N} \tag{5}$$

10. Reset the value of n as in [8].
11. Calculate [8,13]

$$\lambda(t) = \frac{1}{k}\sum_{i=1}^{k} \lambda_i(t) r_i \tag{6}$$

12. Repeat steps 3 to 12 Q (simulation steps) times.
13. Calculate node reliability

$$\frac{1}{Q}\sum_{q=1}^{Q} \lambda(q,t) \tag{7}$$

In step 6.2.1 when we say that an agent tries to migrate from (say) node A to B, we mean that the agent sends its replica to B. If the replica works successfully at B, then a reply would come and the copy at A would kill itself. But if a timeout occurs, the agent will once again send another replica and the process goes on unless r_{max} is reached. Normal distribution is used to replicate this fact. That is whenever the next destination is found to be reachable, the agent tries to migrate according to a Normal distributed variable. Normal distribution is generally used to represent the scenario in which random variables tend to cluster around a mean. In this case the mean represents a situation where the agent will make a transition successfully that is there is no transient error. The standard deviation (SD) links to the number of retries. Thus

higher the number of transient errors, greater will be the SD, more will be the number of necessary retries for a successful agent migration.

3.4 An Example

We have taken an instance where there are six nodes in the network. Six agents from six different owners start their journey from nodes MN_1, MN_2, MN_3, MN_4, MN_5 and MN_6 respectively and roam around the network to accomplish its task. Each agent represents a group. The groups are sorted in such a manner that group 1 has the tightest delay requirement and group i+1 can tolerate longer delay than groups 1 to i. The weight α is taken to be 0.5. The nodes are taken close enough so that they form a connected network. Every 3 seconds the positions of the nodes are updated according to SRMM. The simulation is carried out for 33 seconds. For clarity we have shown migrations of 4 agents only for 11 successive time instants in table I. In this scenario the agents tried to move to their next destination (selected randomly). Initially the network graph is found to be connected and all agents make their migrations successfully. Since delay bound was about to expire, agent 1 retracts after time instant t+5Δt. Agent 2 finishes its task at the same time so it reports back to its owner. But time instant t+7Δt onwards the graph becomes so partitioned that all the unvisited nodes (for the agents 3, 4) becomes unreachable. At time instant t+2Δt agent 1 wanted to visit MN_2but due to transient error the migration was not successful. So, depending on equation 2, agent 1 waits till t+4Δt. Here we have assumed that an agent needs to visit all the nodes in the network in order to accomplish its task (as in service discovery). The mean of Normal distribution is taken to be 4 and the value of SD is taken to be 1. After simulation the reliability comes out to be 0.75.

Table 1. Agent Migrations at 11 successive time instants

Agent ids	Time t	Time t+Δt	Time t+2Δt	Time t+3Δt	Time t+4Δt	Time t+5Δt	Time t+6Δt	Time t+7Δt	Time t+8Δt	Time t+9Δt	Time t+10Δt
Agent1	MN_1	MN_5	MN_5	MN_5	MN_5	MN_6	Reports back to the owner				
Agent2	MN_2	MN_1	MN_3	MN_4	MN_6	MN_5	Reports back to the owner				
Agent3	MN_3	MN_5	MN_6	MN_6	MN_6	MN_6	MN_6	MN_6	MN_6	MN_6	MN_6
Agent4	MN_4	MN_5	MN_6	MN_1	MN_1	MN_3	MN_3	MN_3	MN_3	MN_3	MN_3

4 Experimental Results

The simulation is carried out in java and can run in any platform. The initial positions of agents are given. All agents of the same group start from the same node, that node is designated to be the owner. The simulation time is taken to be 120 minutes. For the rest of the experiments, the N is taken to be 40 unless stated otherwise. The variation between delays bounds of group i and group i+1 is 15 seconds unless stated otherwise. The other parameters like mean of Normal distribution are kept the same as mentioned in the example (section 3.4). We have seen that if node movements are allowed only at the beginning before the mobile agents start their task route then

performance of the algorithm does not vary appreciably with no. of agents deployed in the system. But here we investigate how the performance of the MAS varies with no. of agent groups. In this experiment first we take a single group of a given no. of agents (say 20). Then keeping the total count same (for example 20) we increase the no. of groups. For example a single group of M agents is divided into G groups with (M/G) agents in each group. An agent group signifies a particular application. So with increasing no. of agent groups our MAS will be more heterogeneous. As figure 1 shows, for a given no. of agents as the heterogeneity increases among them the overall reliability of the system shows a slow downward trend and eventually stabilizes. So, for greater heterogeneity with smaller sized groups, reliability does not fall sharply rather reaches a steady state. In the next experiment, for a group of agents, we increased the tolerance limit of maximum delay (D_{max})per group-the lower bound to delay tolerance. That is if variation =15 seconds (figure 2) then D_{max}value of group (i+1) is 15 seconds more than that of group i, 30 seconds more than group (i-1) and so on. Thus

$$D_{max}^{group_i} = D_{max}^{group_{i-1}} + \text{Variation} \tag{8}$$

Thus group 1 represents the tightest delay bound hence may simulate real time traffic and for group 6 the delay bound is quite relaxed indicating non real time traffic. When the overall lower bound of delay tolerance is lower, overall reliability improves significantly for higher delay variation. But for larger delay bounds, increasing variation between the groups does not affect overall system reliability much. Rather all the groups manage to perform quite reliably. It can also be observed that for long running applications the system reliability gradually drops as is the case. Because with time the node failure probability increases as well as that the agent software also has a probability of wear and tear with time. But as figure 3 shows the rate of decrease lowers with increasing time which indicates stability of the system. Thus despite initial sudden fall in the curve (may be due to network partitioning) the system stabilizes gradually. Here 6 agent groups are considered each having 5 agents and their delay requirement gradually increases with variation =15 seconds. Thus group 1 represents perfectly real time traffic and group 6 can tolerate the highest amount of delay among the agent groups.

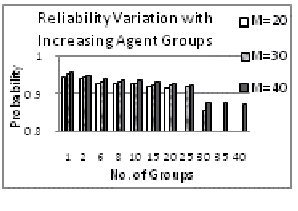


Fig. 1. Reliability variation with no. limit for migration delay

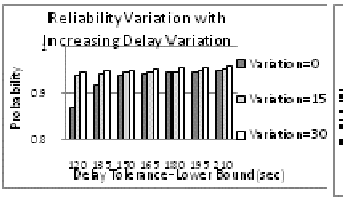


Fig. 2. Reliability variations with increasing reliability of agent groups

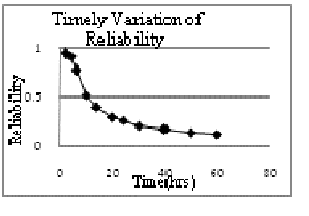


Fig. 3. Timely variation of

Not only the inherent environmental noise is taken into account but we have also considered the effect of transient environmental characteristics by associating a

probability to agent migration even when there is a path between the current agent site and its next probable destination. As mentioned in step 7 of the algorithm in section 3.2.4, a Normal distributed variable represents the ease with which an agent migrates. In our simulation, lower values of mean represents noisy environments, and relatively lower interfering environments are represented by higher mean values. Our experimental results show that for highly stable environments, the reliability of MAS is not much dependent on the nature of the environment as the system will eventually learn to adjust with the background noise.But as the transient environmental error increases, a sharp decrease in reliability can be observed for inherently noisy environments (figure 4). But MAS in a less noisy environment (represented by higher values of mean) would be able to tackle sudden rise in background noise (represented by high standard deviation).

Now we show comparison of MAS performance w.r.t QoS parameters of demanded capacity [8] and delay as MANET gets bigger. The capacity demanded by the agents is Normal distributed with mean=3.16kbps which is the minimum link capacity that a link should support for agent migration in our experiments. The minimum delay bound is kept at N*avgDelay. The results show that for smaller network with higher capacity, agents with demanded capacity perform better as they do not have a delay bound. But as the network gets wider resulting in a drop of average link capacity, the hosts could not fulfill capacity demands of agents thus making them less reliable. But as an agent in a bigger MANET (with longer trails to visit) is designed to tolerate longer delay (proportional to N), graph in figure 5 shows a slight downward trend that ultimately reaches a steady state.

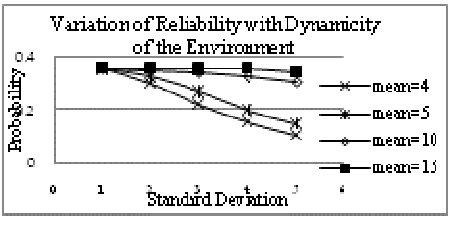


Fig. 4 Variation of reliability with changing environment characteristics

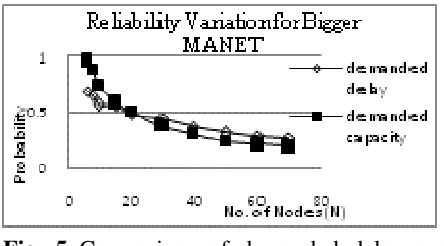


Fig. 5 Comparison of demanded delay and capacity [8] for bigger MANET

5 Conclusion

In this a paper, a scalable approach to estimate the reliability of MAS with QoS for MANET is presented. The reliability calculation depends heavily on conditions like QoS requirements of the agents, transient characteristics of the network and of course node mobility. The QoS requirement (maximum tolerated delay) is considered for agents pertaining to different applications and is shown that reliability increases as the QoS is relaxed (higher value of delay tolerance). The transient characteristics of failure are also addressed in this work as the agents migrate with a given probability and if fails, it will retry. The maximum number of retries is found to be (as in

equation (2)) a function of the amount of delay that the application can tolerate and the progress of the agents. The results show that the agents with delay tolerance can handle sudden noise bursts (figure 4). Moreover, a comparison is given between the agents demanding for capacity [8] and delay(present work). The curve in figure 5 clearly indicates that agents with delay tolerance can perform better in bigger networks. We may also study the effect of QoS parameters where mobile agents are designed specifically for an application.

References

[1] Cao, J., Feng, X., Lu, J., Das, S.K.: Mailbox-Based Scheme for Designing Mobile Agent Communications. Computer 35(9), 54–60 (2002)
[2] Migas, N., Buchanan, W.J., McArtney, K.: Migration of mobile agents in ad-hoc, Wireless Networks. In: Proc. 11th IEEE International Conference and Workshop on the Engineering of Computer-Based Systems, pp. 530–535 (2004)
[3] Meier, R.T., Dunkel, J., Kakuda, Y., Ohta, T.: Mobile agents for service discovery in ad hoc networks. In: Proc. 22nd International Conference on Advanced Information Networking and Applications, pp. 114–121 (2008)
[4] Laamanen, R.H., Alonko, T., Raatikainen, K.: Dependability issues in mobile distributed system. In: Proc. of the Pacific Rim International Symposium on Dependable Computing, pp. 7–14 (1999)
[5] Murphy, A.L., Picco, G.P.: Reliable communication for highly mobile agents. Autonomous Agents and Multi-Agent Systems 5(1), 81–100 (2002)
[6] Urra, O., Ilarri, S., Mena, E.: Agents Jumping in the Air: Dream or Reality? In: Cabestany, J., Sandoval, F., Prieto, A., Corchado, J.M. (eds.) IWANN 2009. LNCS, vol. 5517, pp. 627–634. Springer, Heidelberg (2009)
[7] Ilarri, S., Trillo, R., Mena, E.: SPRINGS: A scalable platform for highly mobileagents in distributed computing environments. In: 4th International WoWMoM 2006 Workshop on Mobile Distributed Computing (MDC 2006), pp. 633–637 (2006)
[8] Chowdhury, C., Neogy, S.: Reliability Estimate of Mobile Agent Based System for QoS MANET Applications. In: The Annual Reliability and Availability Symposium, pp. 1–6 (2011)
[9] Cook, J.L., Ramirez-Marquez, J.E.: Two-terminal reliability analyses for a mobile ad-hoc wireless network. Reliability Engineering and System Safety 92(6), 821–829 (2007)
[10] Cook, J.L., Ramirez-Marquez, J.E.: Mobility and reliability modeling for a mobile ad-hoc network. IIE Transactions, 1545-8830 41(1), 23–31 (2009)
[11] Daoud, M., Mahmoud, Q.H.: Reliability estimation of mobile agent systems using the Monte Carlo approach. In: Proc. 19th IEEE AINA Workshop on Information Networking and Applications, pp. 185–188 (2005)
[12] Daoud, M., Mahmoud, Q.H.: Monte Carlo simulation-based algorithms for estimating the reliability of mobile agent-based systems. Journal of Network and Computer Applications, 19–31 (2008)
[13] Chowdhury, C., Neogy, S.: Estimating Reliability of Mobile Agent System for Mobile Ad hoc Networks. In: Proc. 3rd International Conference on Dependability, pp. 45–50 (2010)

Network Event Correlation and Semantic Reasoning for Federated Networks Protection System

Michał Choraś[1,2] and Rafał Kozik[1,2]

[1] ITTI Ltd., Poznań, Poland
mchoras@itti.com.pl
[2] Institute of Telecommunications, UT and LS Bydgoszcz, Poland

Abstract. In this paper we present semantic approach to network event correlation for large-scale federated intrusion detection system. The major contributions of this paper are: network event correlation mechanism and semantic reasoning based on the ontology. Our propositions and deployments are used in Federated Networks Protection System as a part of the Decision Module.

1 Introduction

Nowadays, especially after successful cyber attacks on Estonia, Georgia, Iran and on companies like Google and Sony, cyber attacks are considered a major threat for critical infrastructures (e.g. power grids) and homeland security (e.g. financing system)[1]. For example, in 2008 successful DDoS (Distributed Denial of Service) attacks were targeted at Georgian government sites, Georgian president site and servers of National Bank of Georgia [1].

Cyber attacks are also considered a threat for military networks and public administration computer systems. The goal of the Federated Networks Protection System developed in the SOPAS project is to protect public administration and military networks which are often connected into a Federations of Systems (FoS). While adpoting the concept of federation of networks, the synergy effect for security can be achieved.

In our approach, we use the capability of the federated networks and systems to share and exchange information about events in the network, detected attacks and proposed countermeasures. Such approach has recently gotten much attention and may replace inefficient approach of "closed security" [2]. The concept of federated networks and systems has gained much attention also in the context of critical systems, military networks and NNEC [3] [4] [5].

Moreover, most currently used IDS (Intrusion Detection System) and IPS (Intrusion Prevention System) systems have problems in recognizing new attacks (0-day exploits) since they are based on the signature-based approach. In such mode, when system does not have an attack signature in database, such attack is not detected. Another drawback of current IDS systems is that the used parameters and features do not contain all the necessary information about traffic

N. Chaki and A. Cortesi (Eds.): CISIM 2011, CCIS 245, pp. 48–54, 2011.

and events in the network [6]. Therefore, in our approach, we propose to use various network sensors and honeypots to analyze diverse network events and correlate them to detect cyber attacks.

The major contribution of this paper is the network event correlation mechanism and semantic reasoning proposition based on the ontology.

This paper is structured as follows: in Section 2 the general architecture of the Federated Networks Protection System (FNPS) is overviewed. In Section 3 network events correlation mechanism is described. In Section 4 ontology-based reasoning is described in detail. Conclusions are given thereafter.

2 Architecture of Federated Networks Protection System

The general architecture of the Federated Networks Protection System (FNPS) is presented in Fig. 1.

It consists of several interconnected domains, which exchange information in order to increase their security level and the security of the whole federation.

Different subnetworks are arranged in domains, according to the purpose they serve (e.g. WWW, FTP or SQL servers) or according to their logical proximity (two networks closely cooperating with each other). In each of the domains, a Decision Module (FNPS-DM) is deployed. Each DM is responsible for acquiring and processing network events coming from sensors distributed over the domain.

If the attack or its symptoms are detected in one domain, the relevant information are disseminated to other cooperating domains so that appropriate countermeasures can be applied.

In the proposed approach (Fig. 1), the basic idea is to use the already available and existing sensors deployed in particular domains, such as:

- Application layer sensors (e.g. SCALP[7], PHPIDS [8], SEC[9]),
- IDS and IPS systems (e.g. we will use SNORT [10]),
- Anomaly Detection Systems [11],
- ARAKIS system [12],
- HoneySpider Network (HSN) system [13].

Currently, IDS or IPS systems are installed in networks as typical means of cyber defense. In FNPS we will use SNORT system, also with additional FNPS preprocessors.

Anomaly Detection Systems rely on the existence of a reliable characterization of what is normal and what is not, in a particular networking scenario. More precisely, anomaly detection techniques base their evaluations on a model of what is normal, and classify as anomalous all the events that fall outside such a model [6,11].

ARAKIS and HSN are existing commercial systems developed by NASK (SOPAS project partner). These systems, if available in the network, will be used as sensors in FNPS.

However, there is no need to install these systems in each domain. The important idea is to use and share important information from the domain in which ARAKIS and/or HSN are installed with other members of the federation.

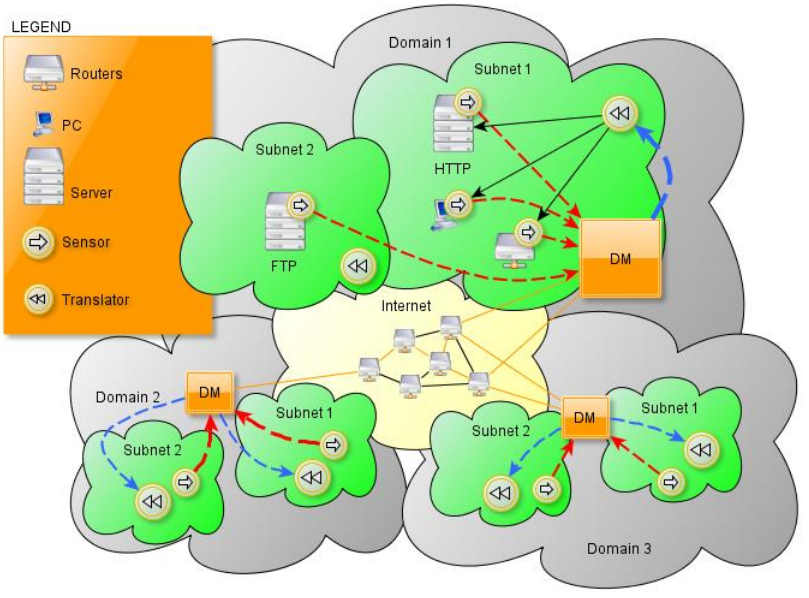


Fig. 1. General Architecture of FNPS

Each Decision Module can react to network events and attacks by sending information to the Translator element. The output information from DMs is the General Reaction Rule describing attack symptoms (information about network events) and particular reaction rule to be applied by reaction elements. Translator has the knowledge about its subnet capabilities and can access the necessary reaction element (e.g. firewalls, filters or IDS). Reaction elements can be reconfigured by Translator in order to apply commands sent by the Decision Module.

All Decision Modules within the federation can also interact with each other and exchange security information. Particularly information about network incidents, like attack in one domain, may be sent to different Decision Modules in order to block the attacker before the consequent attack takes place on another domain. The communication between decision modules is P2P-based (Gnutella algorithm) in order to increase communication resiliency and enable data replication. The communication channels are encrypted using the SSL algorithm. This allows to protect the communication against the packet sniffing (by third persons). Additionally payload is encrypted with different key and it can be only decrypted by domains that belongs to the same distribution group (nodes relaying the message can not read the payload).

In the following sections we will focus on network events correlation and ontology-based reasoning omitting other aspects and building blocks of FNPS.

3 Network Events Correlation Mechanism

The event correlation process is applied in Decision Module. Events are generated by sensor distributed over the domain and sent to Decision Module. The Decision Module consists of the following components (see Fig.2):

- Correlation Engine (e.g. based on the Borealis system),
- CLIPS rule engine,
- Ontology (in OWL format),
- Graphical User Interface.

The events coming from the sensors are received by Borealis system (distributed stream processing engine) in form of streams. Each stream is built of multiple tuples (events). Each tuple, depending on sensor type, may have different schema. Borealis allows to process streams in order to correlate information coming from different sources and to detect network incidents more efficiently. The query that is executed over the multiple streams consists of operators. There are different kinds of operators provided by the Borealis engine that allow for aggregation, filtering and joining data coming from different streams.

Borealis provides mechanisms that allows the Decision Module to efficiently execute multiple queries over the data streams in order to perform event correlation. Particularly, Borealis allows the DM to efficiently aggregate the same kind of events coming from multiply sources generated by particular IP address. The result of a correlation process is an intermediate event that is further processed by CLIPS rule engine.

CLIPS uses the ontology that describes broad range of network security aspects. CLIPS engine identifies whenever some attacks or malicious network behaviors have been discovered. The information describing the network incident

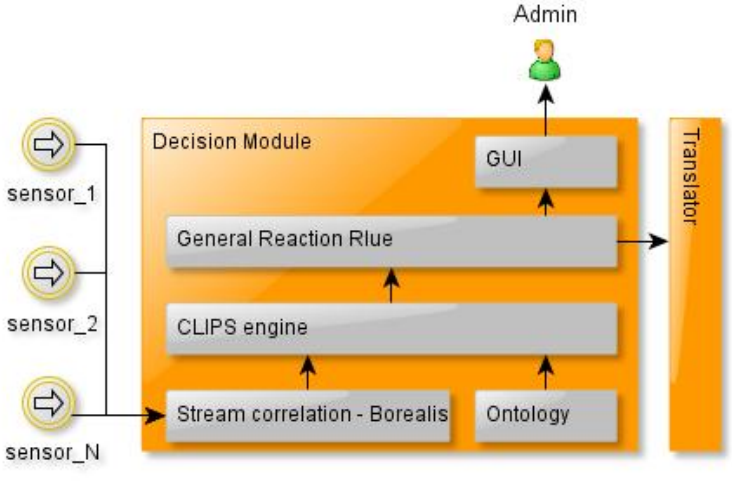


Fig. 2. Information flow diagram

and reconfiguration procedures are sent to Translator (see Fig. 2). Moreover, detailed information in human readable format are generated and visualized to network administrator via FNPS-DM GUI.

4 Ontology-Based Reasoning

In the proposed Federated Networks Protection System, we use our network security domain ontology created in the SOPAS project. The knowledge about the security aspects such as threats, attacks, reactions, policies etc. is modeled and formalized in the OWL format [15]. Moreover, semantic rules are developed in SWRL language [16]. The main classes and relations of the proposed ontology are shown in Fig. 3.

Each intermediate event received by CLIPS rule engine is considered as attack symptom and as such is matched with knowledge obtained from ontology in order infer the most probably attack. In other words estimate the probability $p(A|o_1, o_2, \ldots, o_n)$ of particular attack (A) given the observations $(o_1, o_2, \ldots, o_n)$. Using the Bayesian theory the probability can be rewritten as 1.

$$p(A|o_1, o_2, \ldots, o_n) = \frac{p(o_1, o_2, \ldots, o_n|A)p(A)}{p(o_1, o_2, \ldots, o_n)} \tag{1}$$

For all known attacks and known symptoms the problem of finding the most probable attack in practice become a MAP (Maximum A-Posteriori) problem (2).

$$\mathrm{A}^* = \arg\max_A p(A|o_1, o_2, \ldots, o_n) \tag{2}$$

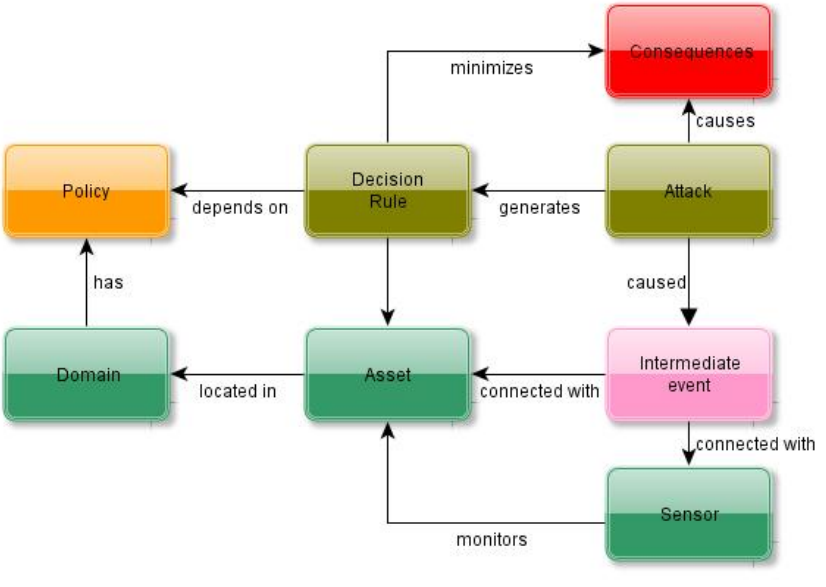


Fig. 3. Main classes in the SOPAS ontology

Assuming observations independence, the MAP problem is equal to finding maximum joint probability $p(A, o_1, o_2, \ldots, o_n)$. Using the Bayes theory the joint probability may be rewritten as 4.

$$p(A, o_1, o_2, \ldots, o_n) = p(o_1, o_2, \ldots, o_n|A)p(A) = p(A)p(o_1|A, o_2, \ldots, o_n) \quad (3)$$

$$p(o_2|A, o_1, \ldots, o_n)(o_n|A, o_1, o_2, \ldots, o_{n-1}) \quad (4)$$

Because observations are mutually independent ($p(o_1|A, o_2, \ldots, o_n) = p(o_1|A)$) the equation 4 is reduced to equation 5.

$$p(A, o_1, o_2, \ldots, o_n) = p(A) \prod_{i=1}^{N} p(o_i|A) \quad (5)$$

The event-observation matching process uses information retrieved from the ontology indicating which event is semantically equal (or is a subclass) of particular observation.

If the event is successfully matched with observation, all possible (matching) attacks are analyzed (the score function from 5 is computed). If the score exceeds the predefined threshold then information about reaction is retrieved from the ontology and General Decision Rule is built (and sent to appropriate Translator and other domains in the federation).

5 Conclusions

In this paper we presented the concept of Federated Networks Protection System that is being developed in the SOPAS national research project. In particular, we focused on the Decision Module and the correlation approach to detect cyber attacks. Moreover, we described ontology-based reasoning approach used to make decisions about the attacks.

The presented system is dedicated for federated networks and systems used by the public administration and military sector. Such systems can increase their overall security and resiliency by sharing and exchanging security related information and General Decision Rules.

The proposed system can for example: correlate various network events from different layers (traffic observations and application logs analysis) in order to detect Injection attacks (e.g. SQLIA) on the Ministry web service (e.g. based on Joomla). Then, the Decision Module will create reaction rules and send it to DM in another domain in federation of systems. Therefore, the same Injection attack targeted at the other network will be blocked.

Acknowledgement. This work was partially supported by Polish Ministry of Science and Higher Education funds allocated for the years 2010-2012 (Research Project number OR00012511).

References

1. Enabling and managing end-to-end resilience, ENISA (European Network and Information Security Agency) Report (January 2011)
2. Choraś, M., D'Antonio, S., Kozik, R., Holubowicz, W.: INTERSECTION Approach to Vulnerability Handling. In: Proc. of 6th International Conference on Web Information Systems and Technologies, WEBIST 2010, vol. 1, pp. 171–174. INSTICC Press, Valencia (2010)
3. NATO Network Enabled Feasibility Study Volume II: Detailed Report Covering a Strategy and Roadmap for Realizing an NNEC Networking and Information Infrastructure (NII), version 2.0
4. El-Damhougy, Yousefizadeh, H., Lofquist, H., Sackman, D., Crowley, R.: Hierarchical and federated network management for tactical environments. In: Proc. of IEEE Military Communications Conference MILCOM, vol. 4, pp. 2062–2067 (2005)
5. Calo, S., Wood, D., Zerfos, P., Vyvyan, D., Dantressangle, P., Bent, G.: Technologies for Federation and Interoperation of Coalition Networks. In: Proc. of 12th International Conference on Information Fusion, Seattle (2009)
6. Coppolino, L., D'Antonio, L., Esposito, M., Romano, L.: Exploiting diversity and correlation to improve the performance of intrusion detection systems. In: Proc. of IFIP/IEEE International Conference on Network and Service (2009)
7. SCALP project homepage, http://code.google.com/p/apache-scalp/
8. PHPIDS project homepage, http://code.google.com/p/phpids/
9. SEC project homepage, http://simple-evcorr.sourceforge.net/
10. SNORT project homepage, http://www.snort.org/
11. Choraś, M., Saganowski, L., Renk, R., Holubowicz, W.: Statistical and signal-based network traffic recognition for anomaly detection, Expert Systems (Early View) (2011) doi: 10.1111/j.1468-0394.2010.00576.x
12. ARAKIS project homepage, http://www.arakis.pl
13. HSN project homepage, http://www.honeyspider.net/
14. Neches, R., Fikes, R., Finin, T., Gruber, T., Patil, R., Senator, T., Swartout, W.R.: Enabling Technology for Knowledge Sharing. AI Magazine 12(3), s.36–s.56 (1991)
15. OWL Web Ontology Language Semantics and Abstract Syntax (June 2006), http://www.w3.org/TR/owl-features/
16. SWRL: A Semantic Web Rule Language Combning OWL and RuleML, W3C Member Submission, http://www.w3.org/Submission/SWRL/

VAMI – A Novel Architecture for Vehicular Ambient Intelligent System

Anupam Saha and Rituparna Chaki

Computer Science and Engineering Department
West Bengal University of Technology
West Bengal, India
{it.anupam.saha,rituchaki}@gmail.com

Abstract. Vehicular communication involves many decision making situations. The existing models for packet routing in VANET fails in cases needing fast decision making. Ambient Intelligence (AmI) system is a relatively new technique which aim at augmenting real environments where users are provided with pervasive virtual services. The ability to learn very quickly from the ambience of an object, and transforming that information into specific response involves a heterogeneous network comprising of vehicles, humans, and road side units. This paper proposes a modular architecture for information sensing, communicating, updation of knowledge-bas, etc, aimed at generating alarms in emergency situations. The emergency situation might involve abnormal physical behaviour, road blockage due to accidents or diversions, etc.

Keywords: Ambient Intelligence, VANET, RSE, Core - conscious model, VAMI.

1 Introduction

Modern society is getting increasingly dependent on machine intervention for a smoother everyday life. In our attempt to gather information from the surroundings, we have arrived at an ambient intelligent world. Such a world has heterogeneous devices working in together to support people in carrying out their everyday activities, tasks and rituals, in an easy and natural manner. The ambient intelligence paradigm has evolved through the steps such as pervasive computing, ubiquitous computing, profiling practices, and human-centric computer interaction design. The three main characteristics of ambient intelligence system are awareness, intelligence and adaptable. Awareness means ability of the system to locate and recognize objects and people, their location and need. Intelligence allows the system to analyze the context learn their behavior and eventually recognize as well as show emotions. Adaptable defined about the environment and the people within it in order to optimize their own behavior. Ambient Intelligence is applicable in many areas such as healthcare [8], smart home, emergency service [7], production oriented service. In emergency services the system provides long term environmental monitoring, alarm event detection and propagation, localization and tracking of objects, assisted navigation as well as fast data dissemination service to be used.

N. Chaki and A. Cortesi (Eds.): CISIM 2011, CCIS 245, pp. 55–64, 2011.

This paper concentrates on designing an ambient intelligence environment for vehicle movement. The proposed system aims at generating alarms depending on a variety of deviations from normal state, such as road condition, road blockage, driver's medical state, etc. There are several modules working in conjugation to achieve the desired level of response generation. The environment itself behaves like an intelligent one which helps the vehicle to respond according to the situation. The underlying mode of communication involves VANET, a form of Mobile ad-hoc network. This helps in providing communications among nearby vehicles and also between passing vehicles and the roadside units. The use of VANET makes communication fast and secure.

Rest of the paper is organized as follows: Section2 of this paper discusses about the state of the art. Section 3 discusses about the proposed methodology. Section4 of this paper discusses about some of the case study work. Section 5 concludes the paper. Section 6 consists of the reference list.

2 Review Work

VANETs in general have a highly dynamic nature [1]; however the vehicles in the network usually follow a predictable pattern. The primary aim of broadcasting in vehicular ad hoc networks is to avoid accident. An AMI system should be able to handle a number of driving conditions that are likely to occur. The normal road conditions i.e., one with normal speed of vehicles via-a-via an abnormal one in case of an accident has to be efficiently handled. The emergency alarm generation is part of the AMI system.

In VANET three types of communication exists; periodic broadcasts, emergency warnings and private application messages. Initially, the periodic broadcasts will be transmitted approximately every 100ms by each vehicle. The periodic broadcast are used to exchange the state of a vehicle to the surrounding vehicles (i.e. location, direction, velocity, etc.), so that the vehicles can passively detect dangerous road conditions. Emergency warnings are exchanged only when a dangerous condition arises. Emergency warnings are used to actively warn other vehicle of an abnormal condition such as an accident on the highway. When an accident takes place warning messages should be disseminated as quickly as possible to all surrounding vehicles. Last, VANETs will be used to send private unicast messages, which will be sent either vehicle-to-vehicle or vehicle-to-roadside. Unicast messages will be used for these value added services such as electronic toll collection.

A number of authors have addressed the problem of sending broadcast messages in MANETs and VANETs. In [1] the authors address the broadcast storm problem in MANETs, with five multi-hop relaying strategies: a probabilistic-based scheme, a counter-based scheme, a distance-based scheme, a location-based-scheme, and a cluster-based scheme. The location-based scheme is performed the best because the scheme eliminated the most redundant broadcast and performed well networks. In [4] the paper shows that the probability of reception of a broadcast message decreases respectively with the distance from the sender increases. Hidden terminal problem is the main reason behind this. A priority access mechanism is implemented to improve the reception rate of broadcast messages, but still fails to achieve reliability anywhere near 100%. In all likelihood, it may be unrealistic to expect every node in an 802.11

based network to successfully receive a broadcast because of the hidden terminal problem. A single-hop broadcast protocol [2] is proposed to increases the probability of a message's reception by sending the message multiple times. But the main problem with this scheme is that it will not scale well when used for multi-hop relaying. The authors of [3] propose the VCWC protocol to transmit emergency warning messages (EWM); this is based on a state machine and a multiplicative rate decrease algorithm. .Other solutions aim at assigning time slots to nodes for them to transmit during, these solution are likely inapplicable to VANETs because it requires the synchronization of nodes which is hard to achieve because of the high mobility of nodes which requires many of the algorithms that need to maintain sets or clusters may not perform well in VANETs because of the high mobility and the large amount of overhead that is necessary to maintain the sets.

[7] gives a detailed technical overview of some of the activities carried out in the context of the "Wireless Sensor networks for city wide Ambient Intelligence" project. The main aim of the project is to demonstrate the feasibility of large scale wireless sensor network deployments. The paper tells about the applications which include long term environmental monitoring, alarm event detection and propagation, single sensor interrogation, localization and tracking of objects, assisted navigation as well as fast data dissemination service to be used.

The overall objective of the WISE-WAI project is precisely to exploit the potential of WSNs by designing and evaluating system architecture for a flexible, heterogeneous and energy efficient network, including the specification of the applications. The project efforts are spent at several design levels, and w driven by the applications the network which require to support. In the context of this project , three classes of applications are of specific interest. Location Based application are among the first and most popular application of WSNs. Environmental monitoring applications are also of primary importance, for ameliorating the quality of life through the definition of comfort criteria and computation of related indices, to increase energy awareness and improve the efficient use of energy resources. Alarm event detection is also of interest, especially in working or living conditions that can be subject to hazardous events.

The network structure required to support these traffic requirements must therefore be very flexible and capable to adapt varying network conditions. The applications supports the communication requirements also for instance, localization and tracking will be performed by communicating to a subset of the sensors available in the proximity of the object to localize, while environmental monitoring and alarm event detection include specific functionalities that reduce the amount of transferred data through aggregation.

In [8] discusses about the body sensor network. It is one of the applications of ambient intelligence. In this paper two separate approaches are said about the health related application of ambient intelligence. In both cases patient's biosignal's are measured by wearable sensor device and communicate wirelessly with a handheld device. Alarms and biosignals can be transmitted over wireless communication link and a remote health professional can view biological data by web application. Two approaches are similar in many respects. The main architecture of the body sensor network is following. The wearable sensors are present in the patient's body and it is connected with mobile base unit called as mbu. The MBU is connected with the proper hospitals or authority using wireless or wired service. One system is called as

Mobihealth and other as PHM. Mobihealth focused on physician care and remote processing whereas PHM focused on patient self care and local processing.

The wanted inspiration can be found in the work of Antonio Damasio where a model[6] for brain and conscious reasoning is reported and motivated on neuro - physiological bases. We are using this model to design a high level complex architecture of AmI for VANET. The model said that the system of this kind need to be aware of the behavior of all objects acting in their scope of sensing as well as of its own components' status and reactions. In this sense awareness of the context in which the system act is not enough, the system has also in some way to be conscious of its own internal state.

3 Proposed Methodology

3.1 VAMI Architecture

After reviewing the current state of affairs, it is observed that AMI system has tremendous potential in emergency response generation as the knowledge base is continuously enriched from the ambience of an object. Here we propose an AMI system for vehicles, VAMI (Vehicular Ambient Intelligence system). The proposed VAMI model aims to generate alarms in case of road blockages, accidents, and also monitors the medical condition of each driver, so as to warn the medical helpline about any abnormal medical symptoms. Figure 1 shows the block diagram of the architecture of VAMI.

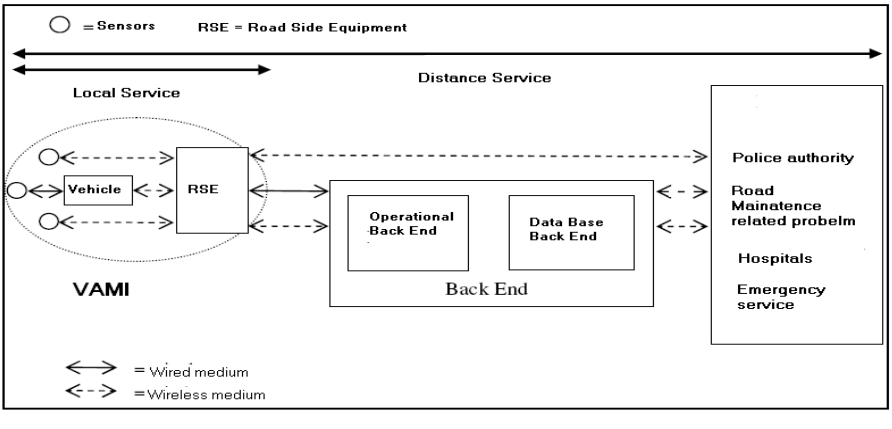


Fig. 1. (a). Generic architecture of proposed system

According to figure 1(b) the main job of Resource manager is to handle resource request of all the subsystems. Sensors and measurement instruments collect and report the values of predetermined quality parameters. A system manager has to coordinate some important activities, such as (i) monitoring quality measurements received from sensing and measurement instruments; (ii) analyzing the measured quality values in accordance with standards, procedures and regulations; (iii) planning actions/responses

using the analysis results, standards, procedures, and regulations; and (iv) executing action plans.

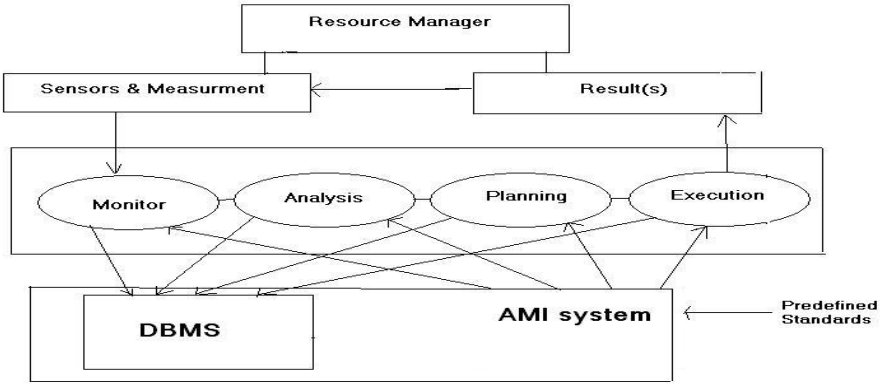


Fig. 1. (b) Detailed architecture of proposed system

The task of knowledge manager is to manage the quality knowledge (i.e. standards, procedures and regulations) and report to the system manager. An information manager manages the quality information generated by various subsystems, and makes the quality information available to system manager.

3.2 Sensing and Measurement Subsystem

The Sensing and Measurement Subsystem contains the following parts,

1. Sensors and measurement instruments;
2. S/M proxy Sub phase; and
3. S/M reporting phase.

Sensors and measurement instruments determine the actual values of the quality parameters. Every sensor (or measurement instrument) is represented by one S/M proxy phase. S/M reporting phase handle requests intended for sensors (or measurement instruments) and manage their interactions of with the rest of the software (e.g. obtain the measured quality values from S/M proxy module and report them to monitoring information receiving module).

3.3 Database Management Subsystem

Though Database management subsystem is a part of AMI system we can call this as Information Management System because this system actually stores the information which will be used as results in some future generated query. This system includes the following sets of modules:

1. Quality Information Database Management System (DBMS);
2. Information proxy sub phase; and
3. Information management sub phase.

3.4 Monitoring Subsystem

Monitor subsystem is discussed in the following,

1. Monitoring information receiving (MIR) Sub Phase;
2. Monitoring knowledge receiving (MKR) Sub Phase;
3. Monitoring module; and
4. Monitoring reporting (MR) Sub Phase.

MIR obtains measured quality values through S/M reporting module, and provide them to monitoring sub phase. MKR obtains quality monitoring standards, procedures and regulations through knowledge reporting module, and provide them to monitoring sub phase. Monitoring sub phase use the received quality monitoring standards, procedures and regulations to vigilantly monitor the system to ensure the measured quality values reported to the monitoring subsystem are valid (e.g. not corrupted), and report the verified quality information to MR. MR packages the verified quality information in a proper form and report it to information management phase as well as analysis information receiving sub phase.

3.5 Analysis Subsystem

The Analysis Subsystem contains the following parts,

1. Analysis information receiving (AIR) Sub Phase;
2. Analysis knowledge receiving (AKR) Sub Phase;
3. Analysis Sub Phase; and
4. Analysis reporting Sub Phase.

AIR obtains quality information from monitoring reporting phase, and provides them to analysis phase. AKR obtains the quality analysis standards, procedures and regulations from knowledge reporting phase, and provide them to analysis phase. Analysis phase analyzes the received quality information in accordance to the received quality analysis standards, procedures and regulations; and report analysis results (including the level and severity of the risks associated with the quality problem) to analysis reporting phase. Analysis reporting phase packages the analysis results in a proper form and report them to information management module as well as planning information receiving sub phase.

3.6 Planning Subsystem

1. Planning information receiving (PIR) Sub Phase;
2. Planning knowledge receiving (PKR) Sub Phase;
3. Planning Sub phase; and
4. Action plan reporting (APR) Sub phase.

PIR obtains quality planning information through analysis reporting phase, and provides them to planning sub phase. PKR obtains quality planning standards, procedures and regulations from knowledge reporting sub phase, and provides them to planning Sub phase. Planning Sub phase makes quality plans using the received quality planning information, standards, procedures and regulations, and send action plans to APR. APR packages actions plans in a proper form and report them to information management sub phase as well as action plan receiving sub phase.

3.7 Plan Execution Subsystem

The Plan Execution Subsystem is described in details in following section,

1. Action plan receiving (APRe) Sub Phase;
2. Execution knowledge receiving (EKR) Sub Phase;
3. Plan execution Sub Phase;
4. Action reporting (ARp) Sub Phase; and
5. Alarm Sub Phase.

APRe receives quality action plans generated and reported by the planning subsystem, and provide them to plan execution sub phase. EKR obtains the quality plan execution standards, procedures and regulations through knowledge reporting sub phase and provide them to plan execution sub phase. Plan execution phase manages the execution of the received quality action plans using the received quality plan execution standards, procedures and regulations. Action reporting phase generates required reports for the appropriate quality control authorities and/or systems, effectors, alarm phase (when needed), and information management phase. Alarm phase generates time-stamped alarm messages to notify appropriate authorities, Alarm phase also notify information management phase of the alarm activation to ensure the event is recorded in the quality information DBMS.

3.8 Logical Framework Description

Figure 2 shows that the system is divided into different part of subsystems. Each subsystem manages its own internal state and behavior while managing its interactions with an environment that consists of numerous signals and messages from other subsystems and outside the system.

The proposed logical model of interconnected logical modules follows a number of Sensing module, Analyzing module, deciding module and acting module. The sensors act as a bridge between the external world and the system. There are multiple numbers of sensors present in the sensing and measurement module. These sensors collect information from inside and outside of the system. This information is sent to the External analyzer through the Monitor module. This acts as continuous stimuli received by the Controller unit to analyze its external environment.

The Internal and External Analyzer inside Analyzer module is devoted to the analysis of different types of heterogeneous data and generation of contextual information. The analysis is to be done with respect to some predefined rules. The autobiographical Memory stores previous information or reports and is used as a reference model by the Analyzer, consists of Short and long term memory for storing the newly generated data and long duration predefined standard respectively. The Memory Manager Module used to store the data that are generated by the event of Monitor and Analyzer module. In addition it forms a Knowledge base for higher level modules. In artificial system this is used to store data in buffer for future use. The data which is stored can be used as analyses criteria in future. The Decision Manager depicted in the diagram uses the context awareness represented by the output of previous modules and has its influence on the Aml domain with its internal and external world. The output of this Planning module is executed by Execution module to show the final result of the system. The result may be any broadcast, unicast message or any alarm to alert the system about some unexpected fault.

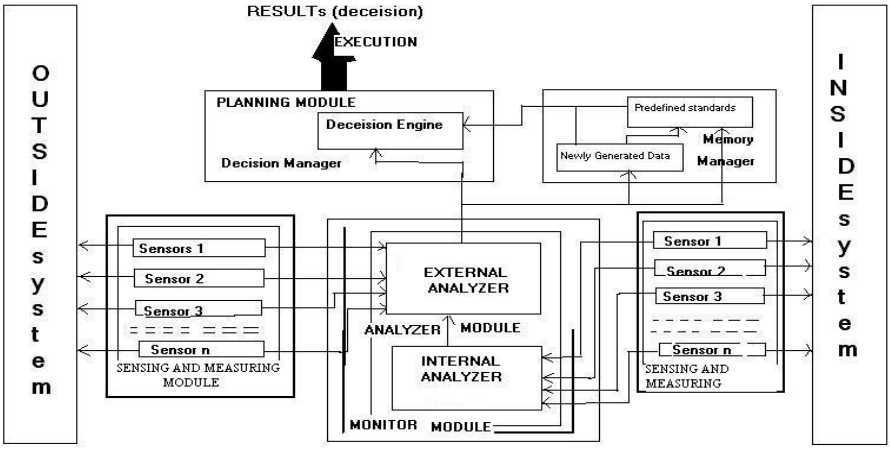


Fig. 2. Proposed Logical Architecture

4 Case Study

Figure 3 desribes a typical road condition. Vehicle A uses an optical sensors for sensing a road blockage along its path and sends this information to the nearest roadside sensing equipment R1. R1 broadcasts this information to other road side equipments and other vehicles as a warning message. The vehicles in the vicinity can thus decide on alternate routes well in advance. Collision warning, road sign alarms and in-place traffic view enables the driver essential tools to decide the best path along the way. The situation is described in details in following section.

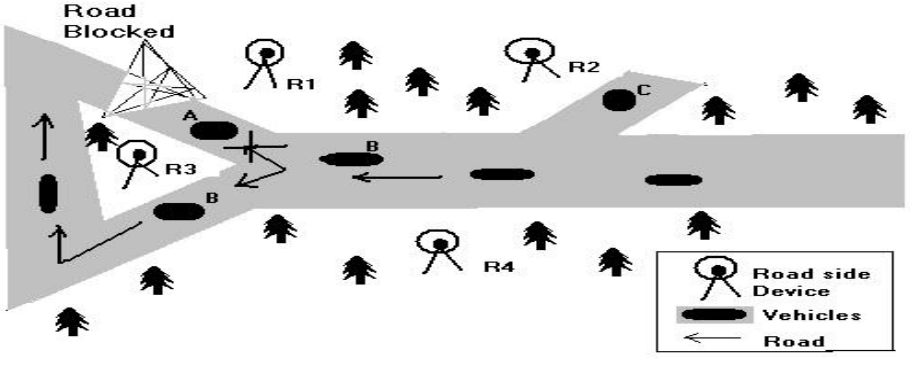


Fig. 3. Typical VANET structure

4.1 Case 1: Accident Occured; No-entry Signal Is Placed on Blocked Road

The sensor (camera) present in front of the vehicle captures the symbol. The sensor has an unique sensor id it attaches this id with message and encapsulates it in a packet. The data packet sends to MIR (Monitor Information Receiving) sub module. MIR only receives the encapsulated data from S/M reporting module and sends this to monitor module. MKR (Monitor Knowledge Receiving) receives the data related to that sensor from sensor database present in the nearest road side unit in distributed manner and sends this information to the monitor sub module. In Monitor sub module it will receive the data from both MKR and MIR. MKR sends the data related to that sensor (i.e.-sensor id, sequence no etc.). MIR sends the data packets which contain the sensor id and sequence no with the data. Monitor will compare both the sensor id if it is same then check for sequence number. If it is right then forward this data packet to AIR (Analysis Information Receiving) sub module otherwise discards it. It only receives the data packets from MR (Monitor Reporting) and extracts the actual data from the encapsulated data packets. AKR (Analysis Knowledge Receiving) retrieves the information from the knowledgebase present in the nearest RSU and sends this message to Analysis sub module. Information related to road signal is received here in this case. Analysis sub module analyze the information receives from AIR with the help of threshold value receives from AKR. The analyze result is send by AR (Analysis Reporting) sub module to the next PIR (Planning Information Receiving) sub module. Suppose here, analysis sub module analyze the data and report that it is signal of “no entry”. PIR sub module receives the information from AR (i.e. “no entry”) and sends it to planning sub module. Planning sub module decides to stop the car and broadcast this message. The plan generated by Planning sub module is delivered to Execution phase by planning reporting sub module. Finally, the execution or alarm generation is done by execution module to stop the car as there is an accident happened.

4.2 Case 2: Abnormal Pulse Rate

A wearable sensor is embedded in the clothing of the driver for sensing the pulse rate of driver. The sensor has a unique sensor id it attaches this id with message and encapsulates it in a packet. The data packet sends to MIR (Monitor Information Receiving) sub module. MIR only receives the encapsulated data from S/M reporting module and sends this to monitor module. MKR (Monitor Knowledge Receiving) receives the data related to that sensor from sensor database present in the nearest road side unit in distributed manner and sends this information to the monitor sub module. In Monitor sub module it will receive the data from both MKR and MIR. MKR sends the data related to that sensor (i.e.-sensor id, sequence no etc.). MIR sends the data packets which contain the sensor id and sequence no with the data. Monitor will check the authentication of the received data. In this sub module sensor id is checked if the it matches with the stored one then information will be forwarded to next phase otherwise discard it. AIR (Analysis Information Receiving) sub receives the data packets from MR and extracts the actual data from the encapsulated data packets. AKR (Analysis Knowledge Receiving) retrieves the information from the knowledgebase present in the nearest RSU and sends this message to Analysis sub module. Information related to that particular driver is received here in this case.

Analysis sub module analyze the information receives from AIR with the help of threshold value receives from AKR. The analyze result is send by AR (Analysis Reporting) sub module to the next PIR (Planning Information Receiving) sub module. Suppose here, analysis sub module analyze the data and report that it is below the threshold value of normal pulse rate. PIR sub module receives the information from AR (i.e. below threshold) and sends it to planning sub module. Planning sub module decides to stop the car and send this nearest health care authority or hospital. The plan generated by Planning sub module is delivered to Execution phase by planning reporting sub module. Finally, the execution or alarm generation is done by execution module to stop the car and informed to proper authority about the health condition of that driver. As a result, "Stop the car" message will be displayed and unicast the message to proper authority.

5 Conclusion

Much work has been done in vehicle to vehicle communication. With the rapid increase in vehicular traffic along highways. This has increased the need for traffic co-ordination, particularly in case of emergency situations. Ambient intelligence is the present day solution for gathering information from heterogeneous systems and linking them together to frame intelligent responses. The application of ambient intelligence is obvious in smart home, smart office, Patient caring system, vanet etc. Researchers had been working for quite some time in the area of VANET for improving highway safety and information services. In this paper we proposed an ambient intelligent architecture for emergency response generation involving vehicular network. Research is now on to formulate a mathematical model for this proposed architecture.

References

1. Ni, S.-Y., Tseng, Y.-C., Chen, Y.-S., Sheu, J.-P.: The Broadcast Storm Problem in a Mobile Ad HocNetwork. In: ACM/IEEE MobiComm (1999)
2. Xu, Q., Mak, T., Sengupta, R.: Vehicle-to-Vehicle Safety Messaging in DSRC. In: ACM VANET (2004)
3. Yang, X., Liu, J., Zhao, F., Vaidya, N.: A Vehicle-to-Vehicle Communication Protocol for Cooperative Collision Warning. Mobiquitous (2004)
4. Torrent-Monero, M., Jeng, D., Hartenstein, H.: Broadcast Reception Rates and Effects of Priority Access in 802.11-Based Vehicular Ad-Hoc Networks. In: ACM VANET (2004)
5. Bettol, C., Campl, C.: Challenges for Ambient Intelligence: Empowering the Users. IOS press (2005)
6. Bosse, T., Jonker, C.M., Treur, J.: Formalisation of Damasio's Theory of Emotion. Feeling and Core Consciousness (2007)
7. Casari, P., Castellani, A.P., Cenedese, A.: Wireless Sensor Networks for City-Wide Ambient Intelligence (WISE-WAI) Project (2009)
8. Jones, V., Gay, V., Leijdekkers, P.: Body sensor networks for Mobile Health Monitoring (2010)

A Cost Efficient Multicast Routing and Wavelength Assignment in WDM Mesh Network

Subhendu Barat, Ashok Kumar Pradhan, and Tanmay De

Department of Computer Science and Engineering,
National Institute of Technology Durgapur, India

Abstract. Multicast Routing and Wavelength Assignment (MRWA) is a technique implemented in WDM optical networks, where dedicated paths are established between a source and a set of destinations, unlike unicasting where a source is connected with only one destination. For a multicast session request a multicast tree is generated to establish a connection from source to all the destinations. A wavelength is assigned to each and every branches of the generated multicast tree to create a light-tree for the session. In this work, we have tried to minimize the wavelength usage to establish multicast sessions for a set of multicast session requests. Our approach is to minimize the size of the multicast tree by sharing branches, as much as possible, to connect all the destinations from the source node. A lesser usage of links minimizes the collision probability for the assignment of wavelength, say w, in each of the selected links to be assigned the wavelength. Secondly, greater sharing implies lesser splitting. As splitters are costly, minimum usage of splitters incurs lesser infrastructure cost in the network. The effectiveness of our approach has been established through extensive simulation on different set of multicast session under different network topologies and comparing with standard Minimal Spanning Tree (MST) based algorithm. The simulation shows our algorithm performs better than the MST based algorithm.

Keywords: Wavelength Division Multiplexing (WDM), Light-tree, Splitting, Multicast Routing and Wavelength Assignment (MRWA), Multicast Session.

1 Introduction

In today's world, communication needs a faster and safer media. As the modern science and technology progresses the various needs of communication are coming into the scenario. Now a days, Optical Fiber Communication is playing a major role in data communication area. As one-to-many communication is a recent need for both business and domestic purpose, the problem to connect multiple recipients with a single sender efficiently is becoming a vital issue for modern day communication. As conventional broadcasting will not suffice the need because of security and traffic issues, hence concept of Multicasting evolved. In fiber optics communication this problem is studied as Multicast Routing and Wavelength Assignment (MRWA) Problem.

In most of the works on MRWA light-tree concept is used to establish multicast session. *Light-tree* [4,5] can be assumed to be UNION of the light paths between each source-destination pair, where the source of the multicast session is common. The concept of *light-path* is vastly implemented in unicast RWA problems, where a light-path is

N. Chaki and A. Cortesi (Eds.): CISIM 2011, CCIS 245, pp. 65–73, 2011.

treated as a logical connection between a source and a destination node in optical layer. Unicast RWA problems are encountered in [16,17,18]. Several RWA [5,8] schemes have been proposed that differ in the assumptions on the traffic pattern, availability of the wavelength converters, and desired objectives. The traffic assumptions generally fall into one of the two categories: static or dynamic [1,2]. In static RWA models we assume that the demand is fixed and known, whereas in dynamic model session requests arrive and terminate randomly. The MRWA problem is an NP-Complete problem [7]. The general approach is to divide the problem into two sub-problems: One is routing, i.e., constructing a multicast tree starting from the source node and covering all the destination nodes. The other one is wavelength assignment problem, which is to assign one or more wavelengths on each link in the multicast tree. If a single wavelength is assigned to all the links of the computed multicast tree to connect *all* the destinations a *light-tree* is generated. Whereas, if a wavelength is assigned to a part of the sub-tree connecting a *subset* of the original destination set, i.e. more than one wavelengths are used to assign all the branches of the multicast tree, hence generating a set of light-trees for a single multicast session request, is known as *light-forest*. In absence of wavelength converters in a network a multicast tree is assigned only one wavelength and the network is known as *single hop* network and if the internal nodes are equipped with wavelength converters the network becomes *multi-hop*. Wavelength converters are costly devices which can convert an incoming laser ray into an outgoing ray of different wavelength. Several works used wavelength converters in wavelength assignment problem [14,19]. A new multicast routing structure *light-hierarchy* was proposed for all multicast routing in [10], which permits the cycle introduced by *Cross Pair Switching* (CPS) capability of *Multicast Incapable* (MI) nodes. To support the WDM multicast function, a switch node should be equipped with a device called a *Splitter*, which can split one incoming light signal from the incoming link to multiple outgoing signals on different outgoing links. In this way, one copy of the incoming information becomes multiple copies of outgoing information and thus the multicast is implemented. In practice, an all-optical WDM network is equipped with a limited number of splitters. Hence, in wavelength assignment for WDM multicast, we need to consider the splitting constraint [3,6]. The different research studies on multicast routing wavelength assignment can be found [3,5,6,7,8,9,10,11,12,13,15].

The sharing based multicast routing (SBMR) algorithm, proposed here, takes into account two important issues - minimization of wavelength usage and minimization of splitting requirement in consideration. Thus the algorithm gives a cost effective solution for the MRWA problem as both the wavelength and splitters are costly resources, and hence there is a need of proper resource management. We have also presented simulation result of our algorithm SBMR and compared with the well-known Minimal Spanning Tree (MST) based algorithm.

The rest of the paper is organized as follows. Section 2 presents the problem formulation. The proposed approach of multicast routing and wavelength assignment problem is presented in Sect. 3. The experimental result and its analysis are described in Sect. 4. Finally, conclusion is given in Sect. 5.

2 Problem Formulation

The objective of this work is to minimize the number of splitters in all optical networks such that the wavelength requirement is minimal. An improper utilization of splitters may lead to poor wavelength resource management which blocks high percentage of connection due to unavailability of wavelength. We assume splitting at a node in the light tree is independent of the splitting for other sessions.

Let G = (V, E) represents an all-optical WDM network with N number of nodes where V is the set of vertices and E is the set of edges. The graph is bi-directional. Let R = r_i and $i = 1,2,...,m$ be sequence of m multicast requests. Let w_i be the wavelength resource usage for the multicast request r_i. If a multicast request requires one wavelength on one link, the wavelength resource usage for the request on the link is one unit of the wavelength resource. Let $l_1, l_2, ..., l_m$ be the splitting requirement for m multicast requests. Let W and L denote the total wavelength resources and splitting requirement of all m multicast sessions respectively.

The objective functions are to minimize the number of wavelengths and splitting requirements for establishing a set of multicast sessions, i.e.,

$$Minimize \quad W = \sum_{i=1}^{m} w_i \tag{1}$$

$$Minimize \quad L = \sum_{i=1}^{m} l_i \tag{2}$$

Constraints

1. All the destinations in a single multicast session request are to be connected by a single light-tree.
2. No node in the multicast tree is traversed twice, i.e. the solution should be free from loops and redundant paths.
3. One single wavelength is assigned to all the branches of the multicast tree to generate the light-tree.

3 Proposed Approach

Wavelength minimization is one of the objective of MRWA problem. The other cost parameter considered here is the cost due to splitting. The cost of splitting is due to installation of power amplifiers in the splitter node to compensate the attenuation of power of the split signal. Sharing reduces the requirement of splitting at a node, hence installation cost of power amplifiers are reduced.

In this work, we have tried to minimize the wavelength requirement by efficient routing. The approach we have taken is to reduce the size of the multicast tree, i.e. the total number of branches in the multicast tree. Our proposed algorithm sharing based multicast routing reduces the tree size by link sharing, hence minimizes the wavelength requirement.

Let us assume, at any instant of time all the links in the network is more or less loaded equally, which is due to randomness of multicast session requests. Hence, the probability that a particular wavelength, say "w", will be available at a link l_i, i=1 to n is $0 \leq p_i \leq 1$. Say, there exist two alternative multicast trees P_x and P_y, both establishing

the same multicast session. Say the size of P_x is S_x and that of P_y is S_y and $S_x > S_y$. Due to a symmetric load across the network we can assume $p_i \approx$ p where, $0 \leq p \leq 1$. So the probability that wavelength "w" will be available throughout the tree P_x and P_y are p^{S_x} and p^{S_y} respectively. Since, $S_x > S_y$; $p^{S_x} \leq p^{S_y}$. Hence a smaller multicast tree is better for successful wavelength assignment.

In this paper, the multicast routing and wavelength assignment (MRWA) problem is divided into two sub-problems: Multicast tree generation (routing) and wavelength assignment. Here our main focus is on efficient routing, where we have tried to minimize the splitting requirement and tree size by means of sharing. In the wavelength assignment phase we have implemented the standard First-Fit technique. As in the wavelength assignment phase no further optimization is tried, basically the efficient routing in the routing phase is responsible for the efficient performance of our algorithm in the field of wavelength minimization.

3.1 Multicast Tree Generation

We have proposed Sharing Based Multicast Routing (SBMR) algorithm to generate a multicast tree. To construct a multicast tree each session is taken one at a time and the corresponding multicast tree is generated for each session. Then all possible alternative routes from source node to each of the destination nodes in the session are explored until all the destinations are reached, in BFS order such that no non-destination nodes are reached twice. Each combinations of alternative routes to reach every destinations are picked one by one. If the picked solution contains a node traversed twice then that solution is rejected for containing a loop, else the metric "link sharing" among the routes are computed for the chosen combination. The combinations are ranked by the computed value of metric "link sharing". The multicast tree is constructed by using UNION operation on the paths of best ranked combination. This routing algorithm SBMR is shown in Algorithm 1.

3.2 Wavelength Assignment

We have used standard First-Fit wavelength assignment technique to generate light-tree. In this technique, at first we select the lowest index available wavelength, say w_0, to assign in all the branches of the multicast tree. If w_0 is not available in any branch of the multicast tree then the wavelength is ignored and the next higher wavelength is selected for assignment and so on. The wavelength which fits first i.e., available in all the branches of the multicast tree is assigned and the light-tree is generated.

3.3 Example

We analyze the algorithm SBMR by an example on 14 nodes NSF network as shown in Fig.1. The number beside each node indicates the node ID and the number beside each link shows the link ID.

We have taken a multicast sessions request: T = {4, 2, 8, 12, 13}. Here the source node is 4 and destination nodes are {2, 8, 12, 13}. The alternative paths explored between (4, 2), (4, 8), (4, 12) and (4, 13) are {(9 -> 5), (6 -> 2 -> 1)}, {(8 -> 10 ->

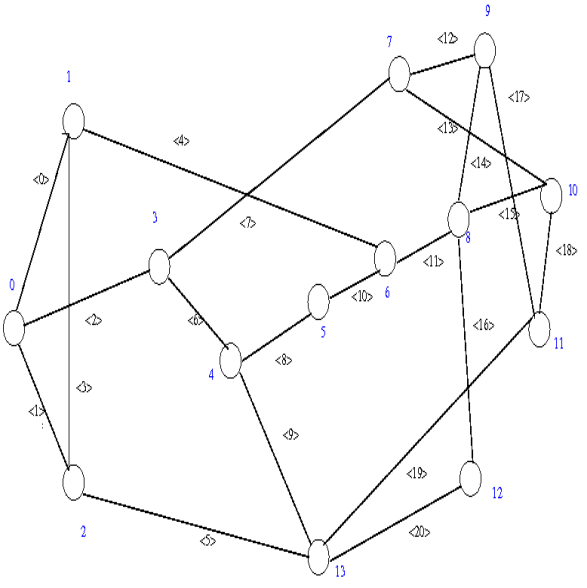


Fig. 1. NSFNET-14

Table 1. Comparison of all possible combinations of explored paths

Alternatives	Extent of link Sharing	Tree size	Selected
1 - 1 - 1 - 1	2	6	x
1 - 2 - 1 - 1	4	4	✓
2 - 1 - 1 - 1	1	8	x
2 - 2 - 1 - 1	3	6	x

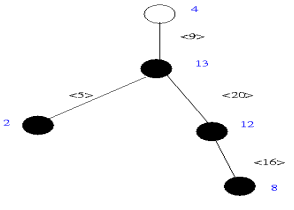


Fig. 2. Multicast tree using algorithm SBMR

11), (9 -> 20 -> 16)}, {(9 -> 20)}, {(9)} respectively. The comparison of all possible combinations of these explored paths is shown in Table 1.

Since option 2 has maximum sharing, it is selected to form the multicast tree {9, 20, 16, 5} of size 4, as shown in Fig. 2. This multicast tree is generated by union of 1st, 2nd, 1st, and 1st alternative paths explored between (4, 2), (4, 8), (4, 12) and (4, 13) respectively.

Algorithm 1: Sharing Based Multicast Routing (SBMR)

Input : One Multicast session request R(s, D) where |D|=N
Output: One Multicast Tree T, corresponding to the input

```
1  Initialize a list L, containing all the destination nodes in a session
2  Mark them as UNVISITED
3  Initialize a queue Q by inserting the Source node of the session
4  Initialize a counter LEVEL by 0 and an indicator IND pointing the first
   (left-most) and last (right-most) node of the explored graph at a particular level
   down from the source node, initially it points (0, 0)
5  Create a 2-D dynamic list ALT, which will contain all the alternative paths to a
   destination from the source, initialized as empty
6  while a node in L is marked UNVISITED do
7      while a node v is in queue Q and within the range of IND associated with the
       current LEVEL do
8          if v ∈ L then
9              mark it VISITED and Retrieve the path from the source node and
               insert it into ALT[i], where L[i] = v
10         Explore the node, i.e. evaluate its children CH[1, n] with the help of
           Network's Adjacency Matrix
11         for i=1 to n do
12             if CH[i] is in Q AND not in L then
13                 continue
14             else
15                 Insert CH[i] into Q
16     Increment the counter LEVEL by 1
17     Update IND
18 Sharing=-1
19 for all combinations of ALT[i], i =1 to N do
20     if the picked combination contains a node more than once then
21         continue
22     else
23         Sh = Calculate Sharing
24         if Sh> Sharing then
25             Sharing = sh
26             Update the current combination as the solution
27 The multicast tree T is generated by doing UNION on the Branch sets of each of
   the paths in the selected combination of paths
28 Exit
```

4 Analysis of Result

We conducted our simulation on various randomly generated multicast sessions on 14-node NSF network, as shown in Fig.1. Here we have shown the performance of our algorithm SBMR with respect to a standard MST based routing and First-Fit wavelength assignment algorithm. The performance metrics are wavelength requirement for a set of multicast sessions, average tree size per session, and average splitting requirement per session. The average value of 50 iterations of simulation on randomly generated multicast requests is shown in this section.

In Fig. 3, we have plotted the performance graph of wavelength usage vs. total number of multicast session requests, where we have varied the number of requests in the range [0, 100] keeping maximum size of a request to 8 nodes. From this figure we can see that our proposed algorithm SBMR seeks lesser wavelength for a particular simulation point. The reason behind this is, our proposed approach generates smaller sized tree than existing.

In Fig. 4, we have plotted the performance graph of average tree size against maximum size of multicast sessions, where we have varied the size of session requests in the range [2, 10] and considering total number of multicast session requests equal to 50. Algorithm SBMR provided multicast tree of smaller size than the standard algorithm for all simulation points.

The relationship between the average splitting requirement per session and maximum size of multicast session requests is shown in Fig. 5. We have considered total number of multicast requests equals to 50 and varied the size of session request in the range [2, 10]. The figure shows that algorithm SBMR needs lesser splitting than the standard algorithm. The result is quiet implicit, as the standard algorithm have not taken splitting constraint in consideration in the routing phase, where our heuristic suits well for the splitting minimization problem, as discussed in detail in the Sec. 3 of this paper.

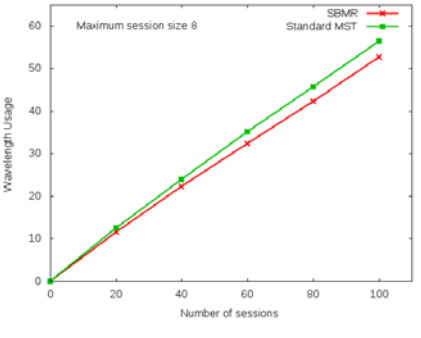


Fig. 3. Wavelength usage vs Number of sessions

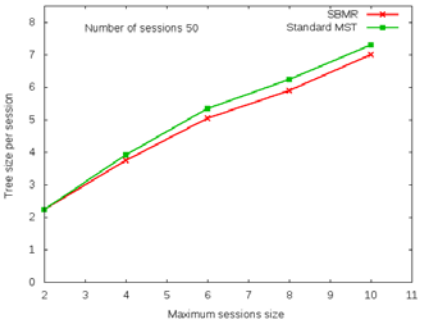


Fig. 4. Tree size per session vs Maximum session size

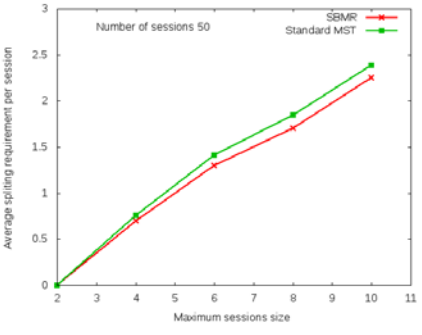


Fig. 5. Average splitting requirement per session vs. Maximum session size

5 Conclusion

In this work, we have proposed a heuristic for multicast routing and wavelength assignment problem in WDM mesh networks which provides a cost efficient solution. We have developed the routing algorithm SBMR that successfully reduces the tree size as well as splitting requirement for routing of a multicast session. As splitting requires costly power amplifiers at the splitting nodes so, more splitting requires installation of more such costly amplifiers. As our approach tries to minimize the usage of splitting while satisfying a multicast session request, the cost of establishing a multicast session is reduced very much.

At the same time wavelength channel requirement is also reduced due to smaller size of the tree. Wavelength channels being a physical resource is constrained by the physical network under usage. Our observation shows when the session size increases, i.e. the network density is denser, wavelength requirement also increases, but the increase rate is very slow. Therefore, it significantly reduces cost of establishing multicast session by reducing number of wavelength usage, which makes our approach workable under high network traffic density. Thus our approach suits well in a low-cost WDM mesh network with constrained splitting and wavelength channel capacity.

References

1. Sivalingam, K., Subramaniam, S.: Optical WDM Networks: Principles and Practice. Kluwer Academic Publishers, Boston (2000)
2. Bandyopadhyay, S.: Dissemination of Information in Optical networks from technology to algorithms (Texts in theoretical computer science an EATCS series)
3. Wang, J., Qi, X., Chen, B.: Wavelength assignment for multicast in all-optical WDM networks with splitting constraints. IEEE/ACM Trans. Netw. 14, 169–182 (2006)
4. Yang, D., Lio, W.: Design of light-tree based logical topologies for multicast streams in wavelength routed optical networks. In: IEEE INFOCOM (2003)
5. Suman, A., Ozdaglar, E., Bertsekas, P.: Routing and wavelength assignment in optical networks. IEEE/ACM Trans. Netw. 11, 259–272 (2003)
6. Zhou, F., Molnar, M., Cousin, B.: Multicast routing and wavelength assignment in WDM mesh networks with sparse splitting. In: The 5th International Workshop on Traffic Management and Traffic Engineering for the Future Internet (EuroNF 2009), France (2009)
7. Libeskind-Hadas, R., Melhem, R.: Multicast routing and wavelength assignment in multihop optical networks. IEEE/ACM Transactions on Networking 10, 621–629 (2002)
8. Banerjee, D., Mukherjee, B.: A practical approach for routing and wavelength assignment in large wavelength routed - optical networks. IEEE J. Sel. Areas Commun. 14, 903–908 (1996)
9. Bharath-Kumar, K., Jaffe, J.: Routing to multiple destinations in computer networks. IEEE Transactions on Communications com-31, 343–351 (1983)
10. Zhou, F., Molnar, M., Cousin, B.: Light-Hierarchy: The optimal structure for multicast routing in WDM mesh networks. In: The 15th IEEE Symposium on Computers and Communications, ISCC 2010, vol. abs/1012.0017 (2010)
11. Wattanavarakul, W., Segkhoonthod, S., Wuttisittikulkij, L.: Design of multicast routing and wavelength assignment in multifiber WDM Mesh networks for asymmetric traffics. In: TENCON 2005 IEEE Region, vol. 10, pp. 1–6 (2005)
12. Poo, G., Zhou, Y.: A new multicast wavelength assignmet algorithm in wavelength-routed WDM networks. IEEE J. Select. Areas Comun. 24(4) (April 2006)
13. Jia, X., Hu, X., Lee, M., Gu, J.: Optimization of wavelength assignment for QoS multicast in WDM networks. IEEE Transactions on Communications 49, 341–350 (2001)
14. Yoo, S.: Wavelength conversion technologies for WDM network applications. J. Lightwave Technol. 14, 955–966 (1996)
15. Takahasi, H., Matsuyama, A.: An approximation solution for the Steiner problem in graphs. Math. Japonica 24, 573–577 (1980)
16. Bermond, J., Gargano, L., Perennes, S., Rescigno, A., Vaccaro, U.: Efficient collective communication in optical networks. In: Presented at the Annu. Int. Colloq. Automata, Languages and Programming (ICALP), Paderborn, Germany (July 1996)
17. Chlamtac, I., Farago, A., Zhang, T.: Lightpath (wavelength) routing in large WDM networks. IEEE J. Select. Areas Commun. 14, 909–913 (1996)
18. Sahasrabuddhe, L., Mukherjee, B.: Light trees: Optical multicasting for improved performance in wavelength-routed networks. IEEE Communication Magazine 37(2), 67–73 (1999)
19. Chatterjee, M., Barat, S., Majumder, D., Bhattacharya, U.: New Strategies for Static Routing and Wavelength Assignment in De Bruijn WDM Networks. In: Third International Conference on Communication Systems and Networks, COMSNETS (2011)

Combination of Decision Support System (DSS) for Remote Healthcare Monitoring Using a Multi-agent Approach

Mohamed Achraf Dhouib, Lamine Bougueroua, and Katarzyna Węgrzyn-Wolska

Esigetel, 1 rue du Port de Valvins,
77210 Avon, France
{achraf.dhouib,lamine.bougueroua,katarzyna.wegrzyn}@esigetel.fr

Abstract. This research is in the field of remote healthcare monitoring systems which propose software solutions to monitor elderly people in their own homes. Our objective is to take advantage of the technological diversity of several Decision Support Systems used to detect distress situations. We propose a multi-agent approach in which each agent encapsulates a decision support system. This encapsulation enables the real-time combination of decisions. In this paper, we present the architecture of our multi-agent system and the real-time scheduling of the collective decision process.

Keywords: multi-agent system, decision support system, collective decision, real-time scheduling, remote healthcare monitoring.

1 Introduction

In the coming decades, many European countries will be confronted with issues relating to an aging population. It is estimated that by 2020, 28% of the French population will be over 60[1]. Traditional solutions of housing the elderly in specialized centers have become too expensive; they also have a negative impact on the independence of the patient. Remote healthcare monitoring systems represent a much more convenient solution which assures the social independence of the elderly as well as their security [2].

There are currently many research projects in this context which implement diverse solutions using various technologies. A corner stone technology in this field is the decision support system (DSS). These systems are able to analyze the data gathered from several ambient captors to generate risk detection decisions. These decisions are sent to a remote monitoring center primed to take action in the case of real threat. Although they use several data modalities (localization, physiological, actimetric…), the DSS usually use a unique type of artificial intelligence (neural network, expert system, fuzzy logic…). The pertinence of each DSS is determined by the occurrence of alarms which are either false or undetected. A real-time combination of these decisions is able to improve the usage of appropriate resources within an acceptable response time. In order to make best use of this technological diversity, we aim to encapsulate each DSS in an intelligent agent. In fact, the multi-agent system (MAS)

N. Chaki and A. Cortesi (Eds.): CISIM 2011, CCIS 245, pp. 74–84, 2011.

architecture enables these DSS to have a uniform view of the decision concept and to exchange both knowledge and intelligence.

This paper presents a theoretical framework for real-time dispatching of a multi-agent approach for collective decisions. In the first section, we present a survey of three DSS used in health care telemonitoring [2],[3]. In the second section, we present the method for encapsulating a DSS in an intelligent agent. This encapsulation is achieved by transforming the generated decision of the DSS into an abstract form that may thus be used as an environmental fact by other intelligent agents (those agents are probably encapsulating other DSS). For this reason, we propose in this section a semantic alignment approach based on the works of [4]. We then define the 2-phase collective decision algorithm of our MAS and the role of the central agent. In the third section we present the real-time dispatcher plugged into the central agent. The first part of this section is dedicated to explaining the dispatching problems of our MAS. Subsequently we present the different types of message and the classification made by the real-time classifier. In the same section, we present dynamic priority assignment and the scheduling algorithm. In conclusion we will present the advantages of our system and finally outline our future work.

2 A Survey of the Remote Healthcare Monitoring Systems

The objective of remote healthcare monitoring systems is to enable people to live in their own homes longer. The use of artificial intelligence is crucial in this context and it has been the subject of many research projects.

Several of these projects are both ambitious and promising. They regroup expert partners from several domains. In this section we describe the risk-detection method used in three pre-selected remote healthcare monitoring projects. We expect to give the reader an overview of the advancement in hardware and artificial intelligence in this area. Next we present a conclusion of this survey in order to establish the foundations upon which our project is based, and the potential advantages of our approach.

The AILISA project [2],[5] (Intelligent Apartments for effective longevity) is an experimental platform for the evaluation of remote care and assistive technologies. This experimental platform is based on an information system that circulates information between the patient residence and the telemonitoring site. The information system collects data from patients and their environment through a local network of sensors connected to a computer (PC) which is in turn connected to the Internet. The computer has software to communicate with sensors and other entities within the system but also to transform data from sensors into an adequate format in the database, and finally to deliver indicators on the status of the patient. These indicators are developed by the fusion of information provided by the sensors. This project is the result of the combined work of several experts on smart homes, networks, software engineering, electronics and signal processing with extremely interesting results.

The S(MA)²D system aims to carry out a generalization of profile patterns in order to classify the people monitored. Its function consists of transforming a selection of sensory data and other personal information into indicators. These indicators are

collected by several distributed agents who then generate pattern classification. This classification is actually multi-agent because the classification result is not the work of a simple agent, as it is the case in other multi-agent systems. It is really a collective work [2]. This multi-agent classification approach can spot an adequate pattern in an open and dynamic environment. It does not depend on the type of the indicators and does not require preliminary categories.

This system may be used in preventative control and the telemonitoring of people suffering from chronic health problems. The most relevant aspect of the S(MA)²D is the organization of the assistance of multiple domiciles for the elderly and the detection of global medical problems as it analyzes the telemonitoring issue in a global, rather than in individual-centered way. In fact, this system is a large scale or global solution "the uninterrupted monitoring of (a) hundred people". It requires a strongly distributed and dynamic system.

The QuoVADis project [6] aims to compensate for losses in cognitive abilities leading to difficulties in communication. This generates social isolation, depression, insecurity and discomfort in daily life. The system aims to restore the emotional connection with loved ones, caregivers and healthcare providers by a mobile interactive robot which accompanies the person in question. The project aims to resolve both problems posed by home support: cognitive stimulation and the medical safety of patients. QuoVADis implements a decision support system called EMUTEM. This DSS is a multimodal platform for remote healthcare monitoring. It uses fuzzy logic to achieve fusion between several data modalities (physiological, sound, actimetric).

We studied three projects able to monitor people in their domicile. The common principle between theses systems is gathering information in order to present, as clearly as possible, the situation of the patient. This diversity helps researchers to lead new projects and make progress in their results. The main drawback is the difficulty of the technological exchange between these projects. In fact, progress in this domain would be greater if we could automatically combine the results of different solutions functioning at the same time. The potential advantages of such a combination are:

1. the possibility of fusing the results in order to make more pertinent decisions
2. the possibility of an automatic online learning adaptation for the DSS, comparing its results with the others
3. the specialization of each system in a precise type of data (modality)

Table 1. Data sources and prevention methods of the three projects

	QuoVADis	**S(MA)²D**	**AILISA**
Data sources	• presence detector • intelligent microphones • physiological captors • ambulant robot	• sensor • nurse notes, questionnaire • software to transform data into indicators	• presence detector • smart shirt • ambulant robot
Distress prevention	EMUTEM, fuzzy logic, data fusion	Multi-agent classification of patterns	Hardware methods

3 Encapsulation of a DSS in an Intelligent Agent and Collective Decision Principle

3.1 The Architecture of a BDI Agent for the Encapsulation of a DSS

An intentional agent is a cognitive agent (able to recognize and represent its environment) with intentional attitudes [7]. These attitudes have been defined by [8] as establishing a strategy capable of interpreting the behavior of an entity (human, animal, machine, etc…).

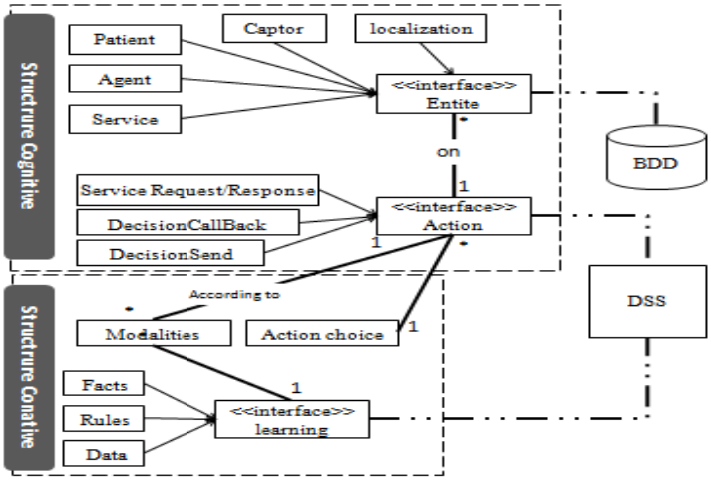


Fig. 1. Inner design of an agent (entity)

In a remote healthcare monitoring system we need such a solution in order to understand the behavior of the patient and the state of their domicile. From this information the system will adapt. According to [9] and [10], the intentional agents are incarnated by the BDI agents (Belief-Desire-Intention). To model a BDI agent, we relied on the works of [11] which propose the following architecture:

- Cognitive: for environmental representation and action planning.
- Conative : for inner state and intention modeling

We designed this architecture as described in Figure 1. In this design, we based our research on the GAIA approach. GAIA is an extension of classical software engineering approaches [11] which proposes several conception models for multi-agent systems. We based our conception on the role model (also known as the control model). This model consists of defining the different roles to be played in the multi agent system. A role is defined by: the permission/rights associated to the role and the role's responsibilities.

The Agent's autonomy in this context is expressed by the fact that a role encapsulates its functionality which is internal and does not affect the environment [7]. The

roles stated in our architecture are initially defined according to the collective decision process. In fact, each agent's role is to make decisions according to modalities used by the DSS it encapsulates. The functionality of the DSS is completely internal to the agent and does not affect the environment.

3.2 Decision Abstraction and Priority Assignation

In intelligent remote healthcare monitoring a decision support system uses the data flow of several modalities to generate decisions about the patient's situation. To standardize the decision concept, we classify the generated decisions by the modalities used. The considered modalities in our system are: sound, speech, physiological data (e.g. activeness and pulse rate), actimetric data (localization, falls), video, sensor states and alarm calls. Generally, every decision is based on global pertinence calculated by combining the pertinence affected to each decision modality. For a d decision, the global pertinence is:

$$Gp(d) = \sum_{m_i} p_i(d) \cdot c_i \tag{1}$$

Where:

- m_i: the modalities used for the decision d
- p_i: the pertinence of the decision d according to the modality mi
- c_i: the coefficient of the modality mi accorded by the DSS.

When a DSS generates a decision, it sends the data concerning this decision to its encapsulating agent. The agent reorganizes these data in a decision report (type, pertinence, arrival date …), which it then sends to the central agent.

The collective decision is made in two phases:

Phase 1: The central agent starts the wait window of phase-1. The duration of the wait window depends on the trigger decision data (agent affinity, modalities used …). In this paper, we do not detail the computing algorithm of the waiting duration. The decision messages received in phase-1 are called SEND decisions. A SEND decision is a spontaneous decision. It is not a response to a previous request. In the case of a trigger decision, we also define the pertinence threshold. The arriving decision reports during this first wait window are fused with the trigger decision. If the final decision's pertinence surpasses the threshold, the decision is confirmed as an alert. If the wait window is terminated without attaining the pertinence threshold, the central agent starts the second phase of decision.

Phase 2: the central agent starts a new wait window. During this wait window, a real-time consensus is launched among the agents concerned by the trigger decision modalities. For this purpose, the central agent assigns to each concerned agent a consensus priority. This is computed as follows:

$$p_i(d) = \sum_{m_j \in d} A_{ij} \cdot c_j \tag{2}$$

Where:

- m_j are the modalities used in the trigger decision d,
- c_j the corresponding coefficient for each modality,
- A_{ij} is the pertinence coefficient of the agent i for the modality m_j

During this second wait window, the received message may be SEND decisions. As they do not concern the launched consensus, they are placed in the wait queue.

The response messages are called CALL BACK decisions. At the end of the second wait window, the central agent computes the global pertinence of the received CALL BACK decisions. If the pertinence threshold is reached, the trigger decision is confirmed otherwise it is rejected and a learning procedure is sent to the responsible agent. In this article, we do not detail the inner learning procedure of such an agent.

4 Real-Time Scheduling of the Collective Decision Process

One of the major problems in the field of multi-agent systems is the need for methods and tools that facilitate the development of systems of this kind. In fact, the acceptance of multi-agent system development methods in industry depends on the existence of the necessary tools to support the analysis, design and implementation of agent-based software. The emergence of useful real-time artificial intelligence systems makes the multi-agent system especially appropriate for development in a real-time environment [12]. Furthermore, the response time of the DSS in a remote healthcare monitoring system is a central issue. Unfortunately the DSS studied in this context does not give a real-time response. For this reason we aim to control, as much as possible, the response time of their encapsulating Agents. The Gaia role model we presented in section 3 guaranties that the agent encapsulation of a DSS makes its response time transparent to the other agents.

4.1 General Operating Principle

This work has focused on a time-critical environment in which the acting systems can be monitored by intelligent agents which require real-time communication in order to better achieve the system's goal, which is detecting, as fast as possible, the distress situation of the patient. The works of [12] define a real-time agent as an agent with temporal restrictions in some of its responsibilities or tasks. According to this same work, a real-time multi-agent system is one where at least one of its agents is a real-time agent. The central agent is the unique decision output of our system. We will apply these definitions by focusing on the real-time scheduling of the central agent tasks. Firstly the different tasks of this agent must be defined. Subsequently, diverse scenarios and the priority assignation rules may be defined.

As explained previously, the central agent receives all the decision reports in the system. The first main issue is thus the scheduling of the treatment of these messages. For each decision received the central agent chooses the concerned agents and assigns a response deadline to each one, based on the degree of expertise of the concerned agent in the modalities used. We propose a scheduling model that enables the reaching of a consensus between the different concerned agents while respecting the defined response deadlines.

4.2 Definition of the Central Agent Tasks

As described in fig.2, an agent has two main functions: conative and cognitive. In the case of the central agent, the cognitive function consists of communicating with the other agents. The conative function consists of making final decisions.

This classification leads us to this list of tasks assigned to the central agent:

- Cognitive tasks :

– Message reception : connection establishing and stream reading
– Message classification: according to the type of the request, this task classifies each message in the appropriate wait queue.
– Entities representation: this task comes into play at each collective decision cycle. Its role is to keep the state of the other agents in the central agent memory, as well as that of the central agent itself.
– Message send : connection establishing and stream writing

- Conative tasks:

– Request analysis and execution: this task executes the selected message requests. Generally it triggers another task of the central agent (representation, message send, decision)
– Decision: this task maintains a fusion buffer in which the message execution task puts the decision message. When this task is activated, it adds all the un-executed messages in the highest priority wait queue to the fusion buffer.
– Deadline assignation: this task assigns an absolute deadline to each message before classification
– Message selection: this task selects the message to be executed from the message buffer.
– Phase manager: this task is responsible for the transition between the collective decision phases. It comes into operation when a wait window is closed or when a collective decision is made. Its main role is changing the priority of central agent tasks.

Each task is executed according to the automaton described in figure 2.

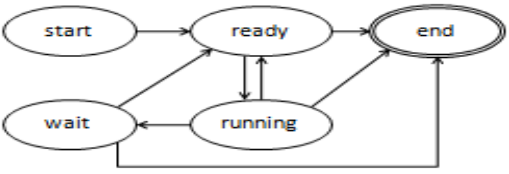


Fig. 2. Execution states of the central agent tasks

4.3 Message Classification

The central agent message buffer consists of 3 different wait queues (WQ): the CALL BACK queue, for the CALL BACK decision messages, the SEND queue, for the SEND decision message and the Best Effort queue, for the other communication messages (decisions, service requests …)

The BE queue is FIFO scheduled (First In First Out). There is no deadline or priority consideration in this queue. The CALL BACK and the SEND queue are EDF scheduled [13]. EDF is the preemptive version of Earliest Deadline First non idling scheduling. EDF schedules the tasks according to their absolute deadlines: the task with the shortest absolute deadline has the highest priority.

Each message deadline must be determined before being classified in a wait queue. For this reason the Deadline assignation task, the message classification task and the message reception task must be fused. In fact, when a message arrives, the message reception task is activated. It cannot then be preempted before assigning the message to its corresponding wait queue.

4.4 Queue Priority and Message Selection

The message queues have dynamic priorities. This priority is assigned by a phase manager task. In phase-1, the SEND queue has the highest priority. In phase-2, the CALL BACK queue has the highest priority. While the message buffer is not empty, the message execution task's state is *Ready*. When it passes to execution, it selects the shortest deadline message from the highest priority queue. During the wait window of phase-1, the received SEND must be executed first. Thus we assign the highest priority to the SEND queue. When this wait window is closed, the decision task gets the highest priority. The CALL BACK queue has the highest priority in phase-2. Thus a phase cannot be terminated until the corresponding wait queue is empty and all the received decisions fused.

4.5 Global Scheduling of the Central Agent

The main scheduling algorithm of the central agent is FP/HPF.

FP/HPF denotes the preemptive Fixed Priority Highest Priority First algorithm with an arbitrary priority assignment [14].

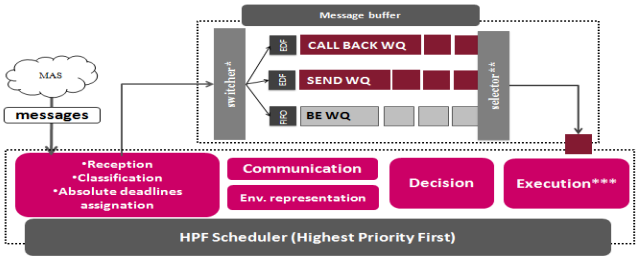


Fig. 3. Real-time scheduling of the central agent

In table 2, we present the priority evolution of each task during the different steps of the 2-phase collective decision. The higher the number, the higher the priority. The phase manager task always has the highest priority. In fact, it is responsible for changing the system phase and the priority assignation.

Table 2. Priority variation of the central agent tasks

Task	*Wait For trigger*	Phase-1		Phase-2	
		wait	*Decision*	*wait*	*Decision*
reception	4	4	2	3	1
Send	1	1	3	4	3
decision	2	2	4	1	4
execution	3	3	1	2	2
phase manager	5	5	5	5	5

4.6 Scheduling Sample

In figure 4, we present a scheduling sample in a system composed of a central agent (CA) and five other agents (A1, A2, A3, A4, A5). The red arrows represent the movement of the task to the ready state. Here we present the priority assigned to each task at the beginning of each phase. We suppose that the message wait queues are initially empty. Our sample scenario goes through these stages: a trigger decision from A3 is received. The execution task treats the received trigger and then requests that the phase manager start a new collective decision process. The phase manager starts the first phase. It opens a new wait window and changes the priority of the CA tasks. During phase-1, two SEND decisions are received (from A1 and A4). The first wait window is terminated by the phase manager task. The highest priority is assigned to the decision task. The pertinence threshold is not reached. The phase manager task starts the second phase. The highest priority in this task is accorded to the send task in order to allow the CA to activate the consensus. During the phase-1 decision process, the CA receives two SEND messages. The reception task is preempted because it has a lower priority. In phase-2, A1, A2, A4 and A5 are involved in the consensus (a choice based on the trigger decision modalities). A SEND and 3 CALL BACK decisions are received (positive: A1 and A5, negative: A4). The final fusion reaches the

pertinence threshold. Two learning procedures are sent to A4 and A2. We suppose that the message buffer is initially empty.

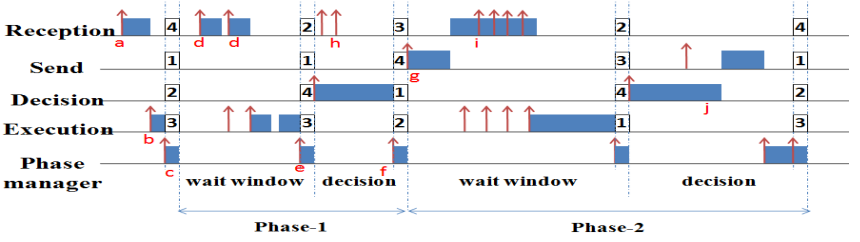


Fig. 4. Temporal diagram of a scheduling sample

The phase manager task is responsible for changing the priority of the central Agent tasks. We can observe on figure 4 the priority assigned to each task at the start of each new phase. The task manager is activated at the end of the wait windows to hand over to the decision task. At the end of its treatment, the decision task hands back to the phase manager task which starts a new phase by changing the priority of the other tasks.

5 Conclusion

In this paper, we propose a new approach for remote healthcare monitoring risk detection. Based on the concept of a multi-agent system, we present architecture which takes advantage of the technological diversity of several DSS used in this research field. We propose a real-time scheduling mechanism by using several priority levels. Our approach is interesting as it proposes the making of collective decisions between heterogeneous DSS which increases the pertinence of the final decision. The real-time aspect guaranties a necessary QoS in this kind of application, which should favor the acceptance of the remote healthcare monitoring system in industry.

References

1. McMorrow, K., Roeger, W.: The economic and financial Market Consequences of global Ageing. Springer, Berlin (2004)
2. Rammal, A., Trouilhet, S., Singer, N., Pécatte, J.M.: An adaptative System for Home monitoring Using a Multiagent Classification of Patterns. International Journal of Telemedcine and Application 2008 (March 2008)
3. Souidene, W., Istrate, D., Medjahed, H., Boudy, J., Baldinger, J.L., Belfeki, I., Delavault, F.: Une plateforme multimodale pour la télévigilance médicale. In: 4th ws AMINA, Tunis (2008)
4. Mazuel, L., Sabouret, N.: Un modèle d'interaction pour des agents sémantiquement hétérogène. In: Lip6, 16eme Journées Francophone Sur Les Systèmes Multi-Agents, Brest (2008)

5. Dokovsky, N.T., Van Halteren, A.T., Widya, I.A.: BANip: Enabling Remote Healthcare Monitoring with Body Area Networks. In: FIDJI International Workshop on Scientific Engineering of Distributed Java Applications, Luxembourg, pp. 27–28 (November 2003)
6. QuoVadis Project, http://quovadis.ibisc.univ-evry.fr/
7. El Fallah Seghrouchni, A., Briot, J.P.: Technologies des systèmes multi-agents et applications industrielles, Lavoisier (2009)
8. Dennet, D.C.: The intentional Stance. MIT Press, Cambridge (1987)
9. Rao, A.A., Georgeff, M.P.: An abstract architecture for rational agents. In: Nebel, N., Rich, C., Swartout, W. (eds.) KR-1992, pp. 439–449. Morgan Kaufmann, San Francisco (1992)
10. Rao, A.A., Georgeff, M.P.: A model-theoretical approach to the verification of situated reasoning systems. In: Bajcsy, R. (ed.) IJICAI 1993, pp. 318–324. Morgan Kaufmann, San Francisco (1993)
11. Wooldridge, M., Jennungs, N.R., Kinny, D.: The Gaia Methodology for Agent-Oriented Analysis and Design. Journal of Autonomous Agents and Multi-Agent Systems 3(3), 285–312 (2000)
12. Julian, V., Botti, V.: Developing real-time multi-agent systems. Integr. Comput.-Aided Eng. 11(2), 135–149 (2004)
13. George, L., Rivierre, N., Spuri, M.: Preemptive and non-preemptive real-time uniprocessor scheduling. INRIA, research Report 2966 (September 1996)
14. Lehoczky, J.P.: Fixed priority scheduling of periodic task sets with arbitrary deadlines. In: Proc. 11th IEEE Real-Time Systems Symposium, FL, USA, December 5-7, pp. 201–209 (1990)

MABHIDS: A New Mobile Agent Based Black Hole Intrusion Detection System

Debdutta Barman Roy[1] and Rituparna Chaki[2]

[1] Calcutta Institute of Engineering and Management
Calcutta, West Bengal, India
barmanroy.debdutta@gmail.com
[2] West Bengal University of technology
Calcutta, West Bengal, India
rituchaki@gmail.com

Abstract. A Mobile Ad Hoc Network (MANET) is a network built on ad hoc demand by some mobile wireless nodes geographically distributed over an area. There is no fixed infrastructure and also no centralized administration. The traditional security measures fail to provide the required security level in MANETs. Specialised detection schemes are to be employed for detecting the intruder attack in MANET. The intrusion detection systems based on mobile agents uses a set of mobile agents moving from one node to another node within a network. The distributed ID consists of multiple mobile agents which cooperate over a large network and communicate with each other. This as a whole reduces network bandwidth usage by moving data analysis computation to the location of the intrusion data & support heterogeneous platforms. In this paper, a mobile agent based IDS have been proposed to detect the black hole attack in MANET, with the aim of reducing computational complexity.

Keywords. Black hole, MANET, Mobile Agent.

1 Introduction

In recent years, the security concern has been of utmost importance in MANETs. The open nature of the wireless medium makes it easy for outsiders to listen to network traffic or interfere with it [3, 4, 10, and 11]. Lack of centralized controlling authority makes deployment of traditional centralized security mechanisms difficult, if not impossible. Lack of clear network entry points also makes it difficult to implement perimeter-based defense mechanisms such as firewalls. Finally, in a MANET, nodes are generally battery-powered and hence suffer from limited resources, which make the use of heavy-weight security solutions undesirable [4, 6, 10, 11, and 15].

IDSs implemented using mobile agents is one of the new paradigms for intrusion detection. Mobile agents are special type of software agent, having the capability to move from one host to another [16]. In addition, we expect an agent that inhabits an environment with other agents and processes to be able to communicate and cooperate with them, and perhaps move from place to place.

N. Chaki and A. Cortesi (Eds.): CISIM 2011, CCIS 245, pp. 85–94, 2011.

Mobile agents offer several potential advantages over static agent based IDS with respect to load reduction, platform independence, dynamic & static adoption, fault-tolerance, code size, etc. IDS employing mobile agents are however found to suffer from increased detection time. One of the major problems facing these systems is the improvement in speed of detecting malicious activities. The effective detection of autonomous attacks is still very low. Another major problem is protecting the protector (MA_IDS) from attacks.

The rest of this paper is organized as follows. Related works are presented in section 2. MABHIDS architecture is described in section 3. The implementation procedure and evaluation of the design are presented in section 4. In section 5 conclusion is presented.

2 Related Work

2.1 Review Stage

Mobile agents have been proposed as a technology for intrusion detection applications [2]. Rationale for considering agents in an IDS ranges from increased adaptability for new threats to reduced communication costs. Since agents are independently executing entities, there is the potential that new detection capabilities can be added without completely halting, rebuilding and restarting the IDS.

The following is the research that has been done in the area of MA-IDS, focusing on architecture, mode of data collection, security and their strengths and weaknesses.

DSCIDS [5] is based on a hierarchical architecture with Central Analyzer and Controller (CAC). CAC allows interactive querying by the network administrator for attack information/analysis and initiates precautionary measures and performs attack aggregation, building statistics, identify attack patterns and perform rudimentary incident analysis. The authors tested the model using different soft computing techniques which consists of neural network, fuzzy inference system, approximate reasoning and derivative free optimization techniques on a KDD cup dataset. In this approach the problem faced with hierarchical architecture is being solved by allowing free communication between the layers. The agents are not well distributed, leading to overload at certain parts of the network.

MSAIDS [7] provides a methodology where intrusion is done at two levels. The first is the Lower Level Detection (LLD), which has the data agents and processing agents. The data agents move around the nodes in the network to collect associated information. The second is the Upper Level Detection (ULD) also known as confirmation level is involved in separate intrusion detection process. At the ULD, the lower level agents gather data from the data agents and inform the Controller and Protector (CP), which acts as the Facilitator agent about the nature of the data gathered and ensures proper communication and delivery of service among agents. MSAIDS maintains security of agents by using asymmetric cryptosystem of the Aglet's framework. In addition to this, agents' states are recorded and authenticated before they are initiated. This scheme focused on (i) improving IDS performance (ii) detection of autonomous attack using its architecture (iii) Reduction in false alarm (iv) IDS agents' security. MSAIDS however fails to provide adequate security for the database, which could be vulnerable to changes by attackers.

Mobile Agent for Network Intrusion Resistance [9] framework includes a manager and a host monitor MA. The manager is the centre of controlling and adjusting other components and it maintains their configuration information. The manager receives intrusion alarms from host monitor MA and executes intrusion responses using intrusion response MA. The host monitor MA is established on every host in the network. This approach changes the hierarchical system structure of traditional distributed IDS. The location of the Manager has to be kept a secret from the intruder, as it might lead to system collapse.

[12] Proposes adaptive IDS comprising of a main intrusion detection processor, a mobile agent platform (MAP) and mobile IDS agent. The main intrusion detection processor is responsible for monitoring network segments (hosts) and collection and correlation of IDS data from distributed IDS mobile agents. The MAP resides in each host and can create, interpret, execute, transfer and terminate/kill agents for accepting requests made by network users and generating IDS mobile agents plus dispatching them into the network to do intrusion detection functions. Each host has a mobile IDs agent roaming at all times. This agent is responsible for detecting intrusion based on data gathered by sniffing on the network traffic. The mobile agents in this work are fully managed and network resources utilization is saved when there is no attack. This approach suffers from high false positive rates.

In MAIDS [13], the host monitor agent residing on every host cooperate with three subagents, namely, network detection subagent, file detection subagent and user detection subagent. If the intrusion can be determined at certain monitored host, HMA reports the intrusion directly to the manager, otherwise it asks manager for aid and it only records the suspicious activity. The manager is the center for controlling and coordinating all other components. It maintains configuration information about all components including HMA, MA platform, Assistant MA, and Response MA.It was reported that it took a long time that agents migrated with authentication and encryption though the transportation of these agents was very fast. In this scheme, the location of the manager is vulnerable to security threats.

APHIDS [14], employs a network-based architecture by placing an agent engine at every location. It is realized as a distributed layer which operates on top of a set of distributed agent engines. APHIDS architecture takes the advantage of the mobile agent paradigm to implement a system capable of efficient and flexible distribution of analysis and monitoring tasks, as well as integration of existing detection techniques. An APHID makes its Analysis Agent lightweight in order to save the bandwidth during the transfer of log data. The use of Distributed correlation scripts in capturing the expert knowledge of security administrator by automating the standard investigative procedures that are performed in response to an incident. The security of the agents is not considered.

3 Proposed Methodology

The initial motivation for our work is to address limitations of current IDS systems by taking advantage of the mobile agent paradigm. Specifically, we address the following limitations of the earlier proposed IDS False Positive Rate, Scalability, Interdependencies, and Centralized Authorization. The assumptions regarding the proposed work are listed below

3.1 Assumptions

The following assumptions are taken in order to design the proposed algorithm.

1. The source node during message sending creates a mobile agent whose life time is proportional to the maximum hop count to the destination
2. Any node in the route from source to destination can create new mobile agent if any malicious behavior is observed by it.
3. The mobile agent is called after a predefined time period.

3.2 Architecture of a Mobile Agent Based System

Figure 1 shows the modular architecture of MABHIDS. The mobile agent (MA) has to collect the raw data from the host machine then it computes the packet delivery ratio (Ri) for ith host and store it to the mobstatus table. The special agent then compares the Ri with ThR and then gives responses to the source node accordingly.

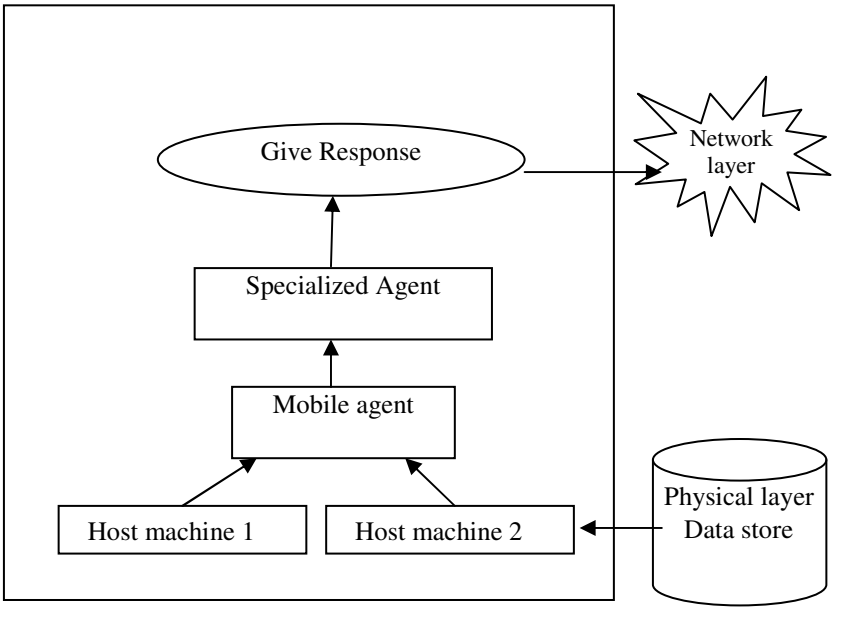


Fig. 1. Architecture of Mobile Agent

3.3 Methodology

The source node S generate the mobile agent and forward to next one hop node (A-B-C-G) then to its another child (D-E-H). The source node calculates the threshold value (ThR). In the figure 2, M is an intruder. Here M behaves as a neighbor to the destination node G by giving false RREP message to the source node against the

Table 1. Data Definition

Black_agent (S_{id} , $D_{,id,}$ $H_{c,\ ount}$ TH_R)	*The black hole detector agent*
Mob_status ($S_{id,}$ $D_{id,}$, $H_{count)}$	*Mob_status table* ,
S_{id}	*Source node ID*
D_{id}	*Destination Node ID*
H_{count}	*HOP Count*
TH_R	*Thresold value of no. of packet forwarded by the node (P_f) and no. of packet receive by the node (P_r)*
N_{id}	*Node IDof i^{th} node*
R_i	*Ratio of no. of packet forwarded by the node (P_f) and no. of packet receive by the node (Pr) for node i*

RREQ message broadcast by the source node S. In reality it has no route to the destination node G. M forward the mob agent to the node E that has no route to the destination node. The source node waits for Tout time and then creates a black agent who performs the black hole attack detection. The node S then forwards it to the route that it follows to send packet to the destination node G.

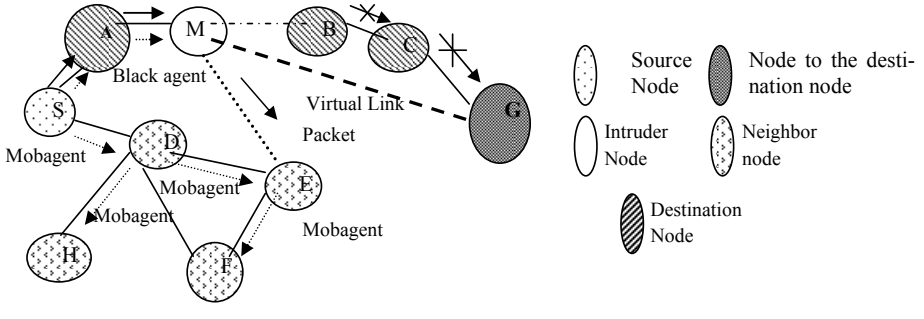


Fig. 2. Scenario of mobile agent traversing through the network with intrusion attack

3.4 Algorithm

Algorithm mobmove ( ) /* the mobmove() method is called whenever a new network is created and a source send packet to the destination */

```
The source node S is pushed into the stack
Until the stack is empty or goal node is reach
    repeat
      pop a node from stack and mark as 'CN'
      look at the route table to find next node
      call the mobagent () to perform the operation
      push the node in the stack
```

Algorithm mobagent( ) /* this mobagent () method is called from mobmove() when a new node is observed */

Observ no. of packet forwarded by the node (Pf) and no. of packet receive by the node (Pr)

. Calculate R=Pf/Pr

. Store node Id and R to mobstatus table

return to mobmove()

Algorithm threholdcalc( ) /*the source calculate the threshold value that is com pared when black agent moves to find malicious node in the network*/

set i=1, Rtotal =0

until i<no. of intermediate hops from source to destination(N)

.Read Ri from mobstatus table for ith node

.Rtotal=Rtotal+Ri

ThR = Rtotal / no. of intermediate hops(N)

Algorithm blackagent( ) /*The source node invoke this agent when destination node fails to acknowledge within Tout to the source node */

Observe Ri for ith node

if R > ThR

send MMSG(Malicious Message) to source node

else

Decrease hop count by 1.

. If hop count=0

. Terminate execution

. Call clearagent()

Call mobilitychange()

4 Performance Analysis

4.1 Simulation Metric

Packet Delivery Ratio (Pr): The ratio between the numbers of packets originated by the application layer sources (Ps) and the number of packets received by the sink at the final destination (Pd).

$$Pr = Pd / Ps$$

Node mobility (Nm) is defined as the speed with which the node is changing it's position in the network, i.e., $Pr \propto 1/Nm$

Average End-to-End Delay is defined as the average delay between the time of sending the packet by the source and its receipt at the corresponding destination. This includes the delays caused during route acquisition, buffering and processing at intermediate nodes, retransmission delays at the MAC layer, etc.

We have used NS2 ver 2.29.2 with Cygwin-1.5.21

Table 2. Simulation Parameter

Simulator	***NS2***
number of mobile node	15
number of malicious node	1
routing protocol	AODV
maximum bandwidth	2Mbps
Traffic	CBR(constant bit rate)
maximum connection	50
Maximum speed	10-100mps
pause time	5s

4.2 Performance Evaluation

Throughput of Packet Delivery: Due to the collision in the network the mobility of the node may vary. So, we have measured the throughput of packet delivery with respect to node mobility. As it can be seen from the figure 3, with MABHIDS the packet delivery is more compared to BHIDS. The node mobility is measured in MPS. When the mobility is maximum means the probability of source comes closer to the destination increases then the performance degrades in both the cases. This is because in both cases the communication overhead increases among the nodes. The distance from source to destination as well as the distance from source to adversary node becomes same.

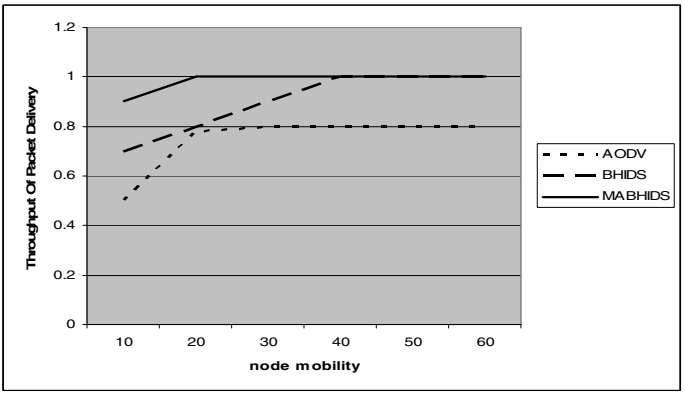


Fig. 3. Packet Delivery Ratios vs node mobility

In figure 4 the packet delivery ratios measured with respect to number of transaction the number of transaction indicates number of flows initiated during a particular duration of time from same or different sources to same or different destinations. The packet delivery ratio increases by using MABHIDS compare to BHIDS. In BHIDS the communication overhead is much more than in MAHIDS. The network is mobile in nature at the transaction 1 and 2 the source and destination comes closer to each other so their communication using mobile agent increase the overhead. The distance from adversary node and that of destination node become very closer to each other. The detection using mobile agent becomes complex, degrading the performance.

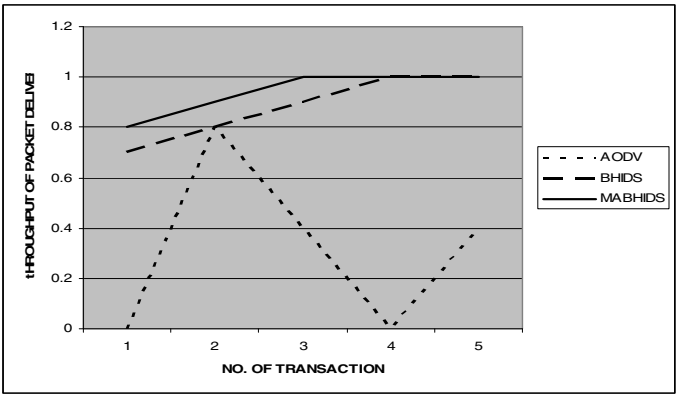


Fig. 4. Packet Delivery Ratios vs no. of transaction

End to End Delay From the figure-5 it can be observed that, when BHIDS protocol is used, there is an increase in the average end-to-end delay, compared to AODV and MABHIDS. This is due to the additional time require by each node before sending the packet to next hop. The node has to check the routing table to search next hop node. Again this is due to the immediate reply from the malicious node. i.e. the nature of malicious node here is it won't check its routing table for the route availability. In case of MABHIDS there is no route discovery overhead. That decrease the End to End Delay in MABHIDS.

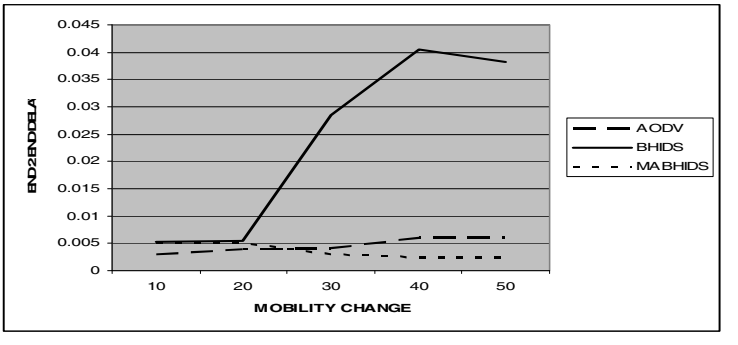


Fig. 5. End to End Delay vs. Mobility Change

5 Conclusion

The design of MABHIDS has shown that MA technology is an efficient tool for building IDS infrastructure. Since MABHIDS has shown an improvement in comparison with some previous works, multi-level intrusion detection using mobile agents has proven to be efficient. From the experiment we can conclude that when source and destination become closer to each other then the use of MA is not effective. In future, MABHIDS has to be tasted in a simulated environment with about 100 nodes and more than one malicious node. Once this has been done, the concept can be extended to the detection of sleep deprivation attacks.

References

1. Roy, D.B., Chaki, R., Chaki, N.: A new cluster-based wormhole intrusion detection algorithm for mobile ad-hoc networks. International Journal of Network Security & Its Applications (IJNSA) 1 (April 2009)
2. Onashoga, S.A., Akinde, A.D., Sodiya, A.S.: A Strategic Review of Existing Mobile Agent- Based Intrusion Detection Systems. Issues in Informing Science and Information Technology 6 (2009)
3. Chaki, R., Chaki, N.: IDSX: A Cluster Based Collaborative Intrusion Detection Algorithm for Mobile Ad-Hoc Network. In: Proc. of the 6th Int'l Conf. on Computer Information Systems and Industrial Management Applications (CISIM 2007), pp. 179–184 (June 2007)
4. Jahnke, M., Toelle, J., Finkenbrink, A., Wenzel, A., et al.: Methodolo gies and Frameworks for Testing IDS in Adhoc Networks. In: Proceedings of the 3rd ACM Workshop on QoS and Security for Wireless and Mobile Networks, Chania, Crete Island, Greece, pp. 113–122 (2007)
5. Abraham, A., Jain, R., Thomas, J., Han, S.Y.: D-SCIDS: Distributed soft computing intrusion detection system. Journal of Network and Computer Application 30, 81–98 (2007)
6. Hu, Y.-C., Perrig, A., Johnson, D.B.: Wormhole Attacks in Wireless Networks. IEEE Journal on Selected Areas of Communications 24(2), 370–380 (2006)
7. Sodiya, A.S.: Multi-level and Secured Agent-based Intrusion Detection System. Journal of Computing and Information Technology - CIT 14(3), 217–223 (2006), doi:10.2498/cit.2006.03.05
8. Mitrokotsa, A., Mavropodi, R., Douligeris, C.: Intrusion Detection of Packet Dropping Attacks in Mobile Ad Hoc Networks, Ayia Napa, Cyprus, July 6-7 (2006)
9. Wang, H.Q., Wang, Z.Q., Zhao, Q., Wang, G.F., Zheng, R.J., Liu, D.X.: Mobile Agents for Network Intrusion Resistance. In: Shen, H.T., Li, J., Li, M., Ni, J., Wang, W. (eds.) APWeb Workshops 2006. LNCS, vol. 3842, pp. 965–970. Springer, Heidelberg (2006)
10. Yang, H., Luo, H., Ye, F., Lu, S., Zhang, U.: Security in Mobile Ad Hoc Networks: Challenges and Solutions. IEEE Wireless Communications 11, 38–47 (2004)
11. Hu, Y.-C., Perrig, A.: A Survey of Secure Wireless Ad Hoc Routing. IEEE Security and Privacy Magazine 2(3), 28–39 (2004)
12. Eid, M., Artail, H., Kayssi, A., Chehab, A.: An adaptive intrusion detection and defense system based on mobile agents. In: Innovations in Information Technologies, IIT 2004 (2004)

13. Li, C., Song, Q., Zhang, C.: MA-IDS: Architecture for distributed intrusion detection using mobile agents. In: 2nd International Conference on Information Technology for Application, ICITA 2004 (2004)
14. Deeter, K., Singh, K., Wilson, S., Filipozzi, L., Vuong, S.T.: APHIDS: A Mobile Agent-Based Programmable Hybrid Intrusion Detection System. In: Karmouch, A., Korba, L., Madeira, E.R.M. (eds.) MATA 2004. LNCS, vol. 3284, pp. 244–253. Springer, Heidelberg (2004)
15. Hu, Y.-C., Perrig, A., Johnson, D.B.: Packet leashes: a defense against wormhole attacks in wireless networks. In: INFOCOM 2003, Twenty-Second Annual Joint Conference of the IEEE Computer and Communication Societies, vol. 3, pp. 1976–1986 (2003)
16. Jansen, W., Mell, P., Karygiannis, T., Marks, D.: Applying Mobile Agents to Intrusion Detection and Response. NIST Interim Report (IR) –6416 (October 1999)

Agent Based Approach for Radio Resource Optimization for Cellular Networks

Rakesh Kumar Mishra[1], Suparna Saha[2], Sankhayan Choudhury[2], and Nabendu Chaki[2]

[1] Feroze Gandhi Istitute of Engineering and Technology, Raebareli, India
[2] Department of Computer Science, University of Calcutta, India
{rakesh.mishra.rbl,sahasuparna,sankhayan,nchaki}@gmail.com

Abstract. Wireless cellular networks are used for data services like browsing, streaming, and conferencing. The increased demand for guaranteed service performance as well as for higher resources as required for bandwidth thirsty applications demand for an efficient call admission control mechanism. Most of the existing solutions targeting to meet this dipole demand follow either a predictive mechanism or some statistical approach. Predictive solutions require extensive computational load over switching centers. Further, the efficiency of solution depends on the degree of accuracy of the prediction. On the contrary, statistical approaches reduce the overhead to some extent through distributed mechanism. However, it needs synchronization among participating components. This paper presents an agent based distributed scheme for efficient management of radio resources in cellular networks such that QoS is assured for every transaction. Agents coupled to network components, like Radio Resource Manager, actively participate in network resource management. In the perspective of call admission control, agents facilitate proactive and reactive capturing requirement of each application's bandwidth. Agents also anticipate the resource requirement continually and optimize allocation of resources with fairness among the contending user applications. Besides, the proposed solution does not require any additional infrastructural and would perform better in terms of call blocking probability.

Keywords: Multi Agent System, Bandwidth pruning, Bandwidth Optimization, Service plane, Network Plane, Resource Plane, Radio Resource Management, QoS.

1 Introduction

QoS implies service performance that determines the degree of user satisfaction for service. Increasing demand and limited bandwidth available for mobile communication services results in higher call drop and blocking rates and thus requires efficient radio resource management [1]. Preserving the interest of the service provider and meeting the expectations of the users is the delicate balance that any call admission policy has to consider [2].

Presently call admission control (CAC) has become prominently a problem of Queuing Models, Markov Chains or k-Erlang Models [3] but these usually induce extensive computational overheads at switching centers.The another class of

N. Chaki and A. Cortesi (Eds.): CISIM 2011, CCIS 245, pp. 95–104, 2011.

solutions, called statistical [5], reduces the computational overhead up to a certain extent but demand synchronization and additional hardware for storing the sampled data. [3,4] Predictive and statistical approaches add overheads accounting towards data sampling for different metrics, caching and performing the analysis. This in turn demands additional resource like cache, processing units etc. for its accomplishment. Thus, an approach involving lesser computation and resource overheads, and with no dependency on historical data but with same level of expected performance will be most beneficial. A truly adaptive solution responding to the prevailing situation ensuring the minimum QoS with respect to each concurrent application in terminal equipment will be more practical and viable solution. In subsequent section related work will be discussed, in section 3 discusses architecture of the proposed system. Section 4 describes the working principle of the solution extensively. Merits and future works are mentioned in section 5.

2 Related Work

[6] introduces two CAC schemes for GSM like TDMA networks based on the load factor, a ratio of total active calls with total available channels. In first proposition calls are being admitted based on the loading factor of neighbouring cells. Second proposition defines a threshold loading factor for taking decision by averaging the spatial load over the interference groups. These approaches involve exhaustive signalling, to be provisioned at basic protocol to support the computation of loading. [7] are the Guard Band schemes where small portion of bandwidth is dedicated for handover. This technique reduces call dropping probability but may incur high call blocking probability. [8] modifies the guard band technique by provisioning shared band within guard band. The dedicated bands are defined for each class of calls with shared band such that on the advent of congestion shared band can be utilised for the allocation. [9] improvises guard band technique by allowing flexible boundaries those can be stretch to other lesser loaded band on congestion. [10] performs the reservation on the basis of availability of mobility information.

Bandwidth borrowing is process wherein a congested cell hires bandwidth from its neighbour or from pre-emption to support call admission. In [11] non interfering channel is borrowed such that it does not interfere with existing calls. This approach leads to restriction of channel usage and causes exhaustive search for locating suitable channel. [12] acquires the bandwidth by making active call to compromised with quality or pre-empt to support new call. [13] introduced the concept of bandwidth reservation. Bandwidth if can be reserved for the call then only call will be admitted. This technique leads to resource wastage if mobility is not assured. [14] extends the reservation concept to guaranteed QoS by only admitting a call when the required spare bandwidth is available for allocation. In this approach guard boundaries were not strictly followed for handover requests. Handoff queuing is suggested in [15] wherein if the target cell does not have adequate bandwidth then a call it will be queued till channel become available. As channel is vacated call is anchored to the channel thus reducing call drop probability. [16] is case of QoS negotiation wherein application if fails to avail desired bandwidth reduces its quality metrics and reattempts for channel acquisition. [17] is the variant of a scheme which works on the principle of assuring minimum transfer rate is guaranteed. Number of users is

restricted in totality or for a class of service to ensure minimum transmission rate for each active call.

Most of the solutions except the guard band policy are more complex because of the involved computation and signalling (channel hiring/reservation). None of the solution are adaptive to the load scenario i.e. a call do not compromise its quality to accommodate a new call. Thus these approaches could not negotiate for quality between the concurrent running applications and the newly admitted call. QoS negotiation can be useful because underestimating QoS will result in waste of scarce resource and overestimating QoS will result in unexpected ongoing call drops ie there always a leverage to comprise with QoS to some extend. [16][17] are negotiating for resource allocation for new call by compromising with the quality of the call itself or restricting number of applications. Quality negotiation is done through repeated computation with modified quality metrics and thus incurs high computational overhead. But, in our proposed solution existing calls compromise their quality within limits to accommodate a new call. A reactive multi-agent system (MAS) takes decision according to instantaneous demands while adaptively adjusts itself to the network load with minimum guaranteed QoS but without extensive computation.

3 Proposed Architecture for Resource Optimization

The proposed solution for bandwidth optimization spans from user equipment side to the network switching centers. All application lying in the purview of the solution are categorized into four classes namely – Realtime–Interactive (RT-I), Realtime–Non–Interactive (RT-NI), Non–Realtime–Interactive (NRT-I) and Non-Realtime-Non-Interactive (NRT-NI). Applications in each class are confined within definite performance bounds called Optimal and Sustainable Level. Optimal level of an application has highest requirement of resource and guarantees best performance whereas the sustainable level has least amount of resource needed for running an application with compromised but acceptable quality. Resources are allocated to applications from two disjoint pools called as Local Pool and the Global Pool controlled by the BSC and MSC respectively.

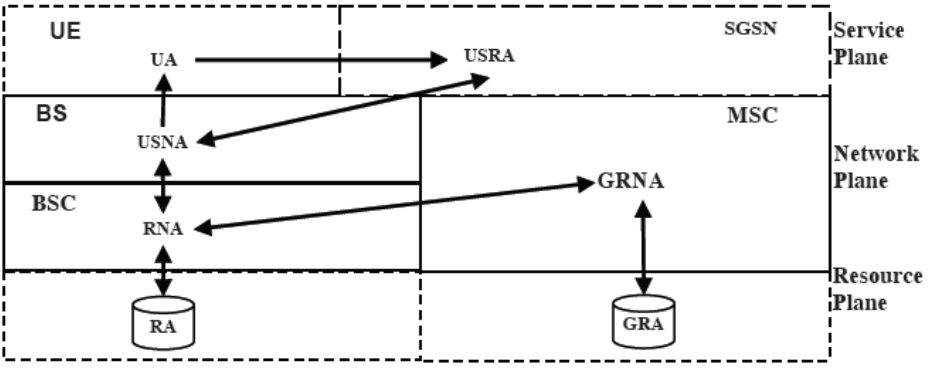


Fig. 1. (UA – User Agent, USRA – User Service Request Agent, USNA – User Service Negotiation Agent, RNA – Resource Negotiation Agent, GRNA – Global Resource Negotiation Agent, RA – Resource Pool, GRA – Global Pool)

Generally, lower precedence applications are being deprived severely for making provisions for higher precedence applications. Fairness needs to be maintained such that for each concurrent application a definite QoS is maintained irrespective of the class of the application. The main objective of the solution is to overcome the biasness in distribution of resources. In the solution resource requirement of each active call and new are assessed. In case of a new call if resource could not be allocated from local and global pool then all the active applications needs to release a small amount of resource from their occupancy to accommodate a new call. The resource will be released only by applications having resource over and above the sustainable level. The active compromised applications will also be periodically submitting their demands for achieving optimal level which will be honored based on the availability of resource. Resource redistribution is being done in an optimal way such that all application belongs within their predefined bounds.

Proposed architecture (Fig – 1) has been divided into three layers namely service plane, network plane and resource plane. Service layer contains active services and is comprises of two parts each at User Equipment (UE) end and Service GPRS Support Node (SGSN). UE side of service plane is a collection of instances of application of various classes. SGSN part of service plane comprises of attributes of each running application instance for the specific UE. This is also performing the computation for optimum acquisition and release of resource.

Network plane comprises of the three constituent sub-planes and each is co-located with the functional component of a network like Base Station (BS), BSC and MSC. BS part of the network plane is responsible for posing the request and communicating the resource allocation to service plane components. BSC part of network plane either allocates the resource or request to MSC. In case both plane fails to allocate resource for the call then BS part will be instructed to release a part of resource with minimum level of QoS is maintained. Resource plane is a passive plane comprising of local and global components.

4 Multi Agent System

In this section we have described the role of agents for achieving the proposed solution described above. All the agents are static in nature and the activity of each agent is described here through following algorithms. (See fig-1)

4.1 User Agent

UA is located at each UE, which is a hand held device or a computer with wireless network interface. The services may be executing from same or different environment, for example while browsing a site (NRT–NI) user invokes a video link (RT–NI) which may start within same environment or in altogether new environment. UA communicates the class of the service to USRA using message new service class. The algorithm for the User Agent is shown in Fig – 2. UA will receive message safe

after resource is allocated to the service. If no response is received by UA after N-tries, service will be dropped.

```
Algorithm: user_agent_UA
Parameter: service_id, class, recvd_msg
 if started(service_id) or service_chng(service_d) then
      for attempt 1 to N do
          send message to USRA for new service and its class
          if no response from USRA before timer expires then
      increase attempt
      done
 if recvd_msg is safe message from USNA then /* safe message implies the
 resource is allocated*/
      start_service( service_id)
 else
      drop_service(service_id)
end
```

Fig. 2. Algorithm for UA

4.2 User Service Request Agent

USRA is residing at SGSN (Serving GPRS Support Node) or any other physical or logical component where the service request from mobile node is granted / managed maintained. It is a computation unit for resource requirement. Service database, maintained with SGSN for keeping track of flows and flow parameters, is accessible to USRA. USRA caches the service set of its domain pertaining to a user from service database. As new service is solicited User Agent communicates to service and its class to USRA which in turn attempts for resource allocation i.e. bandwidth. Service information will be communicated for a new service or an old service which has changed its resource requirement. USRA periodically submit differential resource requirement to maximize its resource share to improve QoS of the services. USRA also prune resource from its service set to facilitate CAC. The operation model is detailed in form of Equations 1-3. Equation -1 is for maximizing the resource share for a UE, similarly Equation-3 is for minimization of resource release during the pruning process. Equation-2 is the minimum increment that is acceptable to the UE for upgradation of its QoS requirement. If ki is the number of active service of class-i, xi is the additional resource requirement / deduction for service class-i, Cmax-i and Cmin--i are maximum and minimum threshold for resource allocation for service class-i, ai is a amount of resource actually now being allocated for service class-i, C_{total} is the total available resource. The algorithm for USRA is shown in fig-3.

$$\text{Max} \sum k_i * x_i$$
Subject to
$$a_i + k_i * x_i < k_i * C_{max\text{-}i}$$
$$\sum (a_i + k_i * x_i) \leq C_{total}$$
$$a_i > 0$$
$$x_i \geq 0$$

Equation 1

$$\text{Min} \sum k_i * x_i$$
Subject to
$$\sum (a_i + k_i * x_i) \leq C_{total}$$
$$a_i + k_i * x_i > k_i * C_{min\text{-}i}$$
$$a_i > 0$$
$$k_i * x_i \geq a_i$$

Equation 2

$$\text{Min} \sum k_i * x_i$$
Subject to
$$a_i - k_i * x_i \leq k_i * C_{max\text{-}i}$$
$$a_i - k_i * x_i > k_i * C_{min\text{-}i}$$
$$a_i > 0$$
$$x_i \geq 0$$

Equation 3

Algorithm: user_service_request_agent_USRA
Parameter: recvd_msg

```
while network is active do
  if recvd_msg is a new service message from UA then
      identify the resource demand according to its class
      send message for new service to USNA with demand
      reset timer
  if recvd_msg is a pruning message from USNA then
      compute resource contribution using equation 3
      send pruned resource message with released quantity to USNA
  else if timer expires and no message either from UA or USNA then
      compute max_optimal and min_optimal from equation 1 and 2
      send message differential demand to USNA to increase service
      resource allocation along with computed values
done
end
```

Fig. 3. Algorithm for User Service Request Agent

4.3 User Service Neotiation Agent

USNA is service negotiation agent, residing at BS. It is a network plane entity for USRA and is negotiating the optimal resource allocation from RNA. It can communicate with UA, RNA (resource pool negotiator) and USRA. USNA informs the UA for the allocated of resource for new service or additional resource for optimal performance of the service, otherwise will be informing about the resource reclaimed for call admission. USRA informs resource requirement for active services and new service to USNA.

In case the resource plane entities fails to allocate required resource for new service USNA will not revert back to UA. RNA, whenever fails to provide the resource USNA immediately request for reallocation to RNA to be passed-on to GRNA. In case of failure allocation of differential request then USRA will continue with its active services and try for differential demand in the next time period. When ever the resource is acquired by USNA from RNA or GRNA or both USNA always informs UA of acquisition through message safe containing the new resource value. Whenever USNA identifies that a new service is about to be abort due to

non-availability resource it raises message pruning to RNA which in turn sends message prune to all the USNA associated with it. Algorithm for USNA is detailed in fig-4.

Algorithm: user_service_negotiation_agent_USNA
Parameter: recvd_msg
if recvd_msg is a prune message from RNA then
send message start pruning to USRA
if recvd_msg is a pruned resource message from USRA then
send message released resource to RNA
send safe message to UA with new resource values
if recvd_msg is a diferenrtial demand message from USRA then
obtain max_resource and min_resource values from message
send new service class message to RNAwith required resource quantity
if recvd_msg is a allocated resource from RNA then
if the allocated value is larger than min_resource then
send message safe with resource quantity to UA
if recvd_msg is a new service demand message from USRA then
send allocate resource message to RNA for resource demand
if recvd_msg is a allocated resource from RNA then
step 1: if allocated resource is less than sustainable threshold value then
send message reallcate to RNs for more resource.
else
send safe message to UA with resource quantity
if recvd_msg is a reallocated resource from RNA then
if resource allocated in step-1 and now do not satisfy sustainable level
thresholdthen
send message start pruning to RNA
else
send message safe to UA
end

Fig. 4. Algorithm for User Service Negotiation Agent

4.4 Resource Negotiation Agent

RNA will be handling request from several USNA and situated at BSC. RNA is network side entity of the architecture and belongs to network plane. USRA is a computational component performing necessary computations for resource allocation to various USNA, while USNA is a communicating component.

USRA informs resource requirement RNA for the resource allocation and accumulation to resource pool after pruning. RNA also collects the differential demands and optimizes the allocation for each requesting USNA by keeping the allocation to individual USNA to minimum but above sustainable level (equation -2). In order to improve the responsiveness the new service demands are immediately catered. The algorithm of RNA is detailed in Fig-5

```
Algorithm: resource_negotiation_agent_RNA
Parameter: recvd_msg
   if recvd_msg is a pruning message from USNA then
      for each USNA
         send message prune to each USNA
   if recvd_msg is a release resource from USNA then
      if local pool is vacant then add to local pool
      else send message add_to_global to GRNA
   if recvd_msg is a new service message from USNA then
      send message allocated resource to USNA with quantity
   if recvd_msg is a reallocated message from GRNA then
      send allocate_global message to GRNA for resource allocation
   if recvd_msg is a allocated message from GRNA then
      send allocated resource to USNA with quantity
   if recvd_msg is a differential demand message from USNA then
      accumulate request for timer interval
      allocate the resource as per the equation 2 and distribute evenly among USNA
      for each USNA
         send message allocate resource to each USNA
end
```

Fig. 5. Algorithm for Resource Negotiation Agent

4.5 Global Resource Negotiation Agent

This agent is analogous to RNA with a difference that it will be doing the optimization of resource for RNA(s). Global pool is for the new call when local pool fails to allocate resource. Global pool will be replenished after the local pool is completes. Optimization function will minimize the allocation to each request such that cumulative demand remains within available resource quantity. Algorithm of the GRNA is detailed in Fig-6

```
Algorithm: global_resource_negotiation_agent_GRNA
Parameter: recvd_msg
   if recvd_msg allocate global message from RNA and timer exists then
      Store recvd_msg from RNA
   if timer expires then
      compute allocation for the requests using equation 2 and
      distributed evenly among RNA(s)
      for each recd_msg from unique RNA do
         send allocate global message for allocated resource
      done
   if recd_msg is add_to_global then
      add resource to global pool
end
```

Fig. 6. Algorithm for GRA

5 Design Merits and Discussion

In this paper we have proposed an overlay agent framework for radio resource optimization. The frame work is focus towards the fairness of resource allocation at UE end with guaranteed minimum QoS. The framework also provisioned for call admission wherein active call donate bandwidth to accommodate new service / call. As design is based on Agents thus has no relative impact on the underlying communication standard consequently there is no as such infrastructural upgradation is required.

Proposed framework is hybrid architecture it include the efficiency of centralized system and the simplicity of distributed approaches. The decision is not taken solely by the hierarchical heads but equally distributed among other components. Agents at service plane defines what is the requirement limits while Network plane agents define what will be allocated this way the decision is jointly taken at service and network plane. Architecture is event driven and the communication between the participating units happens to be through messaging passing. Message passing eliminates need for synchronization between the peer components of the solution.

Proactive assessment for resource is complimented by the reactive adjustment of bandwidth of active applications between the prescribed limits to accommodate a new service call. Further, agents auto adjust bandwidth at lesser load to improve QoS. To make the framework more relevant different class of application and the QoS requirement is kept at the core of design and operation.

The resource allocation by provisioning global pool is identical to bandwidth hiring with a difference that interference problem is tackled by MSC hence avoiding costly signaling for frequency redistribution. Optimization function are have constant complexity and independent of the number of application actives as the allocation is done with respect to the class thus largest objective equation has 4 variables and atmost 6 constraints. In this proposal we succeed to keep complexity consistent, fairness is also assured as at none of instant QoS beyond lower threshold is compromised and no amendment is required either at protocol level or hardware level.

References

1. Fang, Y.: Thinning algorithms for call admission control in wireless networks. IEEE Transactions on Computers 52(5), 685–687 (2003)
2. Chen, D., Elhakeem, A.K., Wong, X.: A Novel Call Admission Control in Multi-Service Wireless LANs. In: Proc. of Third International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks, pp. 119–128 (2005)
3. Ghaderi, M., Boutaba, R.: Call Admission Control in Mobile Cellular Networks: A Comprehensive Survey. Wireless Communications and Mobile Computing 6(1), 69–93 (2006)
4. Naghshineh, M., Schwartz, M.: Distributed Call Admission Control in Mobile/Wireless Networks Personal, Indoor and Mobile Radio Communications. In: Proc. 6th IEEE Int'l. Symp. Pers., Indoor and Mobile Radio Commun. (PIMRC 1995), vol. 1, pp. 289–293 (1995)

5. Abdulova, V., Aybay, I.: Predictive mobile-oriented channel reservation schemes in wireless cellular networks. Springer, Science+Business Media, LLC 2010 (Published online: August 29, 2010)
6. Beming, P., Frodigh, M.: Admission control infrequency Hopping GSM System. In: Prc. Of Vech Tech. Conf. 1997, pp. 1282–1286 (1997)
7. Hong, D., Rappaport, S.: Trafffic Model and performance Cellular Mobile Radio Telephone System with Prioritized and Non-Prioritized Handoff Procedures. IEEE Trans. Vech. Tech. 35(3), 77–92 (1986)
8. Haung, Y., Ho, J.: Distributed Call Admission Control for a Heterogeneous PCS Network. IEEE Trans. Comp. 51(12), 1400–1409 (2002)
9. Yu, O., Leung, V.: Adaptive Resource Allocation for Prioritized Call Admission Control over a ATM based wireless PCN. IEEE JSAC 15(7), 1208–1225 (1997)
10. Bartolini, N., et al.: Improving Call admission Control Procedures by using Handoff Rate Information, Willey Wireless and Mob. Comp. J., 257–268 (2001)
11. El-Khadi, M., et al.: A Rate Based Borrowing Scheme for QoS Provisioning in Multimedia Wireless Network. IEEE Tran. Parallel and Distributed System 13(2), 156–166 (2002)
12. Chang, C.J., et al.: A channel Borrowing Scheme in Cellular Radio Syatem with guard channel and finite queues. In: Proc. of IEEE ICC 1996, vol. 2, pp. 1168–1172 (1996)
13. Kim, J., et al.: An Adaptive Bandwidth Reservation Scheme for High Speed Multimedia Wireless Network. IEEE JSAC 15(6), 858–872 (1998)
14. Zhao, D., et al.: Call Admission Control for Heterogeneous Services in Wireless Network Communication. In: IEEE Int'l Conf. on Comm., vol. 2, pp. 964–968 (2000)
15. Haleem, M., et al.: Fixed Wireless Access System and Autonomous Resource Assignment. In: Proc. of PIMRC 1998, vol. 3, pp. 1438–1442 (1998)
16. Psychis, S., et al.: Call Admission Control traffic Policing Mechanism for the Wireless Transmission of Layered Video Conferencing Traffic from MPEG-4 and H.263 Video Coders. In: PIMRC 2002, vol. 5, pp. 2155–2159 (2002)
17. Naghshineh, M., Acampora, A.: QoS Provisioning in Micro Cellular Network Supporting Multimedia Traffic. In: Proc. 14th Ann. Joint Conf. IEEE Comp. and Comm. Societies, vol. 3, pp. 1075–1084 (1995)
18. Naghshineh, M., Acampora, A.: Design and Control of Micro Cellular Network with QoS provisioning for Data Traffic. Wireless Network 3(4), 249–256 (1997)

Influence of Database Quality on the Results of Keystroke Dynamics Algorithms

Piotr Panasiuk and Khalid Saeed

AGH University of Science and Technology
{panasiuk,saeed}@agh.edu.pl

Abstract. In this paper authors present a general review of the history of keystroke dynamics introducing some selected known approaches and bring attention to the modern trends in this domain of biometrics. Although there is a large number of works on this subject, it appears they are incomparable. That is why authors pay a large attention to testing such algorithms with other databases. Many unique tests were conducted and many observations were described. The influence of using varieties of databases on keystroke dynamics is given in detail in this paper. Important conclusions on databases and training set selections are presented.

1 Introduction

We live in the XXI century and the world we know has been overwhelmed by computers. With the expansion of the Internet and development of new social networks, there is plenty of sensitive data circulating in the worldwide network. Many researchers are looking for ways to improve the data security. However many of the solutions are either invasive for the user or they are very expensive.

In general security methods can be classified on the basis of: something that we know, such as passwords or PIN codes; something that users possess, e.g. chip or smart cards, magnetic cards or the simplest example - the keys; or biometrics, which are based on the physical characteristic features or the behavior of the person. The most popular method of securing digital data is the password dependent. However, recent researches have shown that many users select very simple and obvious passwords to protect their privacy or even money, for example "password", "123456", "qwerty", the user's name, birth date or the combination of those which makes them less secure. Chip or smart cards do not provide high security too because they can easily be stolen. The third group of methods can be used in case of verification or identification of the user.

Biometrics is a science about measuring features of living organisms. During the past few decades there was a noticeable increase of biometrics popularity, especially in the data security domain. Measured features can be divided on the basis of their origin. They could be both physical as well as behavioral. Physical features are those that derive from how our body is built. The most popular physical feature known for almost everyone is fingerprint. Behavioral features,

N. Chaki and A. Cortesi (Eds.): CISIM 2011, CCIS 245, pp. 105–112, 2011.

on the other hand, are those, that derive from how we perform some activities. There may also occur someone without a biometric feature and in that case multi-biometric systems are the desirable solution.

2 General Background

2.1 Beginning

Keystroke dynamics is a behavioral biometric feature that describes one's typing pattern. It is similar to the handwritten signature but using computer keyboard. The history of this method starts with the invention of the telegraph and its popularization in 1860s. Soon the telegraph operators were so experienced in sending "dot" and "dash" signals that they could recognize their co-workers basing on their way of sending messages. The next stage began when personal computers were popularized. With the ease of gathering data and large amount of different keys known from typewriters, the full potential of keystroke dynamics was quickly revealed in password hardening and increasing protection of computer systems.

First information about the possibility of using keyboard to identify people was published in 1975 by RJ Spillane [10], who brought further researchers attention to keystroke dynamics. However, as the actual birth date of the keystroke dynamics, it is considered the publication of a report made in 1980 by Gaines, Lisowski, Press and Shapiro for RAND Corp. [2]. Their research goal was to verify whether an individual can be identified by the statistical typing characteristics. The samples for the experiment were acquired from seven professional typists. The obtained false rejection rate (FRR) was 5.5% and 55% for false acceptance rate (FAR).

Results of predecessors encouraged other scientists for research in the keystroke dynamics area. Among the most important we should mention [5] where Leggett and Williams in 1988 described their first experiment on user verification using non-fixed text. It means that users were working normally typing their own unique datasets, without any imposed phrase. One year later, in 1989 Hussien et al [3] used Artificial Neural Networks in their approach for the first time.

2.2 Now

Latest researches focus in general on the user verification to secure personal computers. There are a few works on user identification. The most common approaches used in classification are neural network algorithms whose main disadvantage is their high dependence on the training database and also time cost of the training process.

The researchers started to look for some new characteristics like keystroke pressure. In fact [9] shows that it is even more significant than the dynamics itself. The main problem is that in the real world we do not use pressure sensitive keyboards. The pioneers in this variation of keystroke dynamics were scientists

from Taiwan [6] who made their own keyboard and carried out some experiments using typing force as an additional information. It turned out to be very helpful in the verification process.

Another chance for developing keystroke dynamics is rapidly expanding touchscreen devices market. Smartphone screens are not as convenient to use as physical full-size PC keyboard, however they allow the device to emulate pressure sensing with scanning the area of the screen being under the finger.

3 Databases

3.1 General Classification

Algorithm testing results highly depend on the quality of the database. Of course, it does not mean selecting the samples to get better results. Good algorithms should also deal with "bad" samples, which means the ones with mistakes or random pauses in user typing. Among the databases we can distinguish those collected in a supervised way. That means every test subject is individually instructed by a supervisor before the start of the samples acquisition. The supervisor can also make notes on how the subject types and what may have the influence on his pattern. It guarantees well described samples. However, this type of database usually does not reflect real world situations. On the other hand, databases that depend only on users honesty may have duplicated accounts. This happens when the user forgets his password or just wants to have two accounts. The typing pattern may be duplicated under two different classes, what may decrease identification accuracy and in hybrid (rank-threshold) based verification method may increase the FRR. Another difference between keystroke dynamics database collecting procedures is the purpose the database is gathered for. When we want to conduct research on user verification we should keep a user identifying token like ID and password assigned to it. In case of testing an algorithm for user identification, the samples should be longer phrases.

What is more, there can be two additional approaches to data acquisition. The first bases on a fixed text. This means that every user has to insert the same phrase. The second way is to use continuous authorization.

3.2 Maxion's Database

Authors have recently found an extremely interesting works of Roy Maxion from Carnegie Mellon University, who summarizes all major efforts in keystroke dynamics. Among algorithms comparison there is also a great concern on databases used along with them. The main problem is that all those databases are not comparable in any way because of all these mentioned above issues.

Another issue is that the event timing may be affected by clock process queuing. Maxion examined this concept by using function and arbitrary waveform generator [4]. He noticed that 18.7 % of keyboard events were registered with 200 us latency. However, that seems not important as while using typical PC we lose a lot of precision due to the operating system event clock which is limited to accuracy of 15.625 ms (64 Hz) on Windows and 10 ms on Linux.

Considering the constraints described above, Maxion has developed the database that is very accurate and have a lot of samples. This allows to conduct many experiments. The database is available online at [11] and is free of charge.

3.3 Authors' Database

Authors data were gathered in non-supervised conditions with the web-browser platform [13]. Authors intention was to create a universal, globally available and operating system independent application using the most popular technologies (HTML, JavaScript), so that everyone with internet access, anywhere in the world can contribute to the research [8]. This platform operates in two language versions: Polish and English. However, because of insignificant contribution to the English version, in this approach we use only Polish version samples.

In the database we have already over 400 users and over 1500 samples stored. Technology used in creating this system can simulate real, low-budget solution due to using popular and free technologies like JavaScript and PHP.

4 Authors' Method

This paper describes authors' identification algorithm operating on a fixed-text samples. It is quite simple and fast created to be able to analyze huge amount of data in short time. Authors' latest approach has been developed to be able to calculate initial weights for all expected key events. However, during tests with Maxion's database it revealed that better results can be obtained with our primary algorithm [8], so we returned to it. The algorithm removes all samples with errors so that we can to compare the results obtained with our database with those produced using Maxion's data. As we use kNN classifier, some initial values of the parameters are necessary to set up first. We choose the k value and then build a training dataset where the amount of samples per each user cannot be less than k. The remaining user samples are assigned to the testing dataset, so the user has to reach the limit of $k+1$ samples. If this condition is not satisfied, the user is not taken into account.

The next step is the classification. During this stage the distance to all training samples is calculated for each test sample using Manhattan metrics between corresponding keyboard event times (1). Additionally, times are divided into flight times and dwell times. Flight times are the times between releasing one key and pressing the another. Dwell time, on the other hand, is the time when a key is in a pressed state. The reason we convert simple event times into those two characteristics is because they are more stable. When the user makes a mistake or hesitates on some key, this only affects the two following keys and not all the remaining times. Both of those times are equally relevant and, if we take only one of them into account, the rate of proper identification decreases dramatically [8].

$$dist(A,B) = \sum_{i=1}^{k} |dwell_{Ai} - dwell_{Bi}| \sum_{i=1}^{k-1} |flight_{Ai} - flight_{Bi}| \qquad (1)$$

After the calculation of all the distances between the given sample and all the training samples, we have $k*users$ results marked with training author ID. This is an example of a rank system. We are experimenting on a closed-world basis, which means we do not consider users from outside the system. Taking into account impostors from the outside world can be solved using hybrid approach (both rank and threshold basis). Among all the results we take the k best ones and then conduct voting procedure on users. The shortest distance gets the highest weight of k, the longest distance gets the lowest weight of 1. If there is more than one vote on a particular user, we merge them adding weights, so at the end there is only one vote per one user. The winner is the user with the greatest weight.

5 Results Comparison

While obtaining the results with our algorithm, we tried each of our approaches and many combinations of initial parameter values. At the beginning it was tested with the authors' newest method, which calculates the weights of each keyboard event basing on Fishers discriminant before classification. Those weights are later used in the calculation of the distances between test and training samples. However, this method gave worse results on Maxion's database than our simplest solution. The experiments were repeated with different methods of training samples selection. This revealed interesting correlation between algorithms accuracy and sample acquisition time.

For the comparison we also tested this algorithm on our database after its proper preparation. For this experiment we used the same amount of users in both databases, the same number of characters per phrase (decreased to the phrase "kaloryfer"), the same amount of training samples and the same algorithm. Below, in Fig. 1 we can see the results of this comparison. In both experiments training data sets were created using random samples. Also all the parameters were the same for both databases, k=2, training set containing 6 samples per each user and 51 classes. The flight times and dwell times were considered as equally important.

Figure 1 shows that even if we have databases with the same data amount, the results will be different. It proves that the results are not comparable even if we test the algorithm on similar databases. This leads us to the conclusion that all the results obtained by any research team on their databases are meaningless if we cannot test other algorithms on the same databases. What is also worth noticing is that the KDS is a database collected under non-supervised way and with less precise technologies as Maxion's, so it should provide worse results.

The next experiment was conducted on artificially expanded samples - two samples of one user were merged together. Effectively, the new samples repeat all the keystrokes twice. This operation decreases the number of samples by half but extends their length. This approach was conducted only on Maxion's samples. As a control group we also took a longer phrase from KDS database (phrase no.5 - an English password from Psylock system) and reduced its length respectively

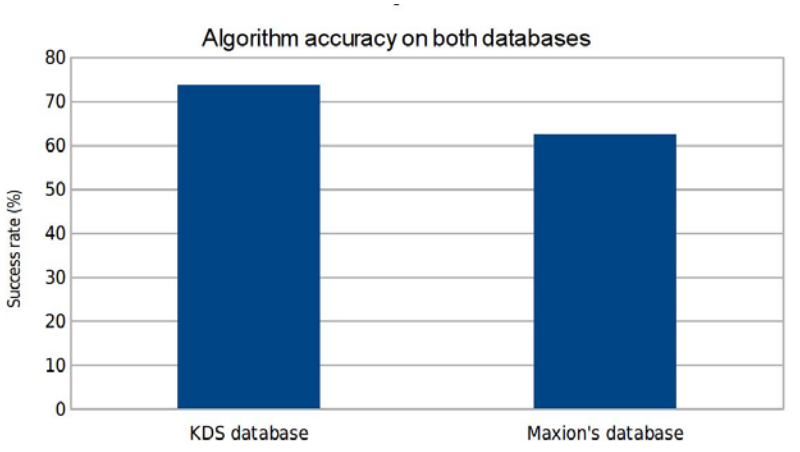


Fig. 1. The results of experiments carried out using authors identification algorithm on both databases under the same circumstances

to 10 and 20 characters, so this phrase in 10 character version was "After some" and in 20 character version "After some considera". The results are presented in Fig. 2.

With Maxion's samples the accuracy increased by 3.6% of the initial value. In the case of KDS samples it was 5.6% of the initial value. The increase of accuracy can be explained by doubling the number of the information in both databases. On the other hand, the reason of such a difference between them may be caused by the fact that in the case of KDS the last 10 characters are different, so it brings more information about the user's pattern.

6 Conclusions

Summarizing, there has been a lot of research done since the beginning of the 1980's. Researchers are trying to find new methods to improve keystroke dynamics accuracy, however, all of their obtained results are incomparable due to the use of different, custom and not publicized databases.

Keystroke dynamics itself cannot be used for forensic purpose, because it does not meet the European access control standards such as EN-50133-1 [1]. However, if one includes other features of typing, the keystroke dynamics results improve. A good idea is also to analyze the pressure of the keystroke. There should be further research made to look for additional features that could describe person's typing pattern.

Along with researchers who invented pressure analysis, the contribution of Maxion to keystroke dynamics analysis, with his comparison of known classifiers on the unified database and his great database itself, shall be widely appreciated. The authors think that all researchers shall test their algorithms using this

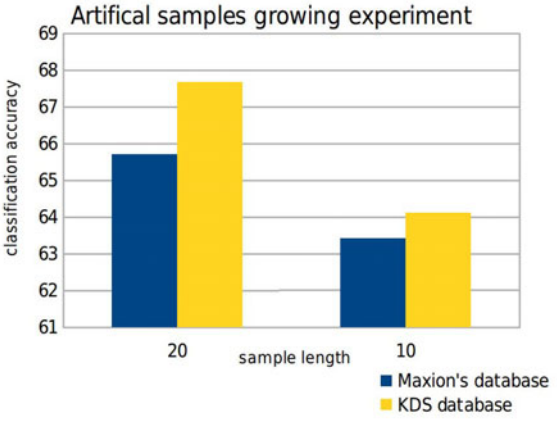


Fig. 2. The influence of artificial sample expanding on the classification accuracy of the authors' algorithm

database because it allows to obtain reliable and comparable results. Also the researchers can download a keystroke dynamics database that includes pressure information [12].

The keystroke dynamics will rather not give satisfying result itself, unless it is merged with some other biometric systems. Especially these, which neither reveal their activity nor bother the user. An example of multimodal system may be the keystroke dynamics and face image recognition system checking the user identity, while the user inserts his PIN number at the ATM. The authors believe that the multi-biometric systems are the future.

Acknowledgement. This work was supported by grant No. 11.11.2010.01 AGH University of Science and Technology in Krakow and partially by WSFiZ in Elk.

References

[1] CENELEC. European Standard EN 50133-1: Alarm systems. Access control systems for use in security applications. Part 1: System requirements, Standard Number EN 50133-1:1996/A1:2002, Technical Body CLC/TC 79, European Committee for Electrotechnical Standardization, CENELEC (2002)
[2] Gaines, R.S., Lisowski, W., Press, S.J., Shapiro, N.: Authentication by Keystroke Timing: Some Preliminary Results. Rand. Corp. Santa. Monica, CA (1980)
[3] Hussien, B., McLaren, R., Bleha, S.: An application of fuzzy algorithms in a computer access security system. Pattern Recog. Lett. 9, 39–43 (1989)
[4] Killourhy, K.S., Maxion, R.A.: The Effect of Clock Resolution on Keystroke Dynamics. In: Lippmann, R., Kirda, E., Trachtenberg, A. (eds.) RAID 2008. LNCS, vol. 5230, pp. 331–350. Springer, Heidelberg (2008)
[5] Leggett, J., Williams, G.: Verifying identity via keystroke characteristics. Int. J. Man-Mach. Stud. 28(1), 67–76 (1988)

[6] Loy, C.C., Lai, W.K., Lim, C.P.: Keystroke Patterns Classification Using the ARTMAP-FD Neural Network. In: Intelligent Information Hiding and Multimedia Signal Processing, Kaohsiung, China, pp. 61–64 (2007)
[7] Panasiuk, P., Saeed, K.: A Modified Algorithm for User Identification by His Typing on the Keyboard. In: Chora, R.S. (ed.) Image Processing & Communications Challenges 2. AISC, vol. 84, pp. 113–120. Springer, Heidelberg (2010)
[8] Rybnik, M., Panasiuk, P., Saeed, K.: User Authentication with Keystroke Dynamics Using Fixed Text. In: 2009 International Conference on Biometrics and Kansei Engineering, Cieszyn, Poland, pp. 70–75 (2009)
[9] Saevanee, H., Bhattarakosol, P.: Authenticating User Using Keystroke Dynamics and Finger Pressure. In: 6th IEEE Consumer Communications and Networking Conference, CCNC 2009, Las Vegas, NV, pp. 1–2 (2009)
[10] Spillane, R.: Keyboard Apparatus for Personal Identification. IBM Technical Disclosure Bulletin 17(3346) (1975)
[11] http://www.cs.cmu.edu/keystroke/ (state from: June 19, 2011)
[12] http://jdadesign.net/2010/04/pressure-sensitive-keystroke-dynamics-dataset/ (state from: June 19, 2011)
[13] http://www.kds.miszu.pl

A New Algorithm for Speech and Gender Recognition on the Basis of Voiced Parts of Speech

Jakub Karwan and Khalid Saeed

karwan@novell.ftj.agh.edu.pl, saeed@agh.edu.pl

Abstract. In this paper a description of an algorithm purposed to a speech recognition problem is presented. Samples were obtained from people in different ages, from 8 to 70 years. Authors' attention was concentrated on finding an efficient voice descriptor for the speech recognition process. To reach this goal Toeplitz matrices were used. The recognition process is based on k Nearest Neighbors algorithm and the analysis is carried out only for voiced parts of speech. Different distance metrics were compared in the aim of kNN optimization. In the research the influence of the sex recognition on final results is confirmed. The algorithm was tested for signals sampled with the rate of 8 kHz to keep all the necessary information contained in human voice.

Keywords: speech recognition, gender recognition, Toeplitz matrices, kNN.

1 Introduction

Both in the industry and science, the popularity of biometrics is growing. There are numerous solutions based on different biometric features, such as fingerprints, iris texture, voice and others, which find their place in social applications. In this paper, the emphasis is pressed on the speech recognition systems.

People, during the usage of the sense of hearing, are able to recognize different people speech and different words. Looking up to this ability, researchers, since the beginning of computer era, have tried to create efficient algorithms allowing for speech or speaker automatic recognition. In this case there are two main problems, the first is an accurate description of the speech signal, the second is efficient classification. The right description of the speech signal is a really complicated task. For many years scientists have struggled with this problem and have found many solutions. Nowadays, the most popular speech signal descriptors are based on LPC (Linear Predicting Coding) and MFCC (Mel-Frequency Cepstral Coefficients). In our approach we want to introduce our modification founded on the Toeplitz matrices. Popular classifiers such as HMM (Hidden Markov Models) or VQ (Vector Quantization) can provide the speaker recognition accuracy on 89% [1] or 92% [2].

2 Samples and Signal Preprocessing

Samples are taken from a control group, counting 26 people. Each object had to utter 10 sentences Polish numbers from 0 to 9 and all of the sentences were repeated 3

N. Chaki and A. Cortesi (Eds.): CISIM 2011, CCIS 245, pp. 113–120, 2011.

times. As a result, our base is built with 780 utterances. Tested group consists of 46% females and 54% males. The age of objects was varies, 6 objects are over age 50, 16 are between 20-50 and 4 are less than 20. Samples are recorded in audio standard (44,1kHz sampling rate) in one channel. 520 samples are used as a learning set, while the remaining 260 are used as a testing set.

2.1 Normalization

It is difficult or even impossible to obtain standardized conditions of voice registration. The distance between mouth and microphone is not constant for different sessions of measurements. What is more, different people speak with varied loudness. This problem can be solved by the normalization process given in [3].

$$F_N = \frac{F - mean(F)}{max|F - mean(F)|} \tag{1}$$

2.2 High-Pass Filtration (55Hz)

Acquisition of samples is carried out with commonly available equipment. Initial quality of samples (before filtration) gained by this device was not satisfying. The influence of the 230 V – 50 Hz electricity was visible and significant. This impact is reduced by 55Hz high-pass filter. The usage of this filter did not affect the information delivered by the speech. It is justified by the fact, that the fundamental frequency of the lowest voices is about 85 Hz [4].

2.3 Low-Pass Filtration (4kHz)

Signal is acquired in standard audio frequency sampling 44.1 kHz. It allows to code signals of frequencies lower than 22.05 kHz according to the Nyquist–Shannon sampling theorem. This range is sufficient to code all hearing band. For this research, the most important information in the signal is represented by frequencies below 4 kHz. To avoid redundancy of data, components containing higher frequencies are removed from samples.

2.4 Resampling

Only samples at frequencies below 4 kHz are used in subsequent data analysis. Thus, sampling with the frequency 44.1 kHz was superfluous. Natural reaction on this fact is the conversion of the signal sampling rate to 8 kHz. It allows to reduce the time for further analysis, especially time of feature extraction.

3 Fundamental Frequency Extraction

The authors of this paper are checking, if the analysis carried out only for voiced parts of speech gives better results than the analysis led for all utterances. Voiced parts of speech are generated during the air flow through the tense vocal folds. It leads to the production of air vibrations, which are characterized by a special frequency. This frequency is commonly called fundamental frequency. The authors decided to extract the fundamental frequency with the autocorrelation function:

$$r_l = \sum_{n=1}^{N-l} x_n \cdot x_{n+l} \tag{2}$$

where: r – value of auto-correlation, l – time shift (expressed in samples), x_n – value of the n-th sample, N – number of samples. Autocorrelation functions computed for voiced parts of speech form curves with characteristic minima and maxima (Fig.1). By contrast, the autocorrelation function of unvoiced parts of speech forms jagged curves (similar to noise functions).

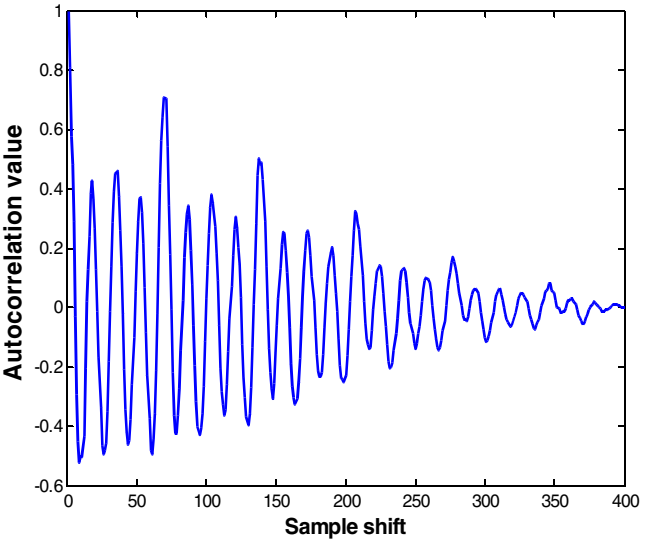


Fig. 1. Autocorrelation function of voiced speech

Location of the global maximum (excluding starting point) determines the value of fundamental frequency

$$F_{fund} = \frac{Fs}{u} \tag{3}$$

where: Fs – sampling rate, u – location of the maximum (expressed in samples). Some phonemes contain voiced and unvoiced components. For these phonemes, the calculations of fundamental frequency are more sophisticated. As a result, of computations, the obtained autocorrelation function is not as smooth as in Fig. 1. However, it still is possible to use it for fundamental frequency extraction. For this purpose a threshold is defined for minimal autocorrelation value and minimal sample shift. These actions reduce the risk of wrong fundamental frequency estimation.

3.1 Gender Recognition

The signal is divided into frames (400 samples long, with 200 samples overlap). For each speech sample, a map of fundamental frequency is created (Fig.2). Then, for each map, the average frequency of the voiced part of utterance is calculated. At the

next step, training set off samples is used for determining the threshold frequencies for each class (number). Every tested sample is classified as a female, if the average frequency of the sample is higher than the threshold and the sample is recognized as a male otherwise. The achieved level of sex recognition was about **90-95%**. Perfect results of sex recognition using this method are impossible to achieve, due to the overlap of the fundamental frequency range for both sexes. However, trials are being conducted to raise the success rate.

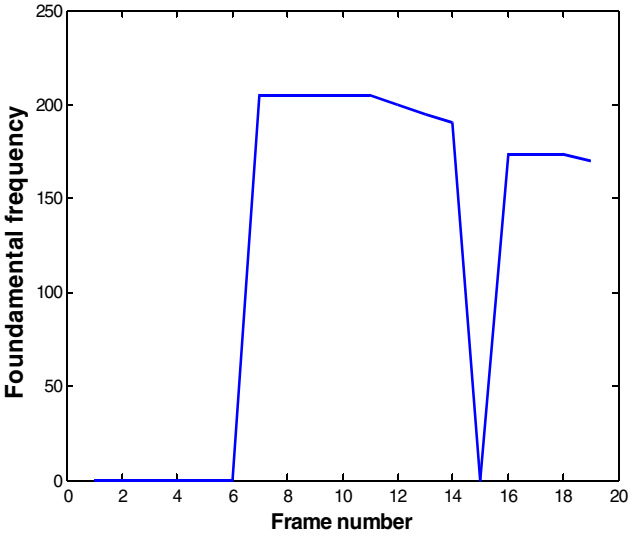


Fig. 2. Fundamental frequency – female, word "four"

4 Feature Extraction

In the authors' approach, the signal is transformed to frequency domain. Fourier transform does not allow to obtain sufficiently smoothed spectra for the presented analysis. This problem is solved by using Burg's coding model (one of LPC methods). At the final stage of feature extraction, the signal is described by Toeplitz matrix lowest eigenvalues.

4.1 Burg's Estimation

The basis of LPC model is an assumption, that speech signal sample value can be represented as a sum of sample values prior to it, multiplied by the weighting factors. Burg's coding is based on forward and backward prediction [3],[5]

$$\begin{cases} ef_p(n) = ef_{p-1} + c_p \cdot eb_{p-1}(n-1) \\ eb_p(n) = eb_{p-1} + c_p \cdot ef_{p-1}(n-1) \end{cases} \tag{4}$$

where $ef_p(n)$ – forward error, $eb_p(n)$ – backward error, c_p – prediction parameter, p – level of the prediction, n – sample number. The solution lies in minimizing the forward and backward errors. Equation (5) is used to calculate the prediction coefficients

$$c_p = \frac{2\sum_{n=p+1}^{N} ef_{p-1}(n)\cdot eb_{p-1}(n-1)}{\sum_{n=p+1}^{N} |ef_{p-1}(n)|^2 + \sum_{n=p+1}^{N} |eb_{p-1}(n-1)|^2} \tag{5}$$

where $c_0 = 1$, N – number of samples.

Burg's coding allows to obtain smoothed spectra. In this experiment, authors gained 30 point Fourier transforms for the prediction coefficient p defined at level 30. FT is carried out for real-valued data, thus the computed spectra are symmetric. As a result, authors obtained 16 point long transforms of all speech samples.

4.2 Toeplitz Matrices

Appropriate description of the signal is a difficult task. In this paper, authors propose an application of Toeplitz matrices as a voice descriptor. If spectrum is treated like an image, we can get n point representation of the signal (n $FT_{length}/2 + 1$), where x is the normalized frequency and y is the normalized amplitude. After normalization values of x and y lie in the range [1,n]. From these data we can obtain the rational function

$$f(s) = \frac{x_0 + x_1 s + x_2 s^2 + \ldots + x_n s^n}{y_0 + y_1 s + y_2 s^2 + \ldots + y_n s^n} \tag{6}$$

where s is a complex number $s = x + jy$. This equation leads to the following power series:

$$T(s) = \alpha_0 + \alpha_1 s + \alpha_2 s^2 + \ldots + \alpha_n s^n + \ldots \tag{7}$$

where

$$\alpha_0 = \frac{x_0}{y_0}, \tag{8}$$

$$\alpha_i = (y_0)^{-i-1} \begin{vmatrix} x_i & y_1 & y_2 & y_3 & \ldots & y_i \\ x_{i-1} & y_0 & y_1 & y_2 & \ldots & y_{i-1} \\ x_{i-2} & 0 & y_0 & y_1 & \ldots & y_{i-2} \\ x_{i-3} & 0 & 0 & y_0 & \ldots & y_{i-3} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ x_0 & 0 & 0 & 0 & \ldots & y_0 \end{vmatrix} \tag{9}$$

for $i = 1,..,n$. The minimal eigenvalues λ^i_{min} of TM are calculated and used at the classification stage.

Detailed description of Toeplitz matrices basis can be found in [6], [7].

5 Classification and Results

Correct classification of speech signals depends not only on proper feature extraction, but also on selecting an appropriate classifier. For classification purpose, authors use

kNN algorithm (k-Nearest Neighbors). k was set at 13. At this point, the signal description and classification is carried out only for voiced parts of speech. The extraction of voiced parts of speech is based on the computations described in section 3.

5.1 Experiment I

In this section, the authors abandoned a preliminary recognition of the sex before the speech classification. The algorithm used several distance metrics (Mahalanobis, standardized euclidean, city block distance, Chebyschev, cosine and correlation), in the aim of optimizing the kNN functioning. Results of the classification for some of them are presented below in Table 1.

Usage of correlation distance metric gave the highest average recognition rate. However, the final score is still not satisfying. In these studies, the signal is treated integrally and it is not divided into frames (at the feature extraction stage). Probably, signal framing and the application of statistical clasifiers would significantly improve the recognition results.

Table 1. Achieved recognition rate for all numbers

Number	Mahalanobis	City block d.	Correlation
0	62.00%	50.00%	**70.00%**
1	34.00%	42.00%	**46.00%**
2	46.00%	**58.00%**	46.00%
3	15.00%	27.00%	**35.00%**
4	**19.00%**	**19.00%**	**19.00%**
5	**38.00%**	31.00%	**38.00%**
6	15.00%	19.00%	**31.00%**
7	31.00%	23.00%	**38.00%**
8	**62.00%**	46.00%	35.00%
9	**58.00%**	46.00%	50.00%
average	38.00%	36.10%	**40.80%**

5.2 Experiment II

In this part of research, the influence of initial sex recognition on final speech recognition result is checked. It means, before appropriate classification, sex is automatically recognized and then the samples are divided into two groups (male group and female group). At the next step, the proper classification was carried out. The scores of the computations are presented in Table 2.

Table 2. Achieved recognition rate for males and females

Number \ Sex	Male group	Female group
0	53.00%	**82.00%**
1	67.00%	45.00%
2	44.00%	55.00%
3	40.00%	45.00%
4	19.00%	40.00%
5	43.00%	50.00%
6	38.00%	36.00%
7	63.00%	60.00%
8	**69.00%**	40.00%
9	63.00%	70.00%
average	**49.90%**	**52.30%**

As we can see, the recognition level is increased by 10%. The results show that the initial sex recognition has significant influence on the final results. However, the successful recognition is achieved only in the half of all cases.

6 Conclusions and Future Work

The research has given the answer that authors' new approach can be used in speech recognition applications. The final identification rate of 10 different classes of words oscillated around 50%. This score is promising and the authors are still improving the algorithm.

The main goal for future work is to achieve effective recognition above the 80% level. For this purpose, authors must complete the algorithm for signal framing and statistical methods of classification. It possibly may lead to significant increase of successful recognition, however computation time would highly be extended.

References

1. Abu Shariah, M.A.M., Ainon, R.N., Khalifa, O.O., Zainuddin, R.: Human Computer Interaction Using Isolated-Words Speech Recognition Technology. In: IEEE Proceedings of The International Conference on Intelligent and Advanced, pp. 1173–1178 (2007)
2. Amin, M.R., Bhotto, M.Z.: Bangali Text Dependent Speaker Identification Using Mel Frequency Cepstrum Coefficient and Vector Quantization. In: 3rd International Conference on Electrical and Computer Engineering, Dhaka Bangladesh, pp. 569–572 (2004)

3. Saeed, K., Szczepański, A.: A study on Noisy Speech Recognition. In: ICBAKE 2009 Proceedings of the 2009 International Conference on Biometrics and Kansei Engineering, Cieszyn Poland, pp. 142–147 (2009)
4. Titze, I.R.: Principles of voice production. Prentice Hall (1994)
5. Grey Jr., A.H., Wong, D.Y.: The Burg Algorithm for LPC Speech Analysis/Synthesis. IEEE Transactions on Acoustic, Speech and Signal Processing 28(6), 609–615 (1980)
6. Nammous, M.K., Saeed, K.: A Speech-and-Speaker Identification System: Feature Extraction, Description, and Classification of Speech-Signal Image. IEEE Transactions on Industrial Electronics 54(2), 887–897 (2007)
7. Saeed, K.: Image Analysis for Object Recognition. Bialystok Technical University Press, Białystkok (2004)

Fast Feature Extractors for Palmprint Biometrics

Michał Choraś and Rafał Kozik

Institute of Telecommunications,
University of Technology & Life Sciences
Bydgoszcz, Poland
{chorasm,rafal.kozik}@utp.edu.pl

Abstract. In this paper we propose to use fast feature extractors for contactless palmprint biometric system dedicated for mobile devices. We present texture feature extraction methods based on box functions, PCA and Matching Pursuit algorithm. Such methods should be effective and computationally robust to be deployable on mobile devices.

1 Introduction and Motivation

Nowadays, most hand and palmprint biometric systems are supervised and require contact with the acquisition device. Such situation contributes to negative opinion and lack of trust to biometric systems in the society. The goal of the biometrics community should be to design biometric systems that could work in a seamless way in the unconstrained environment. Another requirement is the "mobility" understood as the mobility of the subject, sensors and services (both embedded in mobile devices such as smartphones).

Currently, only few studies have been devoted to unsupervised, contactless palm images acquisition and hand pose invariance [1] [2]. In [3] authors proposed a system that uses color and hand shape information for hand detection process. Authors also introduced a new approximated string matching techniques for biometric identification, obtaining promising EER lower than 1.2%. In [4] authors proposed sum-difference ordinal filters to extract discriminative features, which allow to verify the palmprint identity in less than 200ms, without losing the high accuracy. Such fast feature extraction algorithms are dedicated for smart phones and other mobile devices.

Hereby, we propose to use palmprint in the contactless biometric system for mobile devices (unsupervised, uncontrolled image acquisition by mobile cameras).

The main contribution of the research described in this paper is a fast method for palmprint feature extraction. The proposed features are based on approximated eigenpalms enhanced with gradient information (HOG features). Results obtained for our palmprint database of right hands images (11) are promising.

The paper is structured as follows: in Section 2 the general overview of the method is provided. In Section 3 the proposed method is described in detail. Results and conclusions are given thereafter.

N. Chaki and A. Cortesi (Eds.): CISIM 2011, CCIS 245, pp. 121–127, 2011.

Fig. 1. Examples of palmprint images acquired by mobile phone camera

2 General Overview of the Method

The palmprint feature extraction methodology, proposed in this paper, is a combination of two techniques used for image description. The first one, adapting box functions, aims at describing the low frequency features, while the second one, engaging gradient and directional histograms, focuses on high frequency features.

The general description of proposed combination is shown in Fig. 2

Firstly, the low frequency features are extracted, then K gradient images of K nearest neighbors are compared to gradient image of the palmprint. The closest match within the system threshold is chosen in order to accept or reject the particular user.

3 Three-valued Box Functions

The box functions (that build Haar-like features) can be adapted in mobile devices or cameras (eg. face detectors) because they are very efficient and computationally effective.

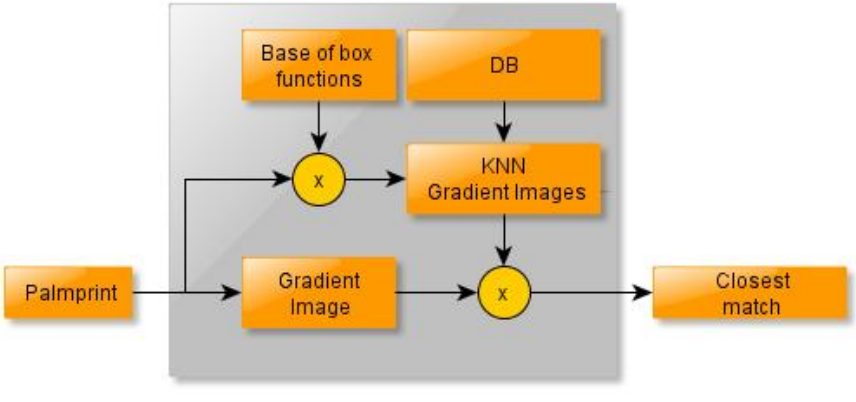


Fig. 2. Overview of the proposed method

The advantage of Haar-like features is that they can be efficiently computed in short and constant time at any scale or location thanks to the integral image (introduced by Viola and Jones in [5]).

However, in this paper, instead of Haar functions, two dimensional three-valued discrete functions are proposed to construct base of vectors:

$\{v_1, v_2, v_3, \ldots, v_N\}$

that will be used to project each of object image (p_k) onto new features space.

In other words, it is determined how much the particular object is similar to v_k (dot product a_{kn}).

The formula is described by equation 1:

$$a_{kn} = (p_k \cdot v_n). \tag{1}$$

The projection coefficients create the feature vector, that is described by equation 2:

$$w_k = (a_{k1}, a_{k2}, a_{k3}, \ldots, a_{KN}). \tag{2}$$

However, the key task is to find (faster than Ada-boost approach) efficient base of box functions than spans the feature space.

The technique proposed in this paper is PCA-guided.

3.1 PCA-Guided Feature Space Approximation

To represent feature space either non-orthogonal or orthogonal functions can be used. The most popular orthogonal basis used by computer vision algorithms are Walsh transform, DWT and PCA. Among these methods the PCA is the most widely and successfully used for palmprint and face recognition.

However, DWT, PCA and Walsh transforms are computationally expensive. The PCA could be even more expensive since it requires firstly to find eigenvectors set and then to compute palmprint projection onto this set.

Eigenvectors can be computed only once in offline mode (in order to decrease the mobile device burden). Nevertheless, the PCA projection is still computationally expensive since the dot product has to be computed (many floating point operations per one eigenvector).

Therefore we decided to represent the orthogonal PCA base with non-orthogonal base of box functions. Such approach allows to significantly reduce the number of multiplications without decreasing system effectiveness. For example the dot product of palmprint of size 512x512 and eigenpalm shown in Fig.5 requires 262144 multiplications. Using box functions (and integral images) this number can be reduced to 24 (4 read operations per one box as it is shown in Fig.6).

In the proposed method, the eigenvectors are first computed using original PCA. Each eigenvector is then normalized to have values within $< -1; 1 >$ range. Afterwards, each eigenvector is approximated by set of box functions that either can have -1,+1 or 0 sign. To solve the task of approximation, the Matching Pursuit algorithm is used.

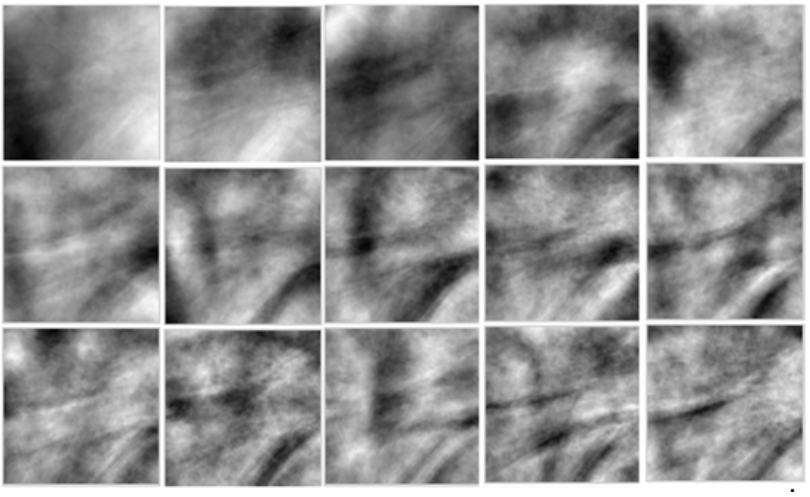


Fig. 3. Examples of eigenpalms

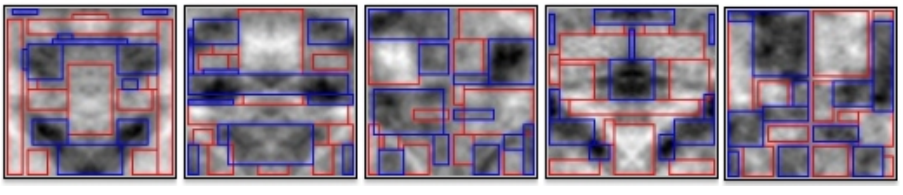


Fig. 4. Examples of eigenpalms approximated with box function (blue indicates -1, while red +1)

3.2 Matching Pursuit Algorithm for Palmprint Approximation

Matching Pursuit (MP) is a greedy algorithm that sequentially selects (in k steps) the base vector f from dictionary $D = \{f_1, f_2, ..., f_n\}$ (and adds it to the solution set $F_k = \{f_1, f_2, ..., f_k\}$), such that:

$$|c_i| = | < x - R_{F_{k-1}}(x), f_k > | \tag{3}$$

is maximized and $R_{F_k}(x) = \sum_{i=1}^{k} c_i f_i$ is an approximation of x after k steps.

The example presenting one of the eigenvector and its approximation is shown in Fig. 5

3.3 Palmprint Gradient Descriptors

The method often used for gradient description is commonly known as HOG (Histograms of Oriented Gradients) descriptors. In this paper 9-bin histograms are used (20 degrees for each bin).

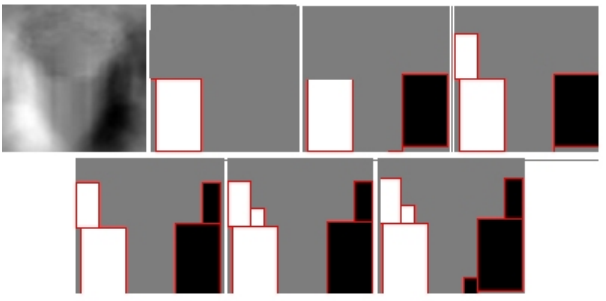


Fig. 5. One of the eigenvectors and its approximation at each iteration of MP algorithm

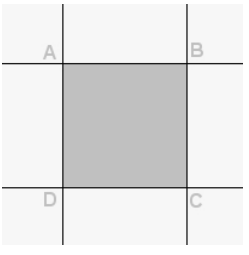


Fig. 6. When using integral images, computing average gradient magnitude for ABCD region of any size requires only 4 access operations

The palmprint image is initially divided into grid. In our case the grid of size 10x10 is used. For each grid cell HOG descriptor is computed. Each descriptor is a vector of length 9. Each vector component describes average magnitude of gradient for particular direction.

In order to speed-up the computations, the integral images are used to compute average magnitude for particular direction. The 9-bin histogram requires 9 integral images. Each integral image aggregates gradient magnitude for one direction.

The x and y components of the gradient are extracted as difference of luminance of two neighboring pixels. The direction of gradient is computed as $\arctan \frac{y}{x}$.

In order to compute average magnitude for block of texture (as it is shown in Fig. 6) it is required to read only 4 values from integral image.

4 Results

We used our own database consisting of 252 images (there are 84 individuals, for each individual there are 3 images of the right hand) for testing method effectiveness. Standard mobile devices have been used (Canon, HTC, Motorola) and the resolution of images is 640x480.

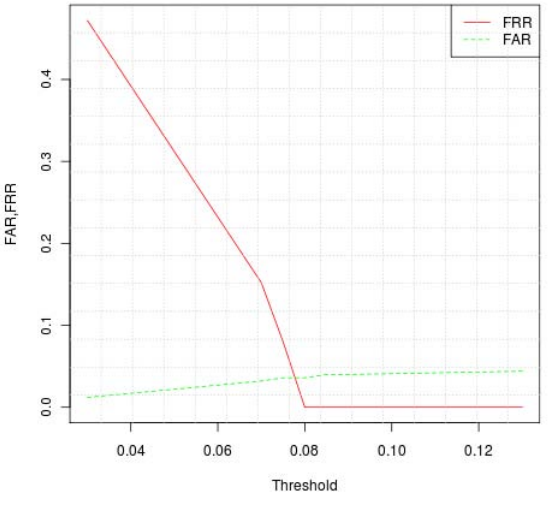


Fig. 7. FRR and FAR versus threshold for box function

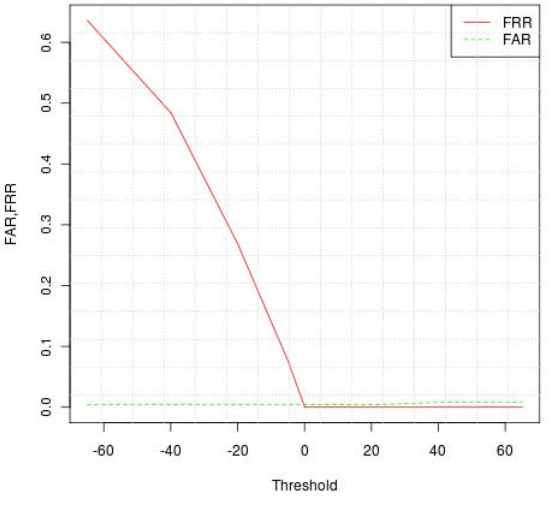


Fig. 8. FRR and FAR versus threshold for box function enhanced by HOG features

During the test classical 10-fold approach was used in order to assess the proposed method effectiveness.

Firstly, we evaluated the effectiveness of palmprint recognition based on box functions, then we assessed how gradient information can increase the effectiveness.

The effectiveness of palmprint recognition based on approximated PCA vectors is shown in Fig. 7. The average Equal Error Rate achieved for this method is 3.6%.

The results for features vector enhanced by HOG features are presented in Fig. 8. The average EER is equal to 0.4%.

5 Conclusions

In this paper our developments in palmprint feature extraction for human identification in biometrics system are presented.

We showed that palmprint texture features may be considered as very promising biometrics modality which can be used in contactless human identification systems. Our goal was to propose efficient feature extractors that can be run on mobile devices.

References

1. Fratric, I., Ribaric, S.: Real-Time Model-Based Hand Localization for Unsupervised Palmar Image Acquisition. In: Tistarelli, M., Nixon, M.S. (eds.) ICB 2009. LNCS, vol. 5558, pp. 1280–1289. Springer, Heidelberg (2009)
2. Methani, C., Namboodiri, A.M.: Pose Invariant Palmprint Recognition. In: Tistarelli, M., Nixon, M.S. (eds.) ICB 2009. LNCS, vol. 5558, pp. 577–586. Springer, Heidelberg (2009)
3. Doublet, J., Lepetit, O., Revenu, M.: Contact less palmprint authentication using circular Gabor filter and approximated string matching. In: Proc. of Signal and Image Processing (SIP), Honolulu, United States (2007)
4. Han, Y., Tan, T., Sun, Z., Hao, Y.: Embedded Palmprint Recognition System on Mobile Devices. In: Lee, S.-W., Li, S.Z. (eds.) ICB 2007. LNCS, vol. 4642, pp. 1184–1193. Springer, Heidelberg (2007)
5. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001, vol. 1, pp. 511–518 (2001)
6. Choraś, M.: Novel techniques applied to biometric human identification. Electronics 3/2009 (March 2009)
7. Choraś, M.: Emerging Methods of Biometrics Human Identification. In: Proc. of 2nd International Conference on Innovative Computing, Information and Control (ICICIC 2007), p. 365. IEEE CS Press, Kumamoto (2007)
8. Kozik, R., Choraś, M.: Combined Shape and Texture Information for Palmprint Biometrics. Journal of Information Assurance and Security 5(1), 58–63 (2010)
9. Bay, H., Tuytelaars, T., Van Gool, L.: SURF: Speeded Up Robust Features. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3951, pp. 404–417. Springer, Heidelberg (2006)

Handwritten Signature Recognition with Adaptive Selection of Behavioral Features

Rafal Doroz and Piotr Porwik

Institute of Computer Science, University of Silesia, Poland, ul. Bedzinska 39
41-200 Sosnowiec
{rafal.doroz,piotr.porwik}@us.edu.pl

Abstract. The presented work focuses on the method of handwritten signature recognition, which takes into consideration a lack of repetition of the signature features. Up till now signature recognition methods based only on signature features selection. Proposed approach allows to determine both the most useful features and methods which these features should be analyzed. In the developed method different features and similarity measures can be freely selected. Additionally, selected features and similarity measures can be different for every person.

Keywords: Signature recognition, biometrics, feature selection, *k*-NN.

1 Introduction

Handwritten signature can be defined as the name and surname written by own hand. The specification of human signatures is one of the greatest problems in the designing the credibly classifiers which work in the identification or verification modes. Both identification and verification are important in biometrics. Repeatability of signatures even the same person characterizes a large discrepancy. For example signatory may put signature every time with different velocities, pen pressures, accelerations, etc. The distribution of pen pressure can be so different in each signature, that the determination of a pattern variation can be very difficult [1], [2]. If handwritten signature will be appropriate measured then it can be treated as a biometric characteristic. Signature belongs to behavioral biometrics and modernly is widely acceptable and collectable biometric characteristic. Currently, there are many measures of determining signatures similarity. Lack of repeatability of the signature features causes problems with arbitrary indication which features should be analyzed. The selection of the signature features is a well-known and frequently described problem. It should be also noticed, that in all previous solutions, only selection of signature features was used [3], [5], [10], [13], [14]. The algorithm, presented in this work, not only performs the selection of the signature features but also indicates the best similarity measures (from the set of available measures), which minimize signature verification error. For every person different signature features and different similarity measures can be chosen. Mentioned algorithm is based on the statistical

N. Chaki and A. Cortesi (Eds.): CISIM 2011, CCIS 245, pp. 128–136, 2011.

analysis of signatures of individuals. In the proposed method of signature classification, the two stages can be distinguished: training and verification mode. The aim of a training stage is to create training sets. Thanks to these sets it is possible to evaluating which signature features and methods of their analysis preferably distinguish the original signature of a given person from the forged signature. The process of signature verification is based only on the signature features and methods determined at the training stage. The best measures, distinguishing signature from others are then collected and connected with a given signature. This information can be used in the future classification attempts.

2 Signature Acquisition

In the first step, the two sets of signatures are created for every person. The set $GS=\{S_1,S_2,...S_{no}\}$ contains only original signatures, while the set $GS'=\{S'_1,S'_2,...S'_{nf}\}$ contains a forged signatures of the same person. As aforementioned, the signature verification can be carried out on the basis of different signature features and different similarity measures. The set of all available signature features that can be used in the signature analyze process will be denoted as $F=\{f_1,f_2,..f_n\}$ and the set of all available methods of signature recognition will be denoted as: $M=\{m_1,m_2,...m_k\}$. It is assumed that each signature feature can be analyzed by means of arbitrary, known method. In this case, the set **FM** containing all possible combinations of pairs „feature-method" used to verify a signature is defined as follows:

$$\mathrm{FM}=\{(f,m)_i: f\in F, m\in M\},\ \ i=1,\ldots,(n\cdot k) \tag{1}$$

where:

- $(f,m)_i$ – the i-th pair „signature's feature (f) – analysis method (m) ",
- n – number of the signature features,
- k – number of methods used in the comparison of the features.

In a further cardinality of the set FM will be denoted by the symbol #.

3 Creation of Training Sets

In the next stage, the training sets are created. The training sets are necessary for a proper operation of the classifier. Based on the learning sets, classifier performs verification of a new, unknown signature. In the proposed method, the learning sets contain the two matrices $\mathbf{U}^{(1)}$ and $\mathbf{U}^{(2)}$. Thanks to them, the classifier is able to distinguish the original signatures of a given person from other forged signatures. The correct creation of the training set (the set of examples) is very important and significantly affects the classifier effectiveness. The matrix $\mathbf{U}^{(1)}$ is built on the basis of a set GS of original signatures of the person. The matrix contains the values of the similarity coefficient Sim calculated between pairs of original signatures. As the similarity coefficient Sim any, normalized to the interval [0,1], symmetric similarity measure can be taken. The matrix $\mathbf{U}^{(1)}$ is built as follows:

$$\mathbf{U}^{(1)} = \left[\left[S_1 \leftrightarrow S_2\right],...,\left[S_1 \leftrightarrow S_{no}\right],...,\left[S_{no-1} \leftrightarrow S_{no}\right]\right] \tag{2}$$

where:

S_i, S_j – the i-th and j-th original signature,

no – the number of original signatures.

The matrix $\mathbf{U}^{(1)}$ consists of the columnar vectors. The each columnar vector contains the values of similarity coefficients between two signatures. The similarity coefficients are calculated successively using all combinations of pairs "feature–method". The first columnar vector of the matrix $\mathbf{U}^{(1)}$ is shown below:

$$[S_1 \leftrightarrow S_2] = \begin{bmatrix} Sim(S_1 \leftrightarrow S_2)^{(f_1,m_1)} \\ \vdots \\ Sim(S_1 \leftrightarrow S_2)^{(f_1,m_k)} \\ \vdots \\ Sim(S_1 \leftrightarrow S_2)^{(f_n,m_1)} \\ \vdots \\ Sim(S_1 \leftrightarrow S_2)^{(f_n,m_k)} \end{bmatrix}_{\#(\mathbf{FM})\times 1} \tag{3}$$

where:

$Sim(S_a \leftrightarrow S_b)^{(f_i,m_j)}$ – similarity coefficient of the i-th feature f_i in the signature S_a and S_b. Similarity was determined by means of the j-th method m_j.

In order to create a matrix $\mathbf{U}^{(2)}$, the set GS of original signatures and the set GS' of forgery signatures are required. The matrix $\mathbf{U}^{(2)}$ is built similarly as before as follows:

$$\mathbf{U}^{(2)} = \left[\left[S_1 \leftrightarrow S'_1\right],...,\left[S_1 \leftrightarrow S'_{nf}\right],...,\left[S_{no} \leftrightarrow S'_{nf}\right]\right] \tag{4}$$

where:

nf – number of the all unauthorized (forged) signatures,.

The individual columns of the matrix $\mathbf{U}^{(2)}$ can be constructed similarly as in the matrix $\mathbf{U}^{(1)}$:

$$[S_1 \leftrightarrow S'_1] = \begin{bmatrix} Sim(S_1 \leftrightarrow S'_1)^{(f_1,m_1)} \\ \vdots \\ Sim(S_1 \leftrightarrow S'_1)^{(f_1,m_k)} \\ \vdots \\ Sim(S_1 \leftrightarrow S'_1)^{(f_m,m_1)} \\ \vdots \\ Sim(S_1 \leftrightarrow S'_1)^{(f_m,m_k)} \end{bmatrix}_{\#(\mathbf{FM})\times 1} \tag{5}$$

where:

S_i	–	the *i*-th original signature,
S'_j	–	the *j*-th forgery signature,
$Sim(S_a \leftrightarrow S'_b)^{(f_i,m_j)}$	–	similarity coefficient of the *i*-th feature f_i in the signature S_a and S'_b. Similarity was determined by means of the *j*-th method m_j.

4 Selection of Signature Features

As was mentioned previously, process of the signature classification can be carried out on the basis of available similarity measures and coefficients or just on several selected ones. The selection should significantly distinguish a given signature features from other signatures collected in the database. It can be done by mean of the well known the Hotelling's discriminant analysis [4]. For a given signature the Hotelling's approach allows removing such features which have the smallest discriminant power. We assumed that analyzed specific data have the normal distribution, what, for signatures, follows from the work [9]. In practice, discriminant analysis is useful to decide, whether selected pair "feature–method" is important for the classification process – if not, the other pair "feature–method" is tested. In this procedure, from the all possible pairs only pairs with the greatest discriminant power will be left.
Nowadays, the biometric methods of the signature analysis are well recognized and widely quoted and represented in the literature [3], [10]. Unfortunately reported the signature recognition levels are still insufficient. It is a challenge for the next investigations. In the present study, a new method of signature features and method of their analysis is proposed. This method will be called as the FMS what is abbreviation of the full name: Feature-Method-Selection.

As was above mentioned the set **FM** contains the all possible pairs „feature-method" which can be applied during recognition process. The set **FM** is separately created for every person. Some recognition methods included in the set **FM** can have better discriminant properties than other.

It means that some signature's features can be better recognized by means of specific methods or some features cannot be classified – then should be rejected. These connections can be discovered by proposed in this paper approach. In practice, it leads to the data reduction in the matrices $\mathbf{U}^{(1)}$ and $\mathbf{U}^{(2)}$.

For a given signature only the best discriminant features and methods of their recognition will be ultimately selected. As was said it will be done by means of reduction of the previously defined the matrices $\mathbf{U}^{(1)}$ and $\mathbf{U}^{(2)}$, respectively. In the first place variance of the all columns in the matrices $\mathbf{U}^{(1)}$ and $\mathbf{U}^{(2)}$ was computed. It is a simple operation, which is reported in many works [8]. The new variance matrices will be denoted as $\mathbf{U}^{*(1)}$ and $\mathbf{U}^{*(2)}$, respectively. The reduced matrices will be denoted as $\tilde{\mathbf{U}}^{(1)}$ and $\tilde{\mathbf{U}}^{(2)}$. At the beginning the matrices $\mathbf{U}^{(i)}$ and $\tilde{\mathbf{U}}^{(i)}$ have the same dimensions. Successively, the joined covariance matrix **K** is formed as follows:

$$\mathbf{K} = \frac{1}{N-2}\sum_{i=1}^{2}\mathbf{K}_i\,, \qquad \mathbf{K}_i = \mathbf{U}^{*(i)}\mathbf{U}^{*(i)\mathrm{T}}\,, \quad N=n_1+n_2 \tag{6}$$

where:

n_1, n_2 – the number of columns in the reduced matrices $\tilde{\mathbf{U}}^{(1)}$ and $\tilde{\mathbf{U}}^{(2)}$, respectively

The algorithm of reduction of the matrices dimension will be executed in the successive several steps:

1) The discriminant measure is calculated by means of the Hotelling's statistic [4]:

$$T^2\left(y_1,\ldots,y_p\right) = \frac{1}{N-2}\sum_{i=1}^{2} n_i\left(\overline{\boldsymbol{u}}_i - \overline{\boldsymbol{u}}\right)^T \mathbf{K}^{-1}\left(\overline{\boldsymbol{u}}_i - \overline{\boldsymbol{u}}\right) \tag{7}$$

where:

y_i – the i-th pair „feature–method",
p – the current dimension of the data vector.

2) The discriminant measure, taking into consideration the i-th pair „feature–method" absence, is then computed [4]:

$$\forall\, i \in \{1,\ldots,\#\mathbf{FM}\} \qquad T_i^2(y_1,\ldots,y_p) = T^2(y_1,\ldots,y_{i-1},y_{i+1},\ldots,y_p) \tag{8}$$

3) The necessity of the each pair „feature–method" is checked:

$$\forall\, i \in \{1,\ldots,\#\mathbf{FM}\}\; U_i = T^2(y_1,\ldots,y_p) - T_i^2(y_1,\ldots,y_p) \tag{9}$$

4) The value of the F-test is calculated. An F-distribution is well known statistical test [4].

$$F = N-p-1\cdot\frac{U_i}{1+T^2\left(y_1,\ldots,y_p\right)-U_i} \tag{10}$$

The value F from (10) is then compared with the table of critical values of the F-distribution. In our approach, in the each loop of the algorithm for the value F significance level $\alpha = 0.05$ was established. The F-distribution critical values were taken from the well-known statistical tables. Because we have two classes, the degree of the freedom is $f = N$-p-1, hence the $F_{1,N-p-1}$ distribution is checked.

If the value of F does not fall into the critical region, the current i-th row of the matrices $\tilde{\mathbf{U}}^{(1)}$ and $\tilde{\mathbf{U}}^{(2)}$ can be removed. In the successive step the parameter p is decreased, and algorithm starts with the new value of parameter p from the beginning. Hence matrices dimension can be reduced.

Finally, remaining elements of the matrices $\tilde{\mathbf{U}}^{(1)}$ and $\tilde{\mathbf{U}}^{(2)}$ might have a significant influence on the classifier's work. The mentioned procedure is executed for the every person Q. Hence, the results of the selection are stored in the set $\mathbf{FM}^{(Q)}$. It means that

for the person Q we obtain the set $\mathbf{FM}^{(Q)} \subset \mathbf{FM}$ in which the pairs $(f \leftrightarrow m)^Q$ are included. These pairs the best distinguish the genuine signature of the person Q from his/her forged signatures. It should be also strongly noted that the previously described process of the data reduction have to be repeated again when the new signature will be collected. It is necessary because the main set **FM** will be changed. It means that recognition process is always conducted in the closed set of data. Hence, the analyzed signature always belongs to one of the classes – class determined by the matrix $\tilde{\mathbf{U}}^{(1)}$ or the class determined by the matrix $\tilde{\mathbf{U}}^{(2)}$.

5 Signature Verification

At the beginning it should be explained that genuine signatures which will be recognized came from the one database, so these signatures forms the first class of objects. The forged signatures were also selected from the mentioned database. These classes have to be correctly recognized during classification process. Determination of the training sets $\mathbf{FM}^{(Q)}$ is a necessary condition for verification of the signatures taken by the person Q. Let S^* be a signature which should be verified.

In the first stage of the verification process various columns of the reduced matrixes $\tilde{\mathbf{U}}^{(1)}$ and $\tilde{\mathbf{U}}^{(2)}$ can be re-written in the form of the multidimensional vectors $\tilde{\mathbf{h}}_1 = [\tilde{h}_1,...,\tilde{h}_n] \in \tilde{\mathbf{U}}_1$ and $\tilde{\mathbf{h}}_2 = [\tilde{h}_1,...,\tilde{h}_n] \in \tilde{\mathbf{U}}_2$ where n is dimension of the reduced matrices. If $\tilde{\mathbf{h}}_1 \in \tilde{\mathbf{U}}^{(1)}$ then $\tilde{\mathbf{h}}_1 \in \pi_1$. If $\tilde{\mathbf{h}}_2 \in \tilde{\mathbf{U}}^{(2)}$ then $\tilde{\mathbf{h}}_2 \in \pi_2$. In the next step, for the classified signature S^* the vector $\mathbf{h}=[h_1,\ldots,h_n]$ is created. Elements of the vector **h** are determined as follows:

$$h_i = Sim(S^*, S_j)^{(f,m)_i^{(Q)}}, \qquad j = 1,...,n \tag{11}$$

where:

- S^* – signature to be verified,
- S_i – the i-th original signature,
- $h_i = Sim(S^*, S_j)^{(f,m)_i^{(Q)}}$ – similarity coefficient of the signatures S^* and j-th original signatures S_j. Similarity was determined with use of the i-th pair $(f,m)_i^{(Q)}$ from the set $\mathbf{FM}^{(Q)}$.

In the next stage, the distances $d(\mathbf{h}, \tilde{\mathbf{h}}_i)$ between the vector **h** and the vectors $\tilde{\mathbf{h}}_i$ are successive calculated. If $\tilde{\mathbf{h}}_i \in \pi_1$, then $d\left(\mathbf{h}, \tilde{\mathbf{h}}_i\right) \in \pi_1$. If $\tilde{\mathbf{h}}_i \in \pi_2$, then $d\left(\mathbf{h}, \tilde{\mathbf{h}}_i\right) \in \pi_2$. Among all distances the smallest k distances are selected. So *k-NN* classifier is applied [7], [15]. Selected distances may belong to the two classes: π_1 (original signature) and π_2 (forged signature). The verified signature S^* is classified to the class π_1 or π_2, respectively by using the voting method.

6 Experimental Results

Aim of the study was to evaluate the proposed method of signature verification based on automatic selection of signature features and methods of their analysis. During the study the three variants of the classifier's work have been compared:

- VAR1 – where only signature features selection was carried out, for one previously assumed method,
- VAR2 – where only signature features selection was used and the most relevant methods of their analysis were chosen,
- VAR3 – in this mode, signature features were not selected.

The study was conducted on a set of real data. Signatures came from the SVC2004 database [16]. It contains both genuine signatures and skilled forgery signatures. All calculations were realized in the Matlab environment. The set of all combinations of the pairs "feature–method" contained the **#FM**=18 pairs. Signature features and methods their analysis, were selected from many well-known from literature similarity coefficients and metrics. The researches were conducted with the use of a database containing 800 signatures coming from 40 people. For each person 10 original signatures and 10 professionally forged were used. The proposed method FMS was compared with other, known from the literature PCA [6], [11] and SVD [12] methods, where reduction of the features has been also performed. Based on [6], [11], [12] and proposed technique the different experiments have been carried out.

In the method FMS it is not possible to arbitrarily determine the dimension of the reduced vector of features. This dimension was determined automatically. The conducted studies shown that in practice, the dimension of the reduced vector is between 3 to 12. In a case of using the PCA and SVD methods the boundary features reduction should be defined *a priori*. Achieved level of the FAR and FRR errors, for different methods of features reduction, presents Tables 1-2. In the Tables 1-2 only the smallest FAR/FRR values have been shown. Those results were obtained for the different number of signatures from the sets *GS* (original signatures) and *GS* ' (forged signatures) for the different number of reduced features. The FRR ratio describes the genuine signatures classified as forged, while the FAR represents the forged signatures recognized as the genuine.

Table 1. False Accepted Rate for different features selection methods, different number of reference signatures, and different dimension of vectors

Number of signatures in the set:		FAR [%]						
GS	*GS'*	FMS	Vector dimension after using PCA[6], [11]			Vector dimension after using SVD [12]		
			2	6	12	2	6	12
3	1	6.34	8.14	6.32	6.75	6.14	7.12	5.62
5	3	1.08	2.22	2.62	2.21	2.52	2.12	2.25
10	4	1.67	3.14	3.44	3.29	3.15	3.21	3.94

Table 2. False Rejection Rate for different features selection methods, different number of reference signatures, and different dimension of vectors

Number of signatures in the set:		FRR [%]						
GS	*GS'*	FMS	Vector dimension after using PCA[6], [11]			Vector dimension after using SVD[12]		
			2	6	12	2	6	12
3	1	7.14	5.31	6.55	5.40	6.33	5.60	4.90
5	3	2.53	2.58	2.78	4.94	4.35	4.78	4.05
10	4	2.60	2.83	3.06	5.25	3.96	4.84	4.60

The Tables 1-2 show that the smallest FAR/FRR coefficients were obtained using proposed in this work method. For this method FAR=1.08%, FRR=2.53%. These results are significantly better than results obtained using the SVD or PCA methods. Differences between obtained results in the PCA/SVD methods follow from the precision of calculations. The smallest FAR/FRR ratio were obtained when the number of original signatures in the set *GS* was 5 and the number of forged signatures in the set *GS* ' was 3.

7 Conclusions

The originality of the proposed approach follows from the fact that classifier utilizes not only extracted signature's features, but also the best (for the analyzed signature) similarity measures. In the signature biometrics such statistical approach has not been applied yet.

References

1. Al-Shoshan, A.I.: Handwritten Signature Verification Using Image Invariants and Dynamic Features. In: Int. Conf. on Computer Graphics, Imaging and Visualisation, pp. 173–176 (2006)
2. Doroz, R., Porwik, P., Para, T., Wrobel, K.: Dynamic Signature Recognition Based on Velocity Changes of Some Features. Int. Journal of Biometrics 1, 47–62 (2008)
3. Doroz, R., Wróbel, K., Porwik, P.: Signatures Recognition Method by Using the Normalized Levenshtein Distance. Journal of Medical Informatics & Technologies 13, 73–77 (2009)
4. Anderson, T.W.: An introduction to multivariate statistical analysis. Wiley (1984)
5. Ibrahim, M.T., Kyan, M.J., Guan, L.: On-line Signature Verification Using Most Discriminating Features and Fisher Linear Discriminant Analysis. In: 10th IEEE Int. Symposium on Multimedia, Berkeley CA, pp. 172–177 (2008)
6. Ismail, I.A., Ramadan, M.A., Danf, T.E., Samak, A.H.: Automatic Signature Recognition and Verification Using Principal Components Analysis. In: Int. Conf. on Computer Graphics, Imaging and Visualization, pp. 356–361 (2008)

7. Józwik, A., Serpico, S.B., Roli, F.: A Parallel Network of Modified 1-NN and *k*-NN Classifiers-Application to Remote-Sensing Image Classification. Pattern Recognition Letters 19, 57–62 (1998)
8. Kirkwood, B.R., Sterne, J.A.C.: Essentials of Medical Statistics, 2nd ed. Wiley-Blackwell (2003)
9. Kovari, B., Charaf, F.: Statistical Analysis of Signature Features with Respect to Applicability in Off-line Signature Verification. In: 14th WSEAS Int. Conf. on Computers, vol. II, pp. 473–478 (2010)
10. Lei, H., Govindaraju, V.A.: Comparative Study on the Consistency of Features in On-line Signature Verification. Pattern Recognition Letters 26, 2483–2489 (2005)
11. LI, B., Wang, K., Zhang, D.: On-line Signature Verification Based on PCA (Principal Component Analysis) and MCA (Minor Component Analysis). In: Zhang, D., Jain, A.K. (eds.) ICBA 2004. LNCS, vol. 3072, pp. 540–546. Springer, Heidelberg (2004)
12. Moravec, P., Snasel, V.: Dimension Reduction Methods for Iris Recognition. In: Proc. of the Int. Workshop on Databases DATESO 2009, pp. 80–89 (2009)
13. Porwik, P., Wrobel, K., Doroz, R.: Signature Recognition Method by Means of the Windows Technique. Int. Journal Image Processing and Communication 14, 43–50 (2009)
14. Richiardi, J., Ketabdar, H., Drygajlo, A.: Local and Global Feature Selection for on-line Signature verification. In: Proc. of the 8th Int. Conf. Document Analysis and Recognition ICDAR 2005, pp. 625–629 (2005)
15. Shakhnarovich, G., Darrell, T., Indyk, P.: Nearest-Neighbor Methods in Learning and 15. In: Vision: Theory and Practice (Neural Information Processing). The MIT Press (2006)
16. Willems, G., Pison, G., Rousseeuw, P. J., Van Aelst, S.: A Robust Hotelling Test, Vol. 55, pp. 125–138. Physica Verlag, An Imprint of Springer-Verlag GmbH (2002)
`http://www.cse.ust.hk/svc2004/`

Robust Algorithm for Fingerprint Identification with a Simple Image Descriptor

Kamil Surmacz and Khalid Saeed

AGH University of Science and Technology,
Faculty of Physics and Computer Science
Department of Applied Informatics and Computational Physics
A. Mickiewicza 30 Ave., 30-059 Krakow, Poland
surmacz@novell.ftj.agh.edu.pl, saeed@agh.edu.pl

Abstract. The paper describes a fingerprint recognition system consisting of image preprocessing, filtration, feature extraction and matching for recognition. The image preprocessing includes normalization based on mean value and variation. The orientation field is extracted and Gabor filter is used to prepare the fingerprint image for further processing. For singular point detection the Poincare index with partitioning method is used. The ridgeline thinning is presented and so is the minutia extraction by CN algorithm. Different Toeplitz matrix descriptions are in the work. Their behavior against abstract set of points is tested. The presented algorithm has proved its successful use in identification of shifted, rotated and partial acquired fingerprints without additional scaling.

Keywords: fingerprint, minutiae, Gabor filtration, thinning, Toeplitz matrices, fingerprint identification.

1 Introduction

The identification of human being is an important task in an information based society. Each person has his own features that can be used to verify or identify them (fingerprint, retina, iris or voice for example). However, not all of them are useful in an automated identification system.

The first traces of fingerprint usage in identification purposes dates back to the beginning of our era [1]. Although in those days no one used scientific methods to compare and classify fingerprint, they were recognized as well as signature. Real development of this method of recognition started with Henry classification system [2]. In the last century an automated fingerprints identification system (AFIS) was developed to meet the extending needs of forensic divisions. There was also an increasing interest in identification coming from commercial market. The result was the growing number of scientific publications on fingerprint area and also more funds to develop better scanning equipments [1, 3].

Today fingerprints are present in many aspects of ordinary people life. For example, work time checks, access to computers or mobile phones, credit cards, cash machines are all protected by fingerprint scanners. Today studies focus on reconstruction of

N. Chaki and A. Cortesi (Eds.): CISIM 2011, CCIS 245, pp. 137–144, 2011.

fingerprints [4], acceleration of processing [5] and increasing security of databases [6]. However, regardless the efforts on developing AFIS there is still a space for further studies.

This paper presents an approach based on Toeplitz matrices as an image descriptor. Natural properties of those matrices, when used to fingerprint representation, can overcome difficulties with recognition of minutiae patterns in shifted, rotated and rescaled images without additional processing. Test of Toeplitz matrix properties was made on artificial created sets of points.

2 Algorithm Description

Normal fingerprints contain different types of features. They can be divided into three levels.

The first level focuses on singular points, ridge flowing, frequency, orientation and shape of fingerprint. The second level approach uses the local ridge characteristics - minutiae. There are many types of minutia but in practice only two of them are error prone and easy to find in no ideal images. They are line ending and bifurcations. Level three features are mainly sweat pores and precise details of ridges (curvature, shape, width, etc.). Acquisition of those details requires scanners with high resolution (1000 dpi) and it is not so popular in commercial applications.

In this approach the minutiae are the main features. To extract them from fingerprint image an algorithm was worked out according to Fig. 1.

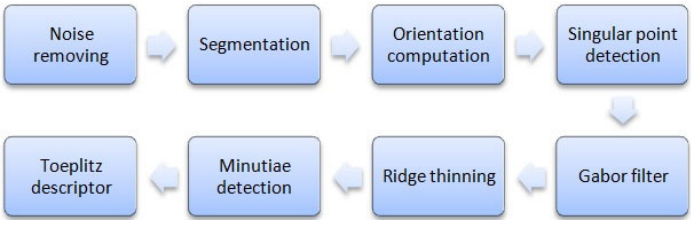


Fig. 1. Algorithm general structure

2.1 Preprocessing and Segmentation

Input images often differ one from another. Thus normalization is needed. Normalized image is defined as follows [7]:

$$N(i,j) = \begin{cases} M_0 + \sqrt{\dfrac{VAR_0(I(i,j)-M)^2}{VAR}}, if\ I(i,j) > M \\ M_0 - \sqrt{\dfrac{VAR_0(I(i,j)-M)^2}{VAR}}, if\ I(i,j) \le M \end{cases} \quad (1)$$

where M_0 and VAR_0 are chosen mean and variance values. I and N are respectively the input and the normalized images. M and VAR are the estimated mean and variation value.

Segmentation is used to determine region of interest on fingerprint images. It is done by calculating the mean value and variation mask. Afterwards for each pixel the affiliation is chosen basing on those two maps. The last step is morphological operation to erase mirror errors.

2.2 Orientation Computation and Singular Points Detection

To determine the orientation map the gradient based method [7] is used. Singular points (SP) are detected by hybrid method compiled from Poincaré index and partitioning-based methods. Orientation image is discretized to ten values. After that the region adjacent to the largest number of different values is marked as candidates. The Poincare index method was used to check the obtained areas. Spurious points are removed by morphological operations while the middle points of other areas are designated. Sample results of this approach are shown in Fig. 2. The type of singularity is not determined because these points are used only as a reference point for determination of minutiae position around them.

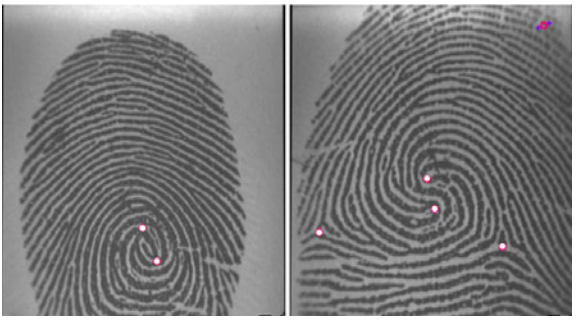


Fig. 2. Examples of SP detection

2.3 Filtration and Thinning

To improve fingerprint quality Gabor filter is used [7]. Its parameters are defined after a series of tests. Next step is adaptive thresholding to remove some noise left by filtering. Thinning is made by K3M algorithm [8]. Results of both steps are shown on Fig. 3.

2.4 Minutiae Detection and Feature Vector

The last step before creation feature vector is the localization of minutiae, which is after thinning, a rather simple task. Cross number algorithm was used to extract minutiae position. For each image point located on the line in region of interest the CN value is calculated by equation 2.

$$CN = 0.5 \sum_{i=1}^{8} |P_i - P_{i-1}| \tag{2}$$

Fig. 3. Image from Fig. 2 after filtration with Gabor filter (left) and thinning with K3M (right)

P_4	P_3	P_2
P_5	P	P_1
P_6	P_7	P_8

Fig. 4. CN algorithm mask

where P_i is defined as the image mask (Fig. 4) around the checked pixel.

The presented algorithm checks only two types of the minutiae – line ending (CN = 1) and line bifurcation (CN = 3). After obtaining the minutiae position they are checked to remove spurious ones. If they are too close to each other and have the same type there is a chance that there is only a gap or a small segment. Those points are removed.

In this approach each SP is given its own feature vector created from the closest minutiae. Thanks to that every fingerprint may have more than one feature vector. This increases the computation time in the comparing part of the algorithm but allows recognizing partial images with more singular points.

3 Toeplitz Matrices in Fingerprint Image Description

Toeplitz matrices were successfully used as an image descriptor in text recognition [9]. They have traits which can be used to improve feature vectors created from localizations of minutiae. They are invariant to shifts, rotations and scaling. Therefore tests were made to check how they can behave as descriptors artificially created points. Basic Toeplitz matrix takes the form given in [9]. Minimal eigenvalues $\lambda_{min}^{(i)}$ form a monotonically nonincreasing sequence which can be used as a feature vector for image description.

The artificial set of points consists of ten different geometrical figures located in the same space. Each point is randomly shifted from original position to simulate errors. The random shift is defined in (3) and (4).

$$x = x_0 + e \cdot \frac{rand()}{100} \tag{3}$$

$$y = y_0 + e \cdot \frac{rand()}{100} \tag{4}$$

where *rand()* are the generated values from 5 to 100, x and y are the new positions of the feature, x_0 and y_0 are the original positions (between 10 and 100), e is scaling factor. Points are sorted with respect to distance from origin (0,0) of coordinate system. They are used to compute the α_i and feature vector according to [9].

Basic Toeplitz matrix takes the following form:

$$D_i = \begin{bmatrix} \alpha_0 & \alpha_1 & \cdots & \alpha_i \\ \alpha_1 & \alpha_0 & \cdots & \alpha_{i-1} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_i & \alpha_{i-1} & \cdots & \alpha_0 \end{bmatrix} \tag{5}$$

where:

$$\alpha_i = (y_0)^{-i-1} \begin{vmatrix} x_i & y_1 & \cdots & y_i \\ x_{i-1} & y_0 & \cdots & y_{i-1} \\ \vdots & \vdots & \ddots & \vdots \\ x_0 & 0 & \cdots & y_0 \end{vmatrix} \tag{6}$$

and x_i and y_i are coordinates of i minutia. Minimal eigenvalues $\lambda_{min}^{(i)}$ of D_i form a monotonically nonincreasing sequence which can be used as feature vector.

After generating database of artificial samples they are matched by kNN classifers using Euclidian distance. Tests show that classical approach results with Toeplitz matrices have not improved the recognition rate (Fig. 5).

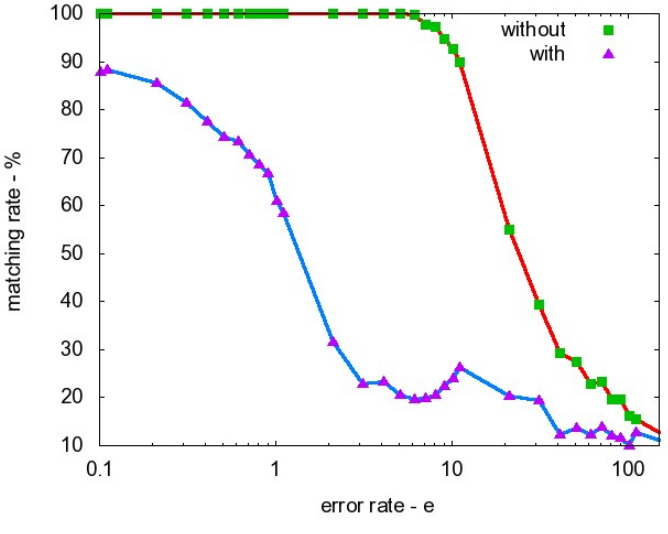


Fig. 5. Matching rate in dependence of the error rate (defined as e in eq. 3 and 4) with the classical Toeplitz matrices based descriptor and without it

Authors also check another approach to fill matrices. In the first, Toeplitz matrices take the following form:

$$D'_i = \begin{bmatrix} d_0 & d_1 & \cdots & d_i & \cdots & d_n \\ d_1 & d_0 & \cdots & d_{i-1} & \cdots & d_{n-1} \\ \vdots & \vdots & \ddots & \vdots & \cdots & \vdots \\ d_i & d_{i-1} & \cdots & d_0 & \cdots & d_{n-i} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ d_n & d_{n-1} & \cdots & d_{n-i} & \cdots & d_0 \end{bmatrix} \tag{7}$$

where:

$$d_i = \sqrt{(x_0 - x_i)^2 + (y_0 - y_i)^2}, for\ i\ \epsilon\ [1, n] \tag{8}$$

$$d_0 = \frac{1}{n}\sum_{i=1}^{n} d_i \tag{9}$$

The results obtained this way are shown in Fig. 6. Including information about distances between points increases matching rate (percentage of correct classified samples) and also increases error tolerance.

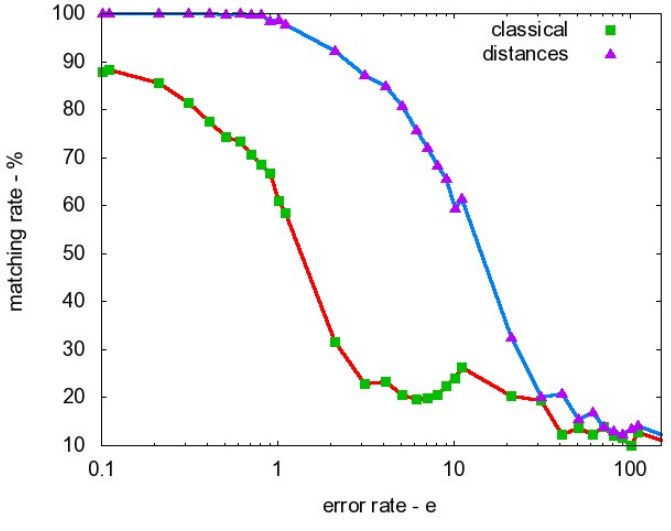


Fig. 6. Matching rate in dependence of the error rate (defined as e in eqs. 4 and 5) for the classical Toeplitz matrices and first tested modification

The second approach is defined as (8).

$$D'_i = \begin{bmatrix} d_{0,0} & d_{0,1} & \cdots & d_{0,j} & \cdots & d_{0,n} \\ d_{1,0} & d_{1,1} & \cdots & d_{1,j} & \cdots & d_{1,n} \\ \vdots & \vdots & \ddots & \vdots & \cdots & \vdots \\ d_{i,0} & d_{i,1} & \cdots & d_{i,j} & \cdots & d_{i,n} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{n,0} & d_{n,1} & \cdots & d_{n,j} & \cdots & d_{n,n} \end{bmatrix} \tag{10}$$

where:

$$d_{i,j} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}\,, for\ i \neq j \tag{11}$$

$$d_{i,i} = \frac{1}{n}\left(\sum_{j=0}^{i-1} d_{i,j} + \sum_{j=i+1}^{n} d_{i,j}\right), otherwise \tag{12}$$

After including information about distances between all points and mean value of each matrix row the obtained results are better. Comparison with the first modification is shown in Fig. 7.

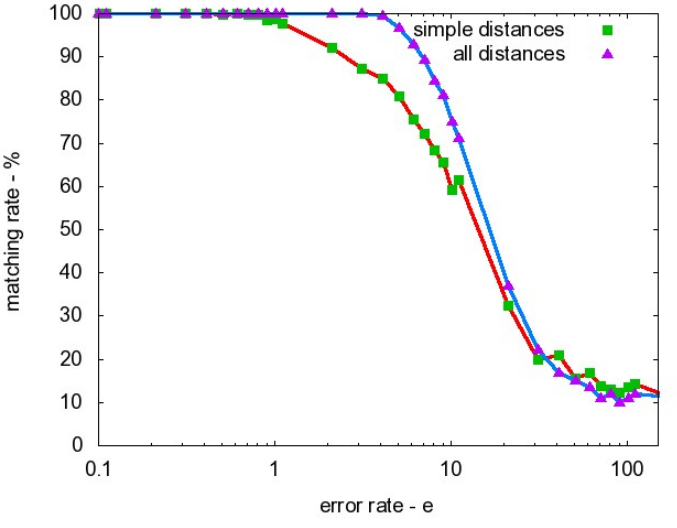


Fig. 7. Matching rate in dependence of the error rate for both tested modifications

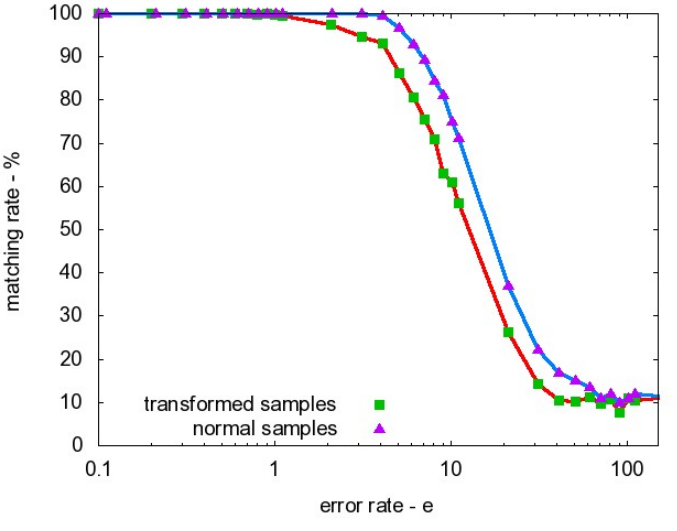


Fig. 8. Matching rate in dependence of the error rate for transformed and normal database

For the final test the database is now shifted, rescaled and rotated. Second approach (eq. 8) was used as image descriptor. As shown in Fig. 8 transformation of database only slightly decreases the matching rate. As a result of using their property there is no need to include the transformation algorithms before feature vector generation. Hence, the additional computations cost of Toeplitz matrices can be neglected. This approach can be easily adapted to create feature vector from fingerprints minutiae. Its properties can simplify the processing of the fingerprint image.

4 Conclusion

The presented algorithm requires more test and tuning to show its complete usefulness. The first author is working on the algorithm that is based on very large data base. The results then should prove the efficiency of the algorithm. Authors will focus now on checking bigger catalog of samples and on finding perfect solutions with Toeplitz matrix as a descriptor. We are aiming at decreasing the number of elements in Toeplitz matrix without information loss about the fingerprint image.

Acknowledgment. This work is supported by grant no. 11.11.2010.01 AGH University of Science and Technology in Krakow.

References

1. Lee, H.C., Gaensslen, R.E.: Advances in Fingerprint Technology, 2nd edn. Elsevier, New York (2001)
2. Tewari, R.K., Ravikumar, K.V.: History and development of forensic science in India. Journal of Postgraduate Medicine 46, 303–308 (2000)
3. Maltoni, D., Jain, A.K., Maio, D., Prabhakar, S.: Handbook of fingerprint recognition. Springer, New York (2003)
4. Liu, E., Zhao, H., Pang, L., Cao, K., Liang, J., Tian, J.: Method for fingerprint orientation field reconstruction from minutia template. IET Electronics Letters 47(2) (January 2011)
5. Fons, M., Fons, F., Cantó, E.: Fingerprint Image Processing Acceleration Through Run-Time Reconfigurable Hardware. IEEE Transactions On Circuits And Systems—II: Express Briefs 57(12) (December 2010)
6. Li, S., Kot, A.C.: Privacy Protection of Fingerprint Database. IEEE Signal Processing Letters 18(2) (February 2011)
7. Hong, L., Wan, Y., Jain, A.: Fingerprint Image Enhancement: Algorithm and Performance Evaluation. IEEE Transactions On Pattern Analysis And Machine Intelligence 20(8), 777–789 (1998)
8. Saeed, K., Tabędzki, M., Rybnik, M., Adamski, M.: K3M – A Universal Algorithm for Image Skeletonization and a Review of Thinning Techniques. International Journal of Applied Mathematics and Computer Science 20(2), 317–335 (2010)
9. Saeed, K.: Image Analysis for Object Recognition. Bialystok Technical University (2004)

A Method for Unsupervised Detection of the Boundary of Coarticulated Units from Isolated Speech Using Recurrence Plot

Arijit Sinharay, Syed Mohd Bilal, and Tanushyam Chattopadhyay

Innovation Lab
Tata Consultancy Services
Kolkata, India
{arijit.sinharay,s.bilal,t.chattopadhyay}@tcs.com

Abstract. One of the major step for Automatic Speech Recognition (ASR) is to mark the boundary of two consecutive co-articulated units. While attempt has been done to use cross recurrence plot (supervised learning) to address similar problems [1], here we propose an unsupervised approach using Recurrence Plots (RP) to address the same problem. The novelty of the work is two fold. First, we report a novel approach on using RP to identify co-articulated boundaries through prominent visual patterns. Second, we use a different quantitative approach rather than usual Recurrence Quantification Analysis (RQA) matrix [2] for automatic detection of transition boundaries. The proposed algorithm is applied on isolated spoken numerals in Bangla, a major Indian language. The results obtained from a considerably large database shows that the proposed method is a potential candidate to address the problem of Co-articulated Units Boundary (CUB) detection.

Keywords: Recurrence plot, ASR, Speech recognition, co-articulated unit boundary detection, Nonlinear signal processing.

1 Introduction

Automatic Speech segmentation is an important step for building speech recognition systems. Since the speech data is non-linear in nature, capturing and dealing with the non-linear speech data is a common and challenging problem. Beside the usual trend of using FFT based analysis or LPC based approaches [4], lots of other approaches have been developed to capture the non-linear speech data and processing the same for segmentation, such as techniques of supervised learning, like support vector machines. More recently Cross Recurrence Plots (CRP) has been used for speech analysis as Recurrence Plot (RP) is a very strong candidate for analyzing non-linear dynamic systems. However, in all the above cases the resultant segmentation of the speech data depends upon the classifier which has been used and the training set of the speech data. For example, there have been several attempts made to use cross recurrence plot for speech processing or analysis, finding the co-articulated or transition boundary between

N. Chaki and A. Cortesi (Eds.): CISIM 2011, CCIS 245, pp. 145–151, 2011.

vowel and consonants in the speech data particularly but the problem associated with supervised learning remained the same. Here we have attempted to devise an unsupervised method to identify the co-articulated boundaries based on RP matrix. The paper first provides a brief review on RP. Next the methodology showing how the RP technique is exploited to distinguish between the consonant and vowel sound that forms the basis for automated unsupervised detection of co-articulated boundaries is presented. Then experimental results are discussed in detail in the next section. Finally, conclusion is presented along with underlining some future scope of work.

1.1 Overview of Recurrence Plot

Recurrence Plots (RP) dates back to 1987, used by Eckmann et. al to visualize the trajectories for dynamical systems in phase space [3]. Recurrence plot measures the recurrence of the state of a dynamical system and is formally defined by a matrix

$$R_{i,j}(\varepsilon) = \Theta(\varepsilon - \parallel (x_i - x_j) \parallel) \quad \forall(i, j \in 1, 2, ..., N) \tag{1}$$

$$R_{i,j} = 1 \quad if \quad x_i = x_j \quad else \quad R_{i,j} = 0 \tag{2}$$

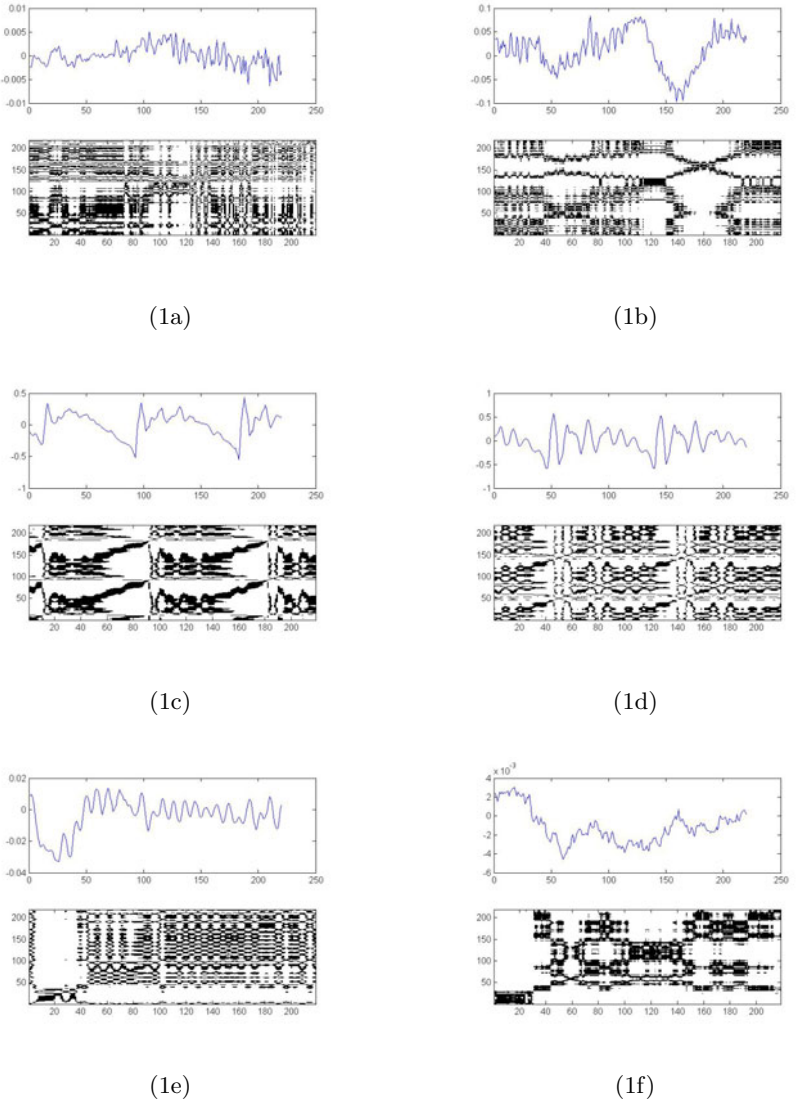
Where $\Theta(.)$ is a Heaviside function, ϵ is a threshold distance, x_i is the i^{th} point in phase space. By definition. The main diagonal is always one and called the line of identity (ROI). The most important fact is that the texture found in the RP reveals useful information of the system dynamics. The use of RP in speech processing is justified as RP is designed for complex dynamic systems and speech is highly dynamic in nature with complex behavior.

2 Proposed Methodology

We used standard 20ms non overlapping windows and carried on RP on the windowed audio signals. Audio signals have been produced by single user with 11025 bit/sec sampling rate. Each window produced a RP plot with embedding dimension as three and delay equal to one. The RP plot shows prominent change in visual patterns around the co-articulated boundaries. One can visually identify the boundaries by observing the change in RP patterns and back track in original audio sequence. To further the work for automatic detection of the boundaries we computed difference between the consecutive windows to capture the change in patterns and flagged the changes (w.r.t. to a threshold) as location of co-articulated boundaries. To verify our results (visual inference and automatic detection algorithm) we used Goldwave Player to hear the different portion of the signals and compared with our observations.

3 Results and Discussions

To prove our approach we explored Bengali numerals pronounced in Bengali to find consonant-vowel and vowel-consonant boundaries. We experimented with



(1a) (1b)

(1c) (1d)

(1e) (1f)

Fig. 1. Case study for "SHAATH" (seven)

many sets of different Bengali numbers and found promising results. To report in this paper we are showing results for Bengali number seven (pronounced as shaath) and six (chaoeh) as the representative ones. The results are presented in two stages. First we showed the windowed RP patterns for visual detection for co-articulated boundaries. Next we tried to quantify the RPs so that automatic detection of co-articulated boundaries can be identified.

3.1 Visual Inspection

RP study for Bengali seven is depicted in Fig 1. Each RP is accompanied by its generating time sequence sub-plotted at the top. At the starting "sh" is present as a consonant sound and RP doesn't show any well defined pattern (in Fig 1a), however, as the transition takes place from "sh" to "a" (i.e. consonant sound to vowel sound) some sorts of pattern starts appearing (as in Fig 1b). When it is completely in vowel sound then a very prominent pattern is achieved (Fig 1c, Fig 1d). Finally, when at the end "aa" to "th" transition takes place the well defined pattern starts disappearing (as in Fig 1e) and the whole pattern disappears when the end is hit, i.e. for the consonant sound "th" (as in Fig 1f) . Here we also note that while staying in "a" sound the different non overlapping RP windows shows the same texture but they simply got smaller and more condensed (Fig 1c vs. Fig 1d). This is because of the variation of the signal strengths/audio characteristics in different windows of "a" sound.

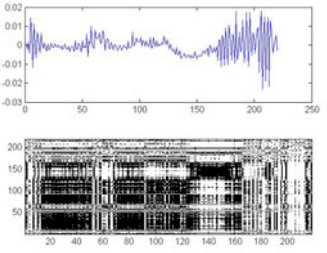
RP study for Bengali six is depicted in Fig 2. Each RP is accompanied by its generating time sequence sub-plotted at the top. As usual, the consonant sound "ch" at the start doesn't shows any pattern (as in Fig 2a). As it enters the "aoe" region (i.e. vowel sound) some pattern appears (as in Fig 2b). Further, in the "aoe" region the "a" sound is longer than "oe" and the effect is captured in the RP as more frames shows the same patterns (as in Fig 2c,Fig 2d). Here it is to be noted that pattern for "a" in seven (discussed previously) and "a" in six shows the same texture. As we enter more into "oe" we find that due to the vowel sound we still see prominent patterns, however, the patterns has changed its texture (w.r.t. "a") as it hits different vowels. For example, the pattern in Fig 2d, Fig 2e and Fig 2f are different. At the end, transition from "oe" to "h" the pattern starts disappearing (as in Fig 2g) and finally arriving at "h" the pattern disappeared (as in Fig 2h).

By locating the interesting areas in RP one can easily backtrack to the actual audio sequence to tag the boundaries. The resolution is defined by the sampling frequency.

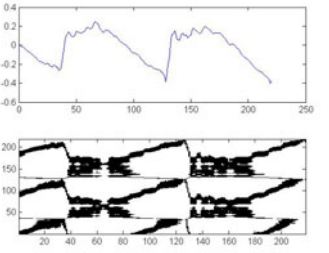
It is to be noted that our approach has advantage over FFT based approach that uses formants to detect vowels in a speech signals. For example, Fig 2c and Fig 2d both represent vowel "a" but looking at the time series we can intuitively guess that the FFT coeffiecients/formants will be different. But the RP shows same visual patterns and differs only in scale and compactness which is easily detectable by visual inspection. Yet another example would be comparing Fig 2c("a") vs. Fig 2e("o") and 2f("e") as the generating time series almost shows same periodicity and FFT/formant analysis might fail to distinguish between "a", "o" and "e" whereas the RP patterns easily reveals the presence of three different vowel sounds through three distinct patterns easily detectable by human eyes.

3.2 Automatic Detection

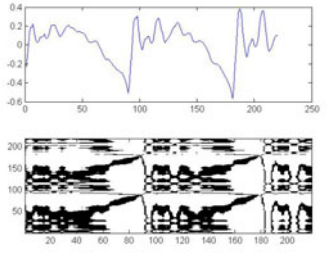
While it is easy for humans to read through the patterns and infer from it (as discussed in the previous section), writing equivalent robust algorithm is



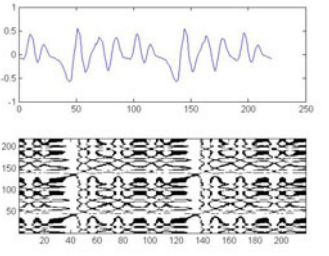
(2a)



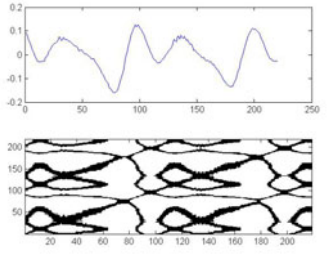
(2b)



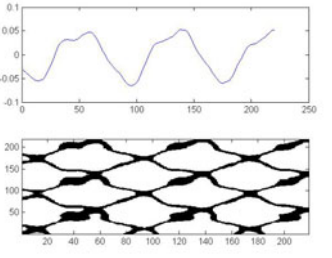
(2c)



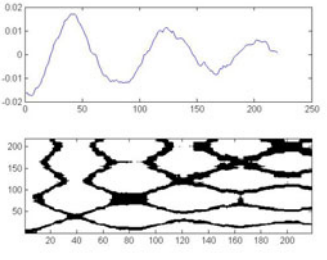
(2d)



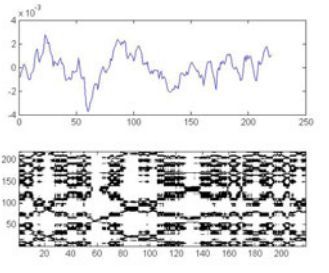
(2e)



(2f)



(2g)



(2h)

Fig. 2. Case study for "CHAOEH" (six)

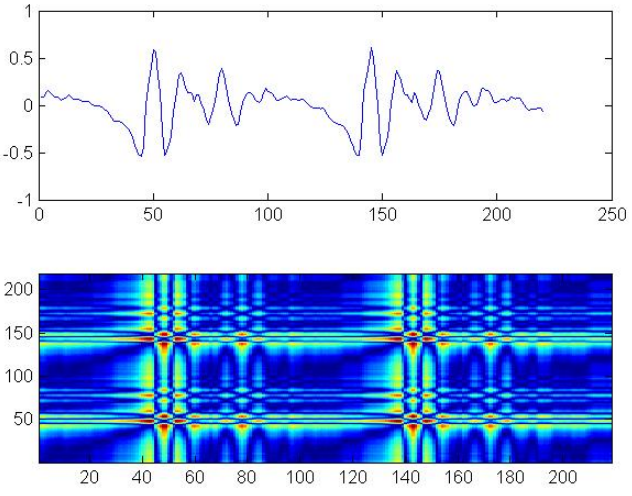


Fig. 3. RP with non-binary matrix elements keeping all the original state-space distances

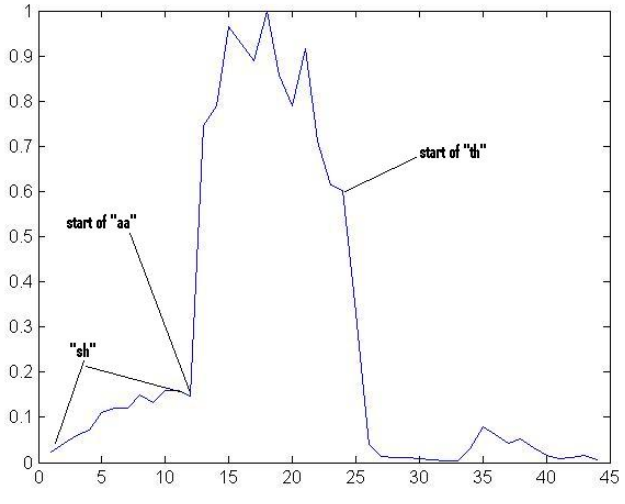


Fig. 4. Automatic detection of consonant-vowel, vowel-consonant boundaries (example for "Shaath")

challenging. However, we made an effort to write a naive algorithm just to prove that co-articulated boundaries can be automatically detected using RP and the results are discussed.

We made minor modification in the RP before we tried to quantify it. Instead of thresholding and binarizing the RP matrix (i.e. distance less than a threshold is set to zero, else kept as one, hence showing black and white patterns) we have retained the actual distances. This removes quantization error and comparison between successive windows gives more accurate results. Fig 3 shows such an example of RP matrix plot. Now we have computed the distances between two consecutive windows simply by subtracting the current window from the previous and summing over the difference matrix to get a single numeric value. We then plotted the values for all windows and calculated the slope for the resulting curve. After some trail and error we fixed a threshold to flag a change (i.e. co-articulated boundaries). Figure 4 depict the result for Bengali Seven. However, in this method the detection resolution is limited by the window size.

4 Conclusion

It has been shown that co-articulated boundaries can be captured by observing the visual patterns emerging through non-overlapping windowed RP. We also showed a completely different way to use RP matrix for information retrieval from dynamic systems (rather than using RQA). This technique has an advantage that it can be used as an unsupervised approach to detect co-articulated boundaries which can then find various applications in speech or music processing. Moreover, our work shows clear advantage of using RP analysis over usual FFT/formant analysis. Although our automatic detection algorithm is very primitive and used here only to show some promise, it is possible to quantify the RPs based on advanced algorithms to come up with some robust distance measurement schemes to accurately detect the change in patterns to identify the boundaries. Here, one can think of using various image processing and pattern recognizing algorithms along with advanced statistical distance measurement algorithms to apply on the RP matrix for better performance.

References

1. Lancia, L., Fuchs, S.: Cross-Recurrence Analysis Of Speech Signals. Centre for General Linguistics (ZAS), Berlin
2. Zbilut, J.P., Thomasson, N., Webber, C.L.: Recurrence quantification analysis as a tool for nonlinear exploration of nonstationary cardiac signals. Medical Engineering & Physics, 53–60 (2002)
3. Eckmann, J.P., Kamphorst, S.O., Ruelle, D.: Recurrence plots of dynamical systems. Europhys. Letter 5, 973–977 (1987)
4. Chattopadhyay, T., Chaudhuri, B.B.: Segmentation of Co-Articulated Basic Units in Isolated Bangla Words Using Multifactorial Analysis. In: IEEE INDICON 2004 (2004)

Czech Text Segmentation Using Voting Experts and Its Comparison with Menzerath-Altmann law

Tomáš Kocyan, Jan Martinovič, Jiří Dvorský, and Václav Snášel

VŠB - Technical University of Ostrava,
Faculty of Electrical Engineering and Computer Science,
17. listopadu 15/2172, 708 33 Ostrava, Czech Republic
{tomas.kocyan,jan.martinovic,jiri.dvorsky,vaclav.snasel}@vsb.cz

Abstract. The word alphabet is connection to a lot of problems in the information retrieval. Information retrieval algorithms usually do not process the input data as sequence of bytes, but they use even bigger pieces of the data, say words or generally some chunks of the data. This is the main motivation of the paper. How to split the input data into smaller chunks without a priori known structure? To do this, we use Voting Experts Algorithms in our paper. Voting Experts Algorithm is often used to process time series data, audio signals, etc. Our intention is to use Voting Experts algorithm for future segmentation of discrete data such as DNA or proteins. For test purposes we use Czech and English text as test bed for the segmentation algorithm. We use Menzerath-Altmann law for comparison of the segmentation result.

Keywords: Voting Experts, Text Segmentation.

1 Introduction

This paper is focused on several seemingly diverse problems. The first problem is processing of semistructured data such as text. It is commonly know fact, that the amount of this data rapidly grows so that there are two subproblems: how to store this kind of data and retrieve any piece of information from the data. To efficiently store the data it is common to use data compression. Data compression methods can treat the data as sequence of bytes, or they can use additional knowledge about the data, such as knowledge of the language of the data. With this additional knowledge data compression methods can improve their results for example by moving from byte oriented alphabet to alphabet of words. The word alphabet is connection to the second subproblem – information retrieval. Information retrieval algorithms usually do not process the input data as sequence of bytes, but they use even bigger pieces of the data, say words or generally some chunks of the data. This is the main motivation of the paper. How to split the input data into smaller chunks without a priori known structure? To do this, we use Voting Experts Algorithms in our paper. Voting Experts

N. Chaki and A. Cortesi (Eds.): CISIM 2011, CCIS 245, pp. 152–160, 2011.

Algorithm are often used to process time series data, audio signals, etc. Our intention is to use Voting Experts Algorithm for segmentation of discrete data such as DNA or proteins. For test purposes we use Czech and English text as test bed for the segmentation algorithm, because the segmentation into words is known without any doubts for Czech or English text so that results of the Voting Experts Algorithm can be easily checked. Or we can think conversely. While using Voting Experts Algorithm, which are originally aimed on signal data, we can successfully segment natural language into small chunks, then we can consider biosignals like language over some alphabet and fulfills some rules i.e. grammar.

The paper is organized as follows: in Sect. 2 a brief introduction of Voting Experts algorithm is given. Section 3 describes Menzerath-Altmann law. Experimental results are provided in Sect. 4, and conclusion is given in Sect. 5.

2 Voting Experts

The *Voting Expert Algorithm* is a domain-independent unsupervised algorithm for segmenting categorical time series into the meaningful episodes. It was first presented by Cohen and Adams in 2001 [4]. Since this introduction, the algorithm has been extended and improved in many ways, but the main idea is always the same. The basic Voting Experts algorithm is based on the simple hypothesis that natural breaks in a sequence are usually accompanied by two statistical indicators [5]: low internal entropy of episode and high boundary entropy between episodes.

The basic Voting Experts algorithm consists of following three main steps[1]:

- Build an nGram tree from the input, calculate statistics for each node of this tree (internal and boundary entropy) and standardize these values at the same level of length.
- Pass a sliding window of length n over the input and let experts vote. Each of the experts has its own point of view on current context (current content of the sliding window) and votes for the best location for a split. The first expert votes for locations with the highest boundary entropy, the second expert votes for locations with a minimal sum of internal spit entropy. By this way, the votes are counted for each location in the input.
- Look for local maximums which overcome selected threshold. These points are adepts for a split of sequence.

Tests showed that the algorithm is able to segment selected input into meaningful episodes successfully. It was tested in many domains of interest, such as looking for words in a text [4] or segmenting of speech record [8].

2.1 Two-way Voting Experts

There are several ways how to improve the basic Voting Experts algorithm. Simply we can divide these improvements into the two main groups. On the one hand,

[1] For detailed explanation of each of mentioned steps see [5].

you can add your own "expert" to voting process (for example Markov Expert in [3]) and receive additional point of view on your input. On the other hand, there are methods based on repeated or hierarchical segmenting the input [6,9].

One of the simplest ways how to slightly improve performance of segmenting is two-way passing of the sliding window. It means using classic voting algorithm supplemented by segmenting of reversed input.

This idea was outlined in [6] where the way how to make high-precision cut points by selection of higher values of a threshold was showed. Additionally, reversing the corpus, and segmenting the reversed input with Voting Experts, generates a different set of backward cut points. The subsequent intersection of sets of cut points offers high precision segmenting. However, on the other hand, this high precision is redeemed by loss of recall.

For this reason, we implemented "two-way voting" idea in another way. After running classic voting experts, we do not make cuts, but we keep the votes for each location. Then we run the backward voting on the reversed input and keep the votes again. Just after that we check the total votes against the selected threshold. Moreover, these received votes can be further balanced by FORWARD and BACKWARD multipliers which can highlight or suppress selected direction of voting.

3 Menzerath-Altmann Law

In 1928, Paul Menzerath observed the relationship between word and syllable lengths: the average syllable length decreased as the number of syllables in the word grew. In its general form, such a dependence can be formulated as follows: the longer is the construct the shorter are its constituents. Later on, this fact was put to mathematical use by Gabriel Altmann [1]. Now, this concept is known as the Menzerath-Altmann law and is considered to be one of the few general linguistic laws with evidences reaching far beyond the linguistic domain itself[2].

Formally, the Menzerath-Altmann law may be written as follows:

$$y = ax^{-b} \tag{1}$$

where y is the distance of constituents, x is the distance of construction and a, b are its parameters.

What do these parameters actually represent? When we demonstrate their relationship, according to Eq. (1), in a graph, we get exponential curves. The greater the absolute value of the negative parameter b, the sharper the demonstrated function y falls. Hence, parameter b controls curve sharpness. As for parameter a, the sample indicates that a expresses the average distance of constituents provided that 1 is substituted for x. This does not mean, however, that we can determine the amount of words containing only one morph[3].

[2] Equation (1) is known in other scientific fields as *the power law*. Mutual relationships among so-called, *scale-free network* nodes are based on this law. The world wide web is a practical example of this type of network.

[3] The question is how to define morph.

Using the Menzerath-Altmann law, we can further define individual language levels within the natural language's words and sentences. If we want to consider an element of a language as a level of a language structure, we will require compliance with the above-mentioned law. Complying with this law also results in the fact that language levels will be self-similar. Hence, text of this type presents a dynamic system with a fractal character!

4 Experiments

As it was mentioned in introduction Voting Experts Algorithm as segmentation algorithm was tested on texts written in natural language. The segmentation of text written in language known to human reader is very easy task. It is also easy task for computer reader, due to the existence of spaces in text. But is these spaces is deleted from the text the task is much harder for human reader and difficult to realize for computer reader. Therefore, the reconstruction of deleted spaces was chosen as test task for Voting Experts Algorithm. To compare performance on different languages same text written in English and Czech language was chosen[4].

For next evaluation we need define precision and recall coefficients. *Precision coefficient* – P and *recall coefficient* – R rank among the most often used indicators for ability to provide relevant documents in the information system. The precision coefficient is understood as the ratio of the amount of relevant documents returned to the entire number of returned documents. Recall represents the ratio of the amount of relevant documents returned to a given query to the entire amount of documents relevant to this query. In order to simplify information about system effectivity, methods have been created to display precision and recall measured in a 1-dimensional space. One of these methods is Van Risjbergen's *F-measure* [10,7]:

$$F_\beta = \frac{1+\beta^2}{\frac{\beta^2}{R}+\frac{1}{P}} = \frac{\left(1+\beta^2\right) R\ P}{\beta^2 P + R} \tag{2}$$

where β indicates the ratio of significance between precision and recall. For example, when β is an even 0.5, it means that the user is twice as interested in precision than in recall and when β is an 2, the users interest is vice versa. β was to 1 in our experiment.

In the first experiment, the segmentation of English text was performed. The best result were provided by following methods:

- Forward segmenting (sliding window size = 6, threshold = 3),
- Backward segmenting (sliding window size = 6, threshold = 3),
- Two-way balanced segmenting (sliding window size = 5 and 7, threshold = 7).

Results of segmentation using methods mentioned above are given in Fig. 1. The two-way segmentation repaired some single way segmentation errors e.g.

[4] We choose novel "Good Soldier Svejk" by Jaroslav Hašek.

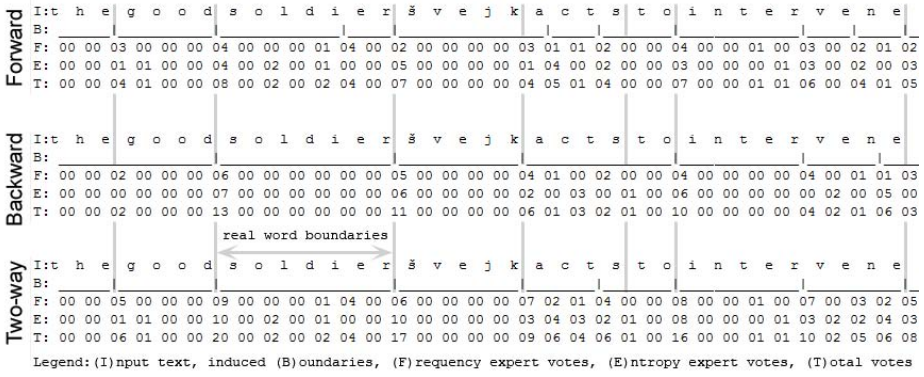


Fig. 1. Segmentation of English Text

space was between the words THE-GOOD, and conversely, the incorrect space was removed from SOLDI-ER. Evaluation of the best obtained results is described in Table 1.

The second experiment was aimed on segmentation of Czech text. The Czech version segmentation achieves worse results in both single and two way voting, see Table 2. It is caused by high number of various suffixes in the Czech language. The word form is very variable in Slavic language, e.g. Czech. There are number of suffixes which change their form according the gender and their position in a tense. Therefore there is a very high boundary entropy between these suffixes and word ends. For this reason it is difficult to find out true word boundaries. Regardless of this, in the two-way segmentation successfully correct some of wrong forward or backward segmentation, see Fig. 2.

Values in Table 2 may seem to be not so satisfactory. However, we have to realize that these results are not directly comparable with the English ones. A better way is to compare changes for the better of the two-way approach against the single way algorithm. In the Table 3 you can see the percentage improvements of both versions of text. The translated version has improvement only about 3.5% whereas the original Czech version reached the improvement about 6.3%.

The third experiment was focused on Menzerath-Altmann law. Natural languages are supposed to fulfill this law. And it can be easily check, that results of

Table 1. English text results

	F-measure	Precision	Recall	Window size	Threshold
Forward	0.75	0.77	0.72	6	3
Backward	0.71	0.74	0.68	6	3
Two way Multipliers:	0.77	0.73	0.82	5 forward	7
$1.0 - 0.9$				7 backward	

Table 2. Czech text results

	F-measure	Precision	Recall	Window size	Threshold
Forward	0.67	0.69	0.66	7	3
Backward	0.69	0.70	0.67	6	3
Two way Multipliers: 1.0 − 0.9	0.71	0.72	0.71	7 forward 6 backward	7

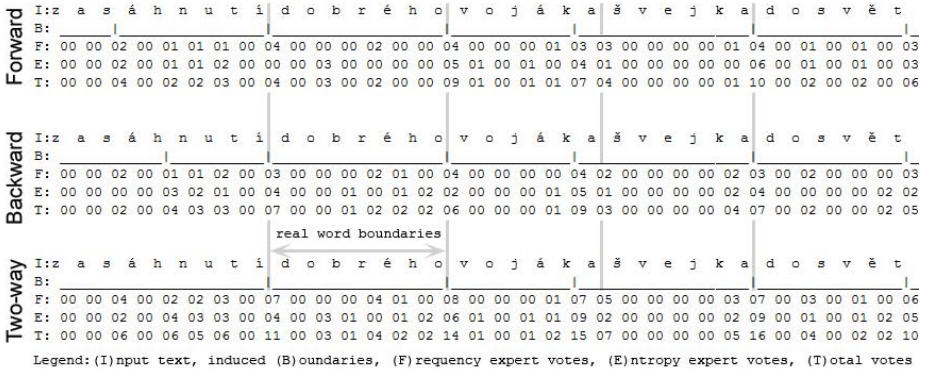


Fig. 2. Segmentation of Czech text version

Table 3. Percentage improvement

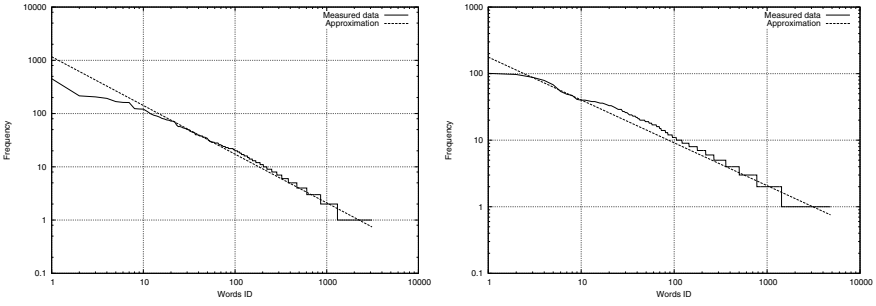
	Voting Expert Algorithm			Percentage
Language	Forward	Backward	Two-way	Improvement
English	0.75	0.71	0.77	3.47%
Czech	0.67	0.69	0.71	6.28%

Table 4. Parameters of Menzerath-Altmann Law

Language	a	b
English	1165.334	0.915
Czech	176.890	0.644

segmentation process from language that has the same property as English and Czech languages.

Menzerath-Altmann law can be interpreted as relationship between length of some text construct measured by a number of its constituents. In this way number and frequency of individual words is observed. Table 4 provides numerical

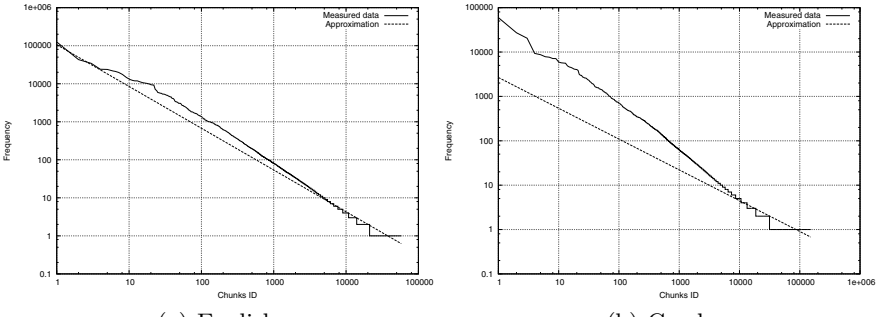


(a) English Text (b) Czech Text

Fig. 3. Curves according to Menzerath-Altman law (both axes have logarithmic scale)

Table 5. E.coli DNA – Parameters of Menzerath-Altmann Law

Language	a	b
English	106201.440	1.098
Czech	2670.092	0.694



(a) English (b) Czech

Fig. 4. Attempt of segmentation of E.coli DNA acid, Menzerath-Altman law curves (both axes have logarithmic scale)

results of the experiment (coefficient a and b) while Fig. 3 provides graphical interpretation of the measured data and approximation curve.

5 Conclusion

The Voting Experts algorithm can be used in many domains in which we want to look for some meaningful episodes. Therefore we are going to test this algorithm on pattern recognition in electroencephalography signal records.

Electroencephalography (EEG) [12] is the measurement of electrical activity produced by the brain. Electroencephalograph exams are based on the principle that the brain emits electrical waves. These brain waves register the electrodes sent from the cap on the patient's head. The EEG machine multiplies the strength of these signals and assigns them a subsequent curve. The shape and character of these curves are the results of actual brain activity.

Our effort is to decompose this brain activity signal by the Voting Experts algorithm. After that, the found sequences can be employed for marking of important brain signal patterns. These patterns can be further used in many ways, such as for controlling robots by human mind.

But there is a little problem with EEG records - human body does not have a precise internal clock, so the measured signals may vary in time and patterns can differ in their length or speed. For this reason we are going to add the third expert (DTW expert) which will deal with the comparison of two distorted patterns.

We will also test the Voting Experts algorithm to segment DNA. We use E.coli DNA taken from Canterbury Compression Corpus[5] [2]. First results are provided in Table 5 and Figs. 4. Parameters of segmentation were set to similar values as in Czech or English text. In this case the Czech configuration of parameters produces three times more segments than the English setting. The question is – segments, words in DNA constitute language without flection?

In general, the Dynamic time warping (DTW) [11] is a method that allows you to find out an optimal mapping between two given time series with certain restrictions. The sequences are non-linearly warped in the time dimension to each other and then the "cost" of mapping is quantified.

Acknowledgment. This work is supported by the grant of Grant Agency of Czech Republic No. 205/09/1079.

References

1. Altmann, G.: Prolegomena to Menzerath's law. Glottometrika 2, 1–10 (1980)
2. Arnold, R., Bell, T.: A Corpus for the Evaluation of Lossless Compression Algorithms. In: Proc. 1997 IEEE Data Compression Conference, pp. 201–210 (1997)
3. Cheng, J., Mitzenmacher, M.: Markov Experts. In: Proceedings of the Data Compression Conference, DCC (2005)
4. Cohen, P.R., Adams, N.: An Algorithm for Segmenting Categorical Time Series Into Meaningful Episodes. In: Hoffmann, F., Adams, N., Fisher, D., Guimarães, G., Hand, D.J. (eds.) IDA 2001. LNCS, vol. 2189, pp. 198–207. Springer, Heidelberg (2001)
5. Cohen, P.R., Adams, N., Heeringa, B.: Voting Experts: An Unsupervised Algorithm for Segmenting Sequences. To Appear in Journal of Intelligent Data Analysis (2007)
6. Hewlett, D., Cohen, P.: Bootstrap Voting Experts. In: Proceedings of the Twenty-First International Joint Conference on Artificial Intelligence, IJCAI (2009)
7. Ishioka, T.: Evaluation of criteria on information retrieval. Systems and Computers in Japan 35(6), 42–49 (2004)

[5] http://corpus.canterbury.ac.nz

8. Miller, M., Wong, P., Stoytchev, A.: Unsupervised Segmentation of Audio Speech Using the Voting Experts Algorithm. In: Proceedings of the Second Conference on Artificial General Intelligence, AGI (2009)
9. Miller, M., Stoytchev, A.: Hierarchical Voting Experts: An Unsupervised Algorithm for Hierarchical Sequence Segmentation. In: Proceedings of the 7th IEEE International Conference on Development and Learning (ICDL) (Best Paper Award, ICDL 2008) (2008)
10. Muller, M.: Dynamic Time Warping. Information Retrieval for Music and Motion, pp. 69–84. Springer, Heidelberg (2007) ISBN 978-3-540-74047-6
11. Van Rijsbergen, C.J.: Information Retrieval, 2nd edn. Department of Computer Science, University of Glasgow (1979)
12. Swartz, B.E., Goldensohn, E.S.: Electroencephalography and Clinical Neurophysiology. Electroencephalography and Clinical Neurophysiology 106(2), 173–176 (1998)

On Creation of Reference Image for Quantitative Evaluation of Image Thresholding Method

Soharab Hossain Shaikh, Asis Kumar Maiti, and Nabendu Chaki

University of Calcutta, India
92 APC Road, Kolkata 700009, India
{soharab.hossain,maiti.asis}@gmail.com,
nabendu@ieee.org

Abstract. A good reference image is important for relative performance analysis of different image thresholding techniques in a quantitative manner. There exist standard methods for building reference images for document image binarization. However, a gap is found for graphic images referencing. This paper offers six different techniques for building reference images. These may be used for comparing different image thresholding techniques. Experimental results illustrate the relative performance of five different image thresholding methods for the six reference image building methods on a set of ten images taken from the USC-SIPI database. The results would help picking up the right reference image for evaluating binarization techniques.

Keywords: Reference image, image thresholding, image binarization, quantitative evaluation, majority voting, misclassification error, relative foreground area error.

1 Introduction

Image binarization is a very important step in image processing and pattern recognition applications. Selecting proper threshold value is an important step for image binarization. The result of binarization depends heavily on the selection of the threshold value. There are different techniques for finding threshold for image binarization. Standard thresholding algorithms are found in the literature for graphic image [1, 2, 9, 10, 11, 12] as well as document image binarization [4, 6, 7, 8].

The performance evaluation is an important criteria whenever proposing a thresholding algorithm. This calls for the requirement of preparing a reference image with which different thresholding techniques should be compared to produce a quantitative result. Some techniques are available for preparing reference or ground truth image for document images [4]. However, there is a gap in the literature for preparing a reference image for comparing graphic images. In most of the cases [2, 3] a reference image has been formed manually which involves human interference. This approach is highly subjective, so the quality of the reference image varies with a large extent from one person to another. Therefore, an alternative technique is required for an objective measure of the reference image.

N. Chaki and A. Cortesi (Eds.): CISIM 2011, CCIS 245, pp. 161–169, 2011.

Keeping this in view, this paper presents some methods for creating the reference image. Five standard image binarization techniques, namely, Otsu [13], Niblack [10], Bernsen [11], Sauvola [6] and Kapur [12] have been chosen for showing how the performance of different image binarization techniques varies on the basis of different reference images with which the binarized image is being compared.

2 Methods for Creation of Reference Image

2.1 A Majority Voting Scheme

The reference image can be formed by using a majority voting scheme as used in [1]. The results of binarization of several techniques are taken and all these images are analyzed pixel by pixel. A pixel is set to 1 if according to majority of the representative methods that particular pixel has a value of 1 otherwise it is set to 0. This technique is an unbiased estimator of the reference image as equal weight is given to all the binarization techniques (Method-1).

2.2 Iterative Local Thresholding

A new method for binarization has been proposed in [1]. This is a local thresholding approach which requires partitioning of the original gray scale image into a number of logical segments. The technique presented in this section is a combination of partitioning as in [1] and finding threshold as in [9] for each partition locally.

Iterative local thresholding requires the original gray scale image to be partitioned into several segments and then finding a threshold for each partition as follows: An initial threshold is chosen to be the average of the grayscale values in that partition. The pixels belonging to this partition are divided into two groups based on this threshold. Next the average gray values of the two groups are calculated and a mean of these two gray values is calculated to get a new threshold. This process is repeated with the new value of the threshold until the changes between the values of the threshold varies less than a predefined parameter in two successive iterations [9]. This approach can be applied to each partition to calculate a local threshold for that partition. (Method -2).

2.3 Local/Global Thresholding Based on Mean Gray Level of the Histogram

This approach requires the original gray scale image to be partitioned into several segments and then finding a threshold for each partition as follows: for each partition a local threshold is calculated by using the mean of all the gray levels in that partition. This value is used as a threshold and applied to each partition locally. (Method-3A). Alternatively the values of these local thresholds can be used to find an average threshold which can be used as a global threshold for the image. (Method -3B).

2.4 Histogram Based Local/Global Thresholding Using Otsu's Method

In this method the original image is partitioned. A threshold value is found for each partition using Otsu's method and the image is binarized by applying these threshold values to each partition locally. The Otsu's method for finding threshold is one of the best algorithms for image binarization for both document graphic images. Experimental results show that, in most of the cases, the reference image generated using this approach is visually very good and close to the one done by any human being manually. This justifies using Otsu's method in each partition in search of a local threshold for image binarization. (Method-4A). This idea can be extended for generating a global threshold by taking the mean of these threshold values for each partition. (Method-4B).

3 Experimental Results

Two quantitative measures have been considered for analyzing the performance of different binarization algorithms on different reference images.

3.1 Misclassification Error (ME)

Misclassification error (ME) reflects the percentage of background pixels wrongly assigned to foreground, and conversely, foreground pixels wrongly assigned to background [5]. For the two-class segmentation problem, ME can be expressed as:

$$ME = 1 - \frac{|B_0 \cap B_T| + |F_0 \cap F_T|}{|B_0 \cap F_0|}$$

where B_0 and F_0 denote the background and foreground of the original ground-truth (reference) image, B_T and F_T denote the background and foreground area pixels in the test image, and $|\,.\,|$ is the cardinality of the set. The ME varies from 0 for a perfectly classified image to 1 for a totally wrongly binarized image.

3.2 Relative Foreground Area Error (RAE)

The next comparison is based on a measure for the area feature; the relative foreground area error as defined in [5].

$$RAE = \begin{cases} \frac{A_o - A_T}{A_o} & if\ A_T < A_o \\ \frac{A_T - A_o}{A_T} & if\ A_T \geq A_o \end{cases}$$

Here A_0 is the area of reference image, and A_T is the area of threshold image. Obviously, for a perfect match of the segmented regions, RAE is zero, while if there is zero overlap of the object areas the penalty is the maximum one. The images from USC-SIPI database [14] have been used for the experimentation purpose.

Table 1. Results of Method-1

Image Name	Otsu		Niblack		Bernsen		Sauvola		Kapur	
	ME	**RAE**	**ME**	**RAE**	**ME**	**RAE**	**ME**	**RAE**	**ME**	**RAE**
Aerial	7.28	8.08	25.28	26.86	**0.02**	**0.02**	3.68	3.88	0.22	0.25
Baboon	10.84	17.32	20.77	7.20	**10.06**	**2.96**	29.89	32.31	13.15	13.18
Barcode	0.28	0.31	33.15	36.67	**0.12**	**0.02**	1.24	0.66	3.28	3.70
Brick	5.20	9.34	14.81	**4.80**	5.20	9.34	16.61	22.43	**5.11**	5.92
F16-Jet	5.62	6.85	26.67	23.56	**1.43**	**0.54**	13.12	13.75	6.98	8.50
House	16.39	29.23	28.13	**3.17**	**7.54**	13.44	39.37	41.23	51.69	92.16
Lena	**12.75**	9.67	25.96	**3.92**	13.25	20.54	29.98	31.66	12.85	13.48
Pepper	16.22	22.47	26.99	**3.93**	16.05	21.88	28.04	29.96	**15.71**	20.65
Texture	37.60	45.82	63.66	67.23	25.69	31.32	**15.52**	**15.64**	17.90	17.91
Tree	**2.00**	3.84	12.22	6.75	**2.00**	3.84	13.95	20.47	2.58	**0.08**
Average	11.42	15.29	27.76	18.41	**8.14**	**10.39**	19.14	21.20	12.95	17.58

Table 1 shows the result of ME and RAE measures for five binarization algorithms where the reference image has been created using Method 1. The minimum entry for each image in each row of the table has been marked in bold face for both ME and RAE measures. Even though out of the ten images Kapur's method scores minimum for four images (for ME) and three images (for RAE), the overall average performance of Bernsen is the best according to both ME and RAE measures.

Table 2. Results of Method-2

Image Name	Otsu		Niblack		Bernsen		Sauvola		Kapur	
	ME	**RAE**	**ME**	**RAE**	**ME**	**RAE**	**ME**	**RAE**	**ME**	**RAE**
Aerial	7.04	7.83	25.12	26.66	0.22	**0.24**	3.88	4.13	**0.01**	0.27
Baboon	13.79	21.03	19.28	11.37	**4.80**	**7.32**	26.95	29.13	6.56	9.09
Barcode	0.51	**0.57**	32.53	36.11	0.78	0.88	**0.50**	0.23	2.50	2.85
Brick	1.25	2.41	16.76	11.57	1.25	2.41	20.40	27.94	**0.66**	**1.26**
F16-Jet	**2.88**	3.62	23.97	20.91	**3.19**	3.87	15.85	16.64	4.24	5.34
House	**0.24**	16.26	38.64	18.16	8.85	**2.36**	55.74	50.33	35.30	90.72
Lena	**0.94**	**1.62**	31.22	7.51	6.07	10.58	37.08	39.27	1.52	2.64
Pepper	0.94	1.82	33.72	17.81	0.56	1.07	41.60	44.70	**0.25**	**0.49**
Texture	19.45	30.44	37.61	57.93	**7.70**	**11.81**	33.38	34.30	36.05	36.07
Tree	1.22	2.04	12.22	8.45	1.22	2.04	14.64	21.92	**0.74**	**1.75**
Average	4.82	8.77	27.11	21.65	**3.46**	**4.26**	25.00	26.86	8.78	15.05

Table 2 shows the similar results where the reference image has been created using Method-2. Out of the 10 images, Bernsen's method scores minimum for 6 images (for ME) and 4 images (for RAE). Bernsen's method again is better than others on the average performance for both ME and RAE measures. Again Otsu is the second best.

Table 3. Results of Method-3A

Image Name	Otsu		Niblack		Bernsen		Sauvola		Kapur	
	ME	**RAE**	**ME**	**RAE**	**ME**	**RAE**	**ME**	**RAE**	**ME**	**RAE**
Aerial	23.02	27.20	**18.73**	**8.51**	29.80	33.07	33.46	35.68	29.59	32.91
Baboon	18.16	**0.91**	**17.86**	10.08	20.53	14.00	40.29	43.52	25.18	27.55
Barcode	2.52	2.85	33.10	34.61	2.80	3.16	2.21	2.52	**0.49**	**0.57**
Brick	**12.25**	5.08	12.68	9.08	**12.25**	5.08	19.04	25.92	12.16	**1.49**
F16-Jet	15.57	13.83	25.17	**4.76**	19.24	20.17	29.39	30.77	**15.01**	12.28
House	22.90	33.69	24.41	**3.24**	**17.49**	18.90	35.62	37.28	55.46	92.65
Lena	23.75	5.05	26.64	10.74	**23.66**	7.34	39.26	41.39	23.85	**0.88**
Pepper	**23.56**	10.39	28.20	9.95	23.58	9.71	36.84	39.41	23.74	**8.28**
Texture	47.59	34.60	56.28	60.45	40.17	**17.09**	29.56	30.12	31.98	32.00
Tree	**6.66**	0.09	12.15	10.24	**6.66**	**0.09**	15.83	23.45	7.04	3.67
Average	**19.60**	13.37	25.52	16.16	19.62	**12.86**	28.15	31.00	22.45	21.23

The reference images have been formed using Method-3A and Method-3B for Table 3 and Table 4 respectively. Results show that the performance of Bernsen and Kapur is better than other methods for highly textured images like *Brick*, *Tree*, etc.

Table 4. Results of Method-3B

Image Name	Otsu		Niblack		Bernsen		Sauvola		Kapur	
	ME	**RAE**	**ME**	**RAE**	**ME**	**RAE**	**ME**	**RAE**	**ME**	**RAE**
Aerial	**12.78**	12.86	18.03	**8.69**	17.92	19.88	21.57	23.01	17.72	19.70
Baboon	**13.36**	6.47	19.70	**4.73**	15.40	8.90	37.16	40.16	19.81	23.24
Barcode	**0.03**	**0.01**	33.03	36.47	0.28	0.32	0.99	0.34	3.00	3.40
Brick	**6.70**	8.09	14.79	6.10	**6.70**	8.09	17.22	23.49	6.54	**4.61**
F16-Jet	6.83	4.42	25.19	14.15	10.60	11.44	22.15	23.21	**6.64**	**2.69**
House	19.27	31.95	27.31	**0.69**	**14.92**	16.77	37.15	38.89	53.93	92.46
Lena	**19.95**	**1.65**	27.01	7.54	20.17	10.56	37.19	39.29	20.12	2.61
Pepper	19.36	18.05	27.82	**1.53**	19.28	17.42	31.58	33.75	**19.13**	16.12
Texture	46.40	34.25	55.57	60.24	39.68	**16.65**	**29.81**	30.49	32.34	32.36
Tree	**4.08**	**0.60**	12.34	9.79	**4.08**	**0.60**	15.43	23.06	4.52	3.18
Average	**14.88**	11.83	26.08	14.99	14.90	**11.06**	25.02	27.57	18.37	20.04

Niblack also gives lowest RAE scores for 5 out of ten images as shown in Table 4. However, according to the average performance, Bernsen's method wins over others in terms of both ME and RAE measures. The reference images have been formed using Method-4A and Method-4B. This has been listed in Table 5 and Table 6 respectively. The results show that for Table 5 Otsu scores minimum for misclassification error for six images out of the ten being considered. Also the average performance of Otsu is the best for this type of errors.

Table 5. Results of Method-4A

Image Name	Otsu		Niblack		Bernsen		Sauvola		Kapur	
	ME	**RAE**	**ME**	**RAE**	**ME**	**RAE**	**ME**	**RAE**	**ME**	**RAE**
Aerial	**19.78**	23.87	20.44	**4.32**	27.04	30.00	30.70	32.73	26.84	29.85
Baboon	5.88	10.20	23.94	**0.77**	**3.10**	5.11	34.85	37.67	14.46	20.05
Barcode	2.52	2.85	33.10	34.61	2.79	3.15	2.21	2.51	**0.49**	**0.57**
Brick	7.35	12.73	16.14	**1.11**	7.35	12.73	15.83	19.42	**5.45**	9.43
F16-Jet	**1.21**	**1.55**	26.83	19.21	4.86	5.89	17.52	18.39	2.57	3.30
House	14.12	26.23	28.12	**7.10**	**5.26**	9.78	41.63	43.62	49.41	91.82
Lena	**5.87**	9.14	30.72	**3.35**	12.87	20.07	30.39	32.06	8.32	12.97
Pepper	4.92	8.88	32.77	11.44	4.54	8.18	37.71	40.41	**3.73**	**6.73**
Texture	12.57	22.05	45.07	52.86	**0.67**	**1.18**	40.48	41.37	42.93	42.95
Tree	1.96	3.76	12.91	6.82	1.96	3.76	14.31	20.53	**0.01**	**0.46**
Average	7.62	12.13	27.00	14.16	**7.05**	**9.99**	26.56	28.87	15.42	21.81

Also the average performance of Otsu is the best for the measure of misclassification errors. One possible reason could be that this method has been used for local threshold calculation for the creation of this type of reference image. However, Bernsen's approach outperforms all other techniques in terms of average performance for RAE measure.

Table 6. Results of Method-4B

Image Name	Otsu		Niblack		Bernsen		Sauvola		Kapur	
	ME	**RAE**	**ME**	**RAE**	**ME**	**RAE**	**ME**	**RAE**	**ME**	**RAE**
Aerial	**7.46**	**9.00**	19.36	12.56	14.72	16.34	18.38	19.60	14.52	16.15
Baboon	5.88	10.20	23.94	**0.77**	**3.10**	5.11	34.85	37.67	14.46	20.05
Barcode	**0.05**	0.88	33.03	36.47	0.28	**0.31**	1.00	0.35	3.00	3.40
Brick	5.43	9.71	16.47	**4.41**	5.43	9.71	17.20	22.11	**3.52**	6.30
F16-Jet	**1.37**	**1.76**	26.81	19.38	4.70	5.69	17.36	18.22	2.73	3.50
House	6.27	13.64	34.10	20.65	**2.58**	**5.32**	49.48	51.84	41.56	90.43
Lena	**4.73**	7.51	30.99	**1.61**	11.74	18.63	31.51	33.26	7.18	11.40
Pepper	5.87	10.41	32.55	9.92	5.49	9.73	36.81	39.39	**4.68**	**8.30**
Texture	27.46	38.19	54.73	62.61	**15.55**	**21.63**	25.86	26.07	28.04	28.06
Tree	**0.80**	**1.56**	13.20	8.90	**0.80**	**1.56**	15.21	22.30	1.16	2.23
Average	**6.53**	10.29	28.52	17.73	6.44	**9.40**	24.77	27.08	12.09	18.98

In Table 6 it is seen that in terms of average performance, Bernsen's algorithm produces the best result for both ME and RE measures.

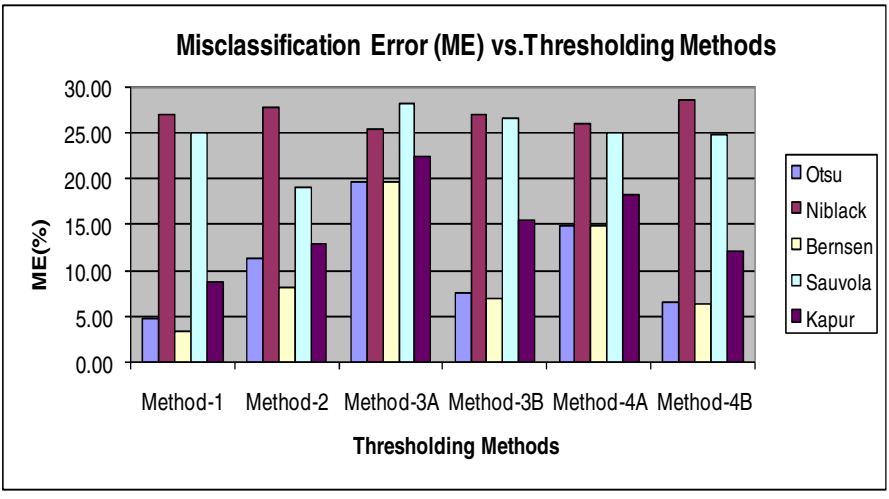


Fig. 1. Performance Analysis for Misclassification Error for six methods

Figure 1 and Figure 2 show the performance analysis (ME and RAE respectively) of all the methods for building the reference with respect to the average performance of the five binarization techniques. For both ME and RAE measures, the average performance of Bernsen's method outperforms others. However, Otsu's method gives very similar results for Method-3B, Method-4A and Method-4B.

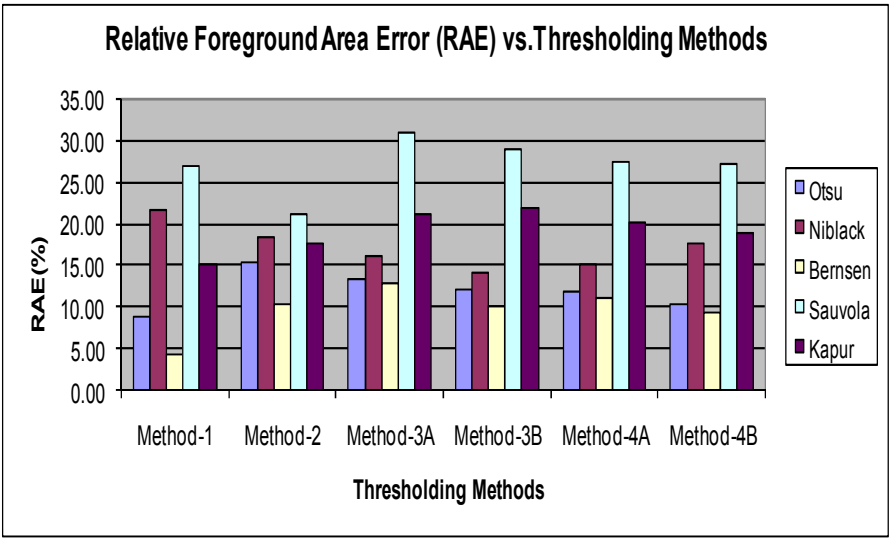
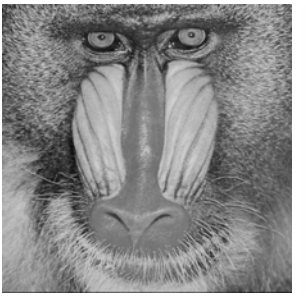
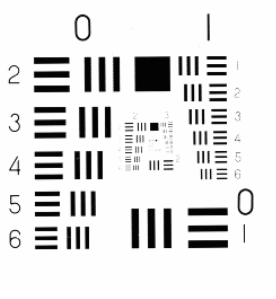


Fig. 2. Performance Analysis Relative Foreground Area Error for six methods

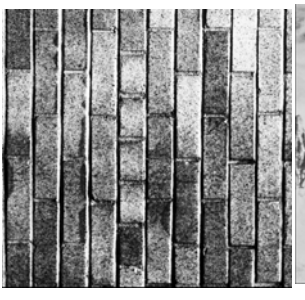
Aerial



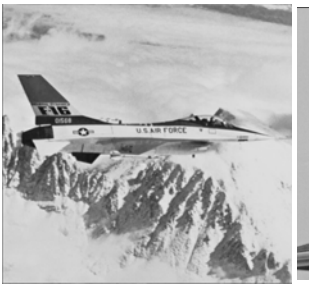
Baboon



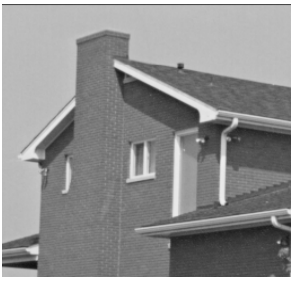
Barcode



Brick



F16-Jet



House

Lena

Pepper

Texture

Tree

Fig. 3. Selected Test Images from USC-SIPI Database

4 Conclusion

This paper presents different techniques for creating reference images to be used for quantitative comparisons of different binarization techniques. Five well-known representative binarization techniques have been selected for experimentation purpose and results show that the performance of a binarization technique varies as the reference image, with which it is being compared, is changed. There are a lot of scopes to extend this work for generating reference images depending on some important properties of the image like texture, spatial distribution of similar gray levels in the images, etc. The authors are consolidating on this issue.

References

1. Shaikh, S.H., Maiti, H.A., Chaki, N.: A New Image Binarization Method using Iterative Partitioning. Springer Journal on Machine Vision and Applications (revised manuscript submitted in July 2011) ISSN: 0932-8092
2. Rodriguez, R.: A Robust Algorithm for Binarization of Objects. Latin American Applied Research 40 (2010)
3. Rodriguez, R.: Binarization of Medical Images based on the Recursive Application of Mean Shift Filtering: Another Algorithm. In: Advances and Applications in Bioinformatics and Chemistry (2008)
4. Ntirogiannis, K., Gatos, B., Pratikakis, I.: An Objective Evaluation Methodology for Document Image Binarization Techniques. In: 8th IAPR Workshop on Document Analysis Systems (2008)
5. Sezgin, M., Sankur, B.: Survey over Image Thresholding Techniques and Quantitative Performance Evaluation. Journal of Electronic Imaging 13(1), 146–165 (2004)
6. Sauvola, J., Pietikainen, M.: Adaptive Document Image Binarization. Pattern Recognition 33(2), 225–236 (2000)
7. Yang, Y., Yan, H.: An Adaptive Logical Method for Binarization of Degraded Document Images. Pattern Recognition 33, 787–807 (2000)
8. Savakis, E. A.: Adaptive Document Image Thresholding using Foreground and Background Clustering. In: Int. Conf. on Image Processing (ICIP 1998), Chicago (October 1998)
9. Gonzalez, R.C., Woods, R.E., Eddins, S.L.: Digital Image Processing Using MATLAB, 2nd edn., ch. 10, p. 513. McGrawHill
10. Niblack, W.: An Introduction to Digital Image Processing, pp. 115–116. Prentice Hall, Eaglewood Cliffs (1986)
11. Bernsen, J.: Dynamic Thresholding of Gray Level Images. In: ICPR 1986: Proc. Intl. Conf. Patt. Recog., pp. 1251–1255 (1986)
12. Kapur, J.N., Sahoo, P.K., Wong, A.K.C.: A New Method for Gray-level Picture Thresholding using the Entropy of the Histogram. In: Graph. Models Image Process., pp. 273–285 (1985)
13. Otsu, N.: A Threshold Selection Method from Gray-Level Histogram. IEEE Transactions on Systems, Man, and Cybernetics 9, 62–66 (1979)
14. University of Southern California, Signal and Image Processing Institute, USC-SIPI Image Database, `http://sipi.usc.edu/database/`

Medical Aid for Automatic Detection of Malaria

Pramit Ghosh[1], Debotosh Bhattacharjee[2], Mita Nasipuri[2], and Dipak Kumar Basu[2]

[1] Department of Computer Science & Engineering, RCC Institute of Information Technology, Kolkata 700015, India
pramitghosh2002@yahoo.co.in
[2] Department of Computer Science. & Engineering. Jadavpur University, Kolkata 700032, India
debotoshb@hotmail.com, {mitanasipuri,dipakkbasu}@gmail.com

Abstract. The analysis and counting of blood cells in a microscope image can provide useful information concerning to the health of a person. In particular, morphological analysis of red blood cell's deformations can effectively detect important disease like malaria. Blood images, obtained by the microscope, which is coupled with a digital camera, are analyzed by the computer for diagnosis or can be transmitted easily to clinical centers than liquid blood samples. Automatic analysis system for the presence of Plasmodium in microscopic image of blood can greatly help pathologists and doctors that typically inspect blood films manually. Unfortunately, the analysis made by human experts is not rapid and not yet standardized due to the operators' capabilities and tiredness. The paper shows how effectively and accurately it is possible to identify the Plasmodium in the blood film. In particular, the paper presents how to enhance the microscopic image and filter out the unnecessary segments followed by the threshold based segmentation and recognize the presence of Plasmodium. The proposed system can be deployed in the remote area as a supporting aid for telemedicine technology and only basic training is sufficient to operate it. This system achieved more than 98% accuracy for the samples collected to test this system.

Keywords: Dilation, Erosion, Field's stain, Gradient operator, HSI colour format, Laplacian Filter, Malaria, Plasmodium.

1 Introduction

Malaria is a mosquito-borne infectious disease of human being caused by a parasite called Plasmodium. It is widespread in tropical and subtropical regions, including much of Sub-Saharan Africa, Asia and America. In the human body, the parasites multiply in the liver (exoerythrocytic phase) [1], and then infect red blood cells (erythrocytic phase). Symptoms of malaria include fever, headache, and vomiting, and usually appear between 10 and 15 days after the mosquito bite. If not treated, malaria can quickly become life-threatening by disrupting the blood supply to vital organs [2]. The accepted laboratory practice for the diagnosis of malaria is the preparation and microscopic examination of blood films stains generated by Giemsa's solution.

N. Chaki and A. Cortesi (Eds.): CISIM 2011, CCIS 245, pp. 170–178, 2011.

Figure 1 shows 1125X Magnified blood sample, which reveals the presence of two Plasmodium vivax parasites; an immature form on the left, and another in a mature form on the right.

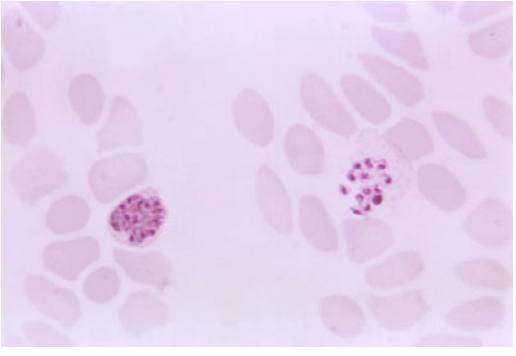


Fig. 1. 1125X Magnified blood sample, with two Plasmodium vivax parasites

Four species of Plasmodium parasites can infect human beings and also they can utilize human beings as temporary repository in transmitting the disease malaria, from one sick individual to another healthy person. Severe disease is largely caused by Plasmodium falciparum [3] and it is also responsible for about 90% of the deaths from malaria. Malaria caused by Plasmodium vivax, Plasmodium ovale and Plasmodium malariae is generally a milder disease that is rarely fatal [4].

The World Health Organization (WHO) reports that over 780,000 people died of malaria in 2009 [5], most of them are children, under the age of five. At a particular time, an estimated 300 million people are said to be infected with at least one of these Plasmodium species. Sometimes it may so happens that malaria becomes an epidemic in the particular time of the year and that generally occurs during rainy season. So, there is a great need for the development of effective treatments for decreasing the yearly mortality and morbidity rates. Once an epidemic like situation emerges, it becomes very difficult to arrange sufficient pathologist and equipments for diagnosis of the Malaria especially in economically backward areas. Some test kits exist like "Malaria home test kit" [6]. But cost of such kit is much higher. As a matter of fact, it is difficult to place such kits in every health care unit. The objective of this work is to design a low cost device which is capable to detect Plasmodium species from blood film images to speed up the diagnosis process.

The rest of the paper is arranged as follows: section 2 describes the proposed system; section 3 presents the results and performance of the system and section 4 concludes the work along with discussions on the future scope of this work.

2 System Detail

The system is explained with the help of a block diagram, shown in figure 2, and all the steps of the system are described next.

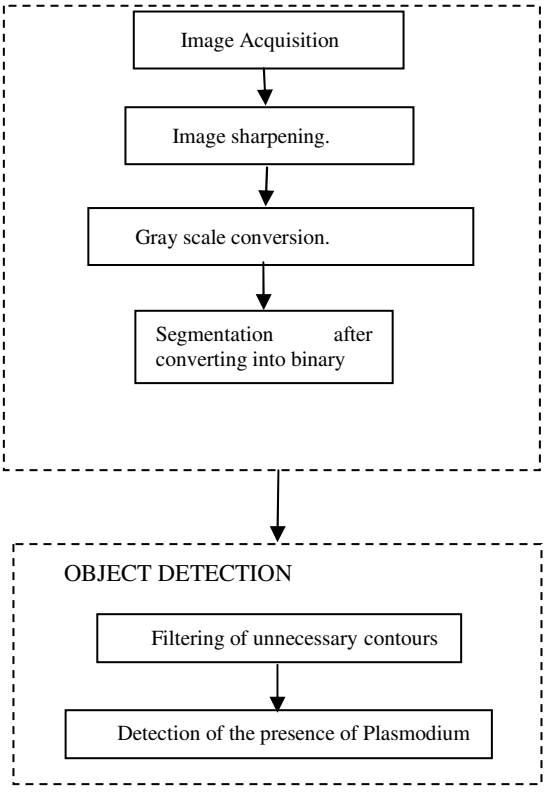


Fig. 2. Block diagram of the system

2.1 Image Acquisition

Image Acquisition is the first step of the system. Digitized images of the blood samples on the slides are acquired with a CCD [7] camera which is mounted upon the microscope. For getting multiple images of a single sample, the glass slide movement is required and it is controlled by two stepper motors [8] in the horizontal and vertical direction shown in Figure 3.

2.2 Image Enhancement

The images obtained from the CCD camera are not of good quality. Laplacian Filter [9] is used to sharpen the edges of the objects in the image. The Laplacian gradient at a pixel position (x,y) is denoted by $\nabla^2 \mathrm{f(x,y)}$ and it is defined as.

$$\nabla^2 \mathrm{f(x,y)} = \frac{\partial^2}{\partial x^2} f(x,y) + \frac{\partial^2}{\partial y^2} f(x,y) \tag{1}$$

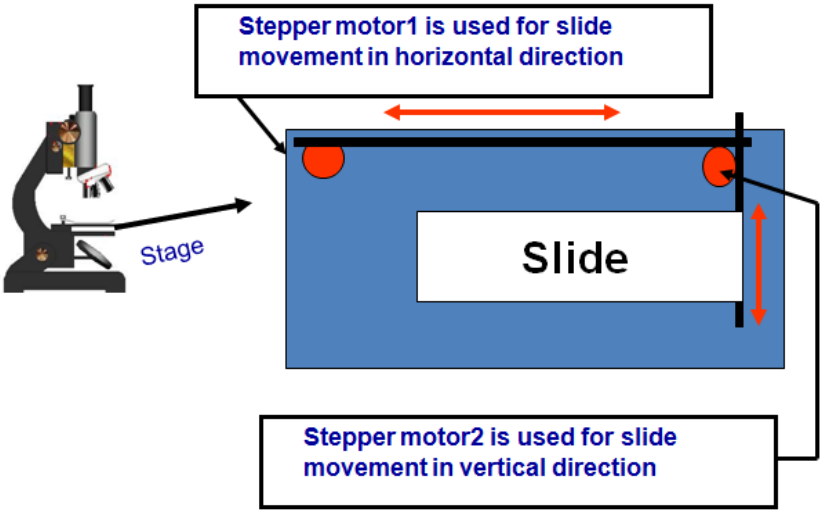


Fig. 3. Slide control through stepper motor

This expression is implemented at all points (x,y) of the image through convolution. The Laplacian filter is applied separately on Red, Green and Blue components of the colour images obtained from the CCD camera. After that, the images are converted into gray scale image by simple average of three components namely Red, Green, and Blue.

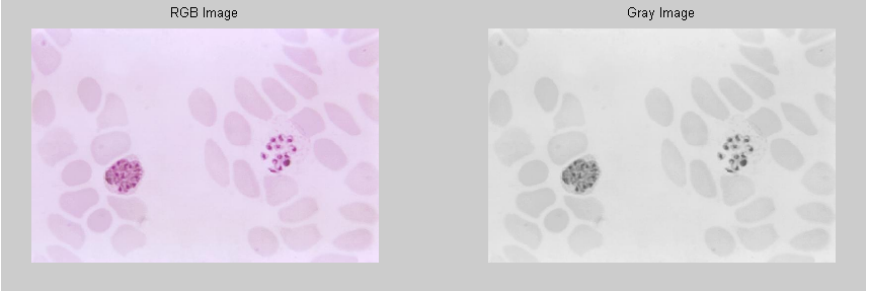


Fig. 4. RGB Image and its corresponding gray scale image

2.3 Segmentation

By analyzing gray scale images it is inferred that the area of the image occupied by the Plasmodium has low gray value and high intensity variation with respect to other area of the image. For the benefit of the implementation of algorithms in better way, all the gray images are converted into corresponding negative images, in which light areas appear dark and vice versa. So in the inverted images the area occupied by the Plasmodium will have high gray value with respect to other area of the image.

All the input blood film images are not illuminated in the same scale. So inverted gray scale image illumination is not in the same scale; this might cause problem in threshold based region segmentation. Uniform scale conversion is required to overcome such type of problem. It is implemented by subtracting a calculated value from each pixel points of the inverted image. The value which is proposed to be subtracted is the 45% of the average gray value of the pixels of the inverted image. This 45% is determined by analyzing data.

Trial and error method reveals that the subtraction is only applicable if the difference between the maximum gray scale value and average gray scale value of the inverted gray image is greater than a threshold value, which is obtained by statistical data analysis. Image histogram is used to find out the maximum value; otherwise noise with high gray value might claim the maximum value.

Algorithm-1: The gray image intensity scaling process.
This algorithm takes a gray image as an input and reduces the intensity of the image for proper scaling of the intensity. The reduction process depends on some parameters.

Step 1: Calculate the mean of the gray values of the inverted image.
Step 2: Calculate the histogram and find out the maximum intensity value of the gray scale where the total numbers of pixels belong to that intensity level is greater than or equal to 1.25% of total number of pixels in the inverted image. This process will eliminate the chance of noise pixel with high intensity value get selected.
Step 3: Find out the difference between two values obtain in step 1 and 2.
Step 4: If result in step3 is greater than predetermined threshold value then subtract 45% of the average value obtained in step 1 from each pixel points of the inverted image.
Step 5: Stop.

After applying algorithm-1, the inverted gray image is converted into binary image using a threshold value [12]. This threshold value varies from image to image. The threshold value calculation algorithm is given next.

Algorithm-2: Calculation of threshold value of the gray image.
This threshold value will be used to convert the gray image into binary image.

Step 1: Select an initial estimate for T (T=threshold value). The initial value for T is the average gray level of the image
Step 2: Segment the images using T. This will produce two groups of pixels: consisting of all the pixels with gray level values > T called as G1 and consisting of pixels with values <= T called as G2.
Step 3: Compute the average gray level values $\mu 1$ and $\mu 2$ for the pixels in regions G1 and G2.
Step 4: Compute a new threshold value: $T = 0.5 * (\mu 1 + \mu 2)$.
Step 5: Repeat steps 2 through 4 until the difference in T in successive iterations is smaller than a predefined parameter T_0.
Step 6: Stop.

2.4 Filter out Unnecessary Contours

The binary image has unnecessary contours that are basically noise. This small noise contours are eliminated by closing, which is dilation[13] followed by erosion [13]. Closing is able to remove unnecessary contours which are small in size but fails to eliminate unnecessary contours of big in size. Figure 5 shows the binary image and its corresponding binary image after removal of small unnecessary contours.

The unnecessary contours which have considerable size are indicating the regions of red blood cells. The difference between red blood cells and Plasmodium is that red blood cells have smooth surface whereas Plasmodium has rough surface area. The gradient operator is helpful to distinguish the picture segment of red blood cell and Plasmodium. The gradient at the center point in a neighborhood is computed as given in [12]

The gradient operator is applied on the inverted gray image and then the output is converted into binary using a threshold value. Figure 6A shows the output along with merging the two binary images given in Figure 6B. The cluster of white region denotes the surface is not smooth.

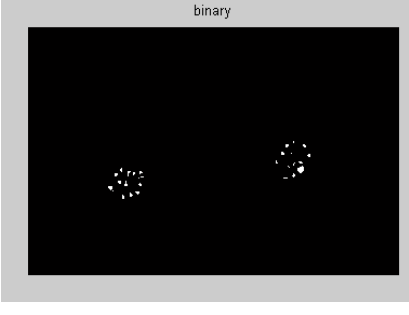


Fig. 5A. Filtered Binary image

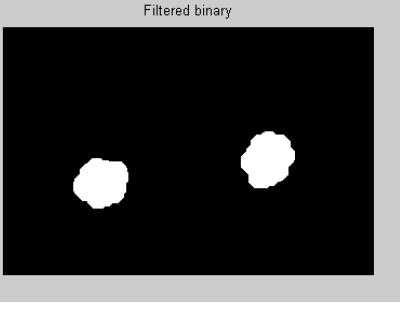


Fig. 5B. After removal of small contours from 5A

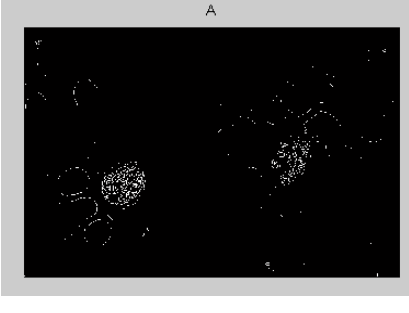


Fig. 6A. Binary image after gradient operator

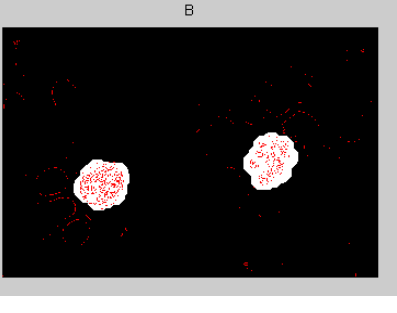


Fig. 6B. Overlapping of figure 5B and 6A

2.5 Detection of the Presence of Plasmodium

By comparing two sets of binary images, obtained from previous sections, the presence of Plasmodium is determined. This is done by the following algorithm.

Algorithm-3: To find out a valid contour of Plasmodium.

Step1: Apply label matrix technique in the binary image obtained from the inverted gray image, and store the output matrix in LB matrix variable. LB will have the same dimension as the binary image.
Step 2: count = maximum integer value stored in LB; So "count" will contain the number of contours in the input image.
Step 3: index = 1
Step 4: val = (total number of points, where value is 1, in the binary image, which is obtained after applying gradient operator) / (total number of points in the binary image)
Step 5 : Find the coordinates of the pixels of the LB where value of the pixel is equal to index;
Step 6: local_Value = (find out the number of points, where value is 1, in the binary image, which is obtained after applying gradient operator and whose coordinates are selected in step 5.) / (total number of coordinates are selected in step 5.)
Step 7: if local_Value > >val
then contour of a Plasmodium is found
Else Not a valid plasmodium contour
Step 8 : index = index + 1;
Step 9: Repeat step 5 to step 8 until index > count
Step 10 : Stop

3 Simulation and Results

For simulation MatLab 7.1 [14] is used. The algorithms are applied on the images available in the database of Centers for Disease Control and Prevention [15]. The image shown in Figure 1 is fed as an input. After analysis, system finds the presence of Plasmodium. Output is shown next.

```
Value = 0.0088
local_Value = 0.1559
ans = Plasmodium found
local_Value = 0.1023
ans = Plasmodium found
```

Figure 7 is another test image where no Plasmodium is present Output is shown next.

```
Value =    0.0228
local_Value =.0.0392
ans =Plasmodium not found
```

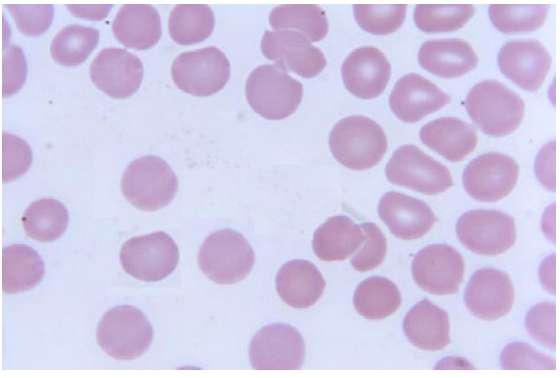


Fig. 7. Blood sample, with no Plasmodium vivax parasite

For testing purpose 160 samples were collected and out of which 157 samples have shown accurate results. So the accuracy of the system is found to be 98.125%.

4 Conclusion

In this paper one novel approach for detection a malarial parasite, called Plasmodium, is proposed. This system is cheaper than other Malaria test kit. This system does not require any special technical skill. So, this can be used by the people of remote places with very basic level of education. It may reduce the probability of wrong treatment which happens due to non-availability of diagnosis systems in remote and economically backward areas. As a next phase authors are trying to design an integrated system for diagnosis of diseases due to such parasites.

Acknowledgment. Authors are thankful to Dr. Abhijit Sen and Dr Saumendu Datta for providing pathological data and the "DST-GOI funded PURSE program", at Computer Science & Engineering Department, Jadavpur University, for providing infrastructural facilities during progress of the work.

References

1. Malaria details, http://www.malaria.com
2. Malaria, World Health Organization, http://www.who.int/topics/malaria/en/
3. Life Cycle of Plasmodium falciparum, World Health Organization, http://www.searo.who.int/en/Section10/Section21/Section340_4269.htm
4. Malaria details, World Health Organization, http://www.searo.who.int/en/Section10/Section21/Section334.htm
5. Malaria Report 2009 from World Health Organization, World Health Organization (2009), http://www.who.int/malaria/world_malaria_report_2009/en/index.html

6. Malaria home test kit, http://www.anytestkits.com/malaria-test-kit.htm
7. Holland, S.E.: Fully Depleted Charge-Coupled Devices. Stanford University, http://www.slac.stanford.edu/econf/C0604032/papers/0009.PDF
8. Stepper motors, Embedded System Design Laboratory, Stanford University, http://www.stanford.edu/class/ee281/handouts/lecture11.pdf
9. Gonzalez, R.C., Woods, R.E.: Digital Image Processing, 2nd edn. Pearson Education (2002)
10. COLOR_CONVERT, NASA, The U.S. Space research organization, http://idlastro.gsfc.nasa.gov/idl_html_help/COLOR_CONVERT.html
11. Ghosh, P., Bhattacharjee, D., Nasipuri, M., Basu, D.K.: Round-The-Clock Urine Sugar Monitoring System for Diabetic Patients. In: International Conference on Systems in Medicine and Biology, pp. 326–330. IEEE Computer Society, IIT Kharagpur (2010)
12. Gonzalez, R.C., Woods, R.E., Eddins, S.L.: Digital Image Processing Using MATLAB, 2nd edn. Pearson Education (2004)
13. Chakravarthy, B.: Mathematical Morphology, Indian Statistical Institute, http://www.isibang.ac.in/~cwjs70/serra-wkshp.pdf
14. MatLab, http://www.mathworks.com/
15. Centers for Disease Control and Prevention, http://phil.cdc.gov/phil/home.asp

Measuring Leanness and Agility Status of Iranian Food Industries Supply Chains Using Data Envelopment Analysis

Pouyeh Rezazadeh

Department of Industrial and System Engineering,
South Tehran Branch, Islamic Azad University, 11518-63411 Tehran Iran
info.rezazade@gmail.com

Abstract. This research proposes a methodology to measure the overall leanness and agility of supply chains considering the most appropriate index of supply chains in food industry. Output and input of proposed model base in Data Envelopment analysis are identified from literature review and the level of them from a survey questionnaire accrued. Using the Data Envelopment Analysis technique, the leanness and agility measures delivers a self-contained, unit-invariant score of the whole supply chain system to support decision making on continuous improvement. And firms could adopt either a lean or an agile strategy or both, depending on the environment. This article provides a DEA method to measure two supply chain strategies and benchmarks their indexes to co-align competitive strategies with the environment to improve performance. This approach has been applied in case of some Iranian food industry supply chains to prove the applicability of the method.

Keywords: Data Envelopment Analysis, Lean Supply Chains, Agile Supply Chains, Iranian Food Industries.

1 Introduction

The risk attached to traditional forecast driven lengthy supply line has become untenable for consumer products A key feature of present day business is the idea that it is supply chains (SC) that compete, not companies [8], and the success or failure of supply chains is ultimately determined in the marketplace by the end consumer. Enterprises are continuously paying attention in responding to the customer demand for maintaining a competitive advantage over their rivals. Supply Chain Management (SCM) has gained as it focuses on material, information and cash flows from vendors to customers or vice-versa. Conceptual and empirical research had great outcomes that there are distinct supplies chains types that may be labeled as lean, agile and 'leagile', which informs strategic choice [8-23]. For supply chain efficient management we have acknowledged decoupling point in the value chain. Models and tools have been developed that may inform decision makers in deciding what is the right supply chain for their business. for choosing Strategy in decoupling point of

N. Chaki and A. Cortesi (Eds.): CISIM 2011, CCIS 245, pp. 179–191, 2011.

both strategies it is needed to study tools and levels of current status of lean and agile supply chain and quantified evaluation for both strategies. but some indexes have conflicts and inconsistent in implementation and tools of other indexes ,this case study research has presented two level quantified for current status and are not consolidate them in one parameter. [9,15]. The supply strategies are denied according to the adoption of the various integration mechanisms, providing definitions and classifications based more on the practice that on the goals of supply. These dimensions are jointly considered by the literature on supply models, such as Lean supply 0 and agile supply [8] which described according multiple dimensions, but each one separately and mainly with case-based evidence.

However, the literature lacks extensive research that considers both supplier selection criteria and integration mechanisms to identify alternative supply strategies. Besides, several authors claim that supply strategies effectively impact the firm's performance; however, very few demonstrated this relationship, and only within the limits previously mentioned [23]. Hence this paper provides a single point of reference for other researchers exploring such issues as supply chain categorization especially with regard to lean, agile system measurement. And try to measure the metrics simultaneously. Thus, firms need to possess a clear strategic planning in order to effectively organize such complicated activities, resources, communications, and processes [5].

There are extensions to the material flow and decoupling point based 'leagile' type that incorporate other considerations, including product type and the associated product life cycle and due consideration of spatial and temporal applications of lean and agile. Recently, lean and agile thinking in manufacturing has been extended to the broader supply chain. the major objective of a lean supply chain strategy is to reduce cost and enhance efficiency through elimination of wastes in both inter- and intra-organizational processes. Lean supply chains are best matched with a relatively stable environment. However, one objective of an agile supply chain is to enrich the customer [24], an agile firm must be able to quickly increase its level of customization and customer service while accommodating different delivery expectations. This paper introduces an approach to measure the key characteristics of supply chains to find the level of preformation and efficiency through model evaluation.

2 Literature Review

Tools used to measure supply chain efficiency had received numerous attentions. The "spider" or "radar" diagram and the "Z" chart are some of the popular tools used to measure supply chain efficiency. These tools are based on gap analysis techniques and they are very graphical in nature. Although the graphical approaches make them easy to understand, it causes inconveniences to the analysts if multiple elements have to be integrated into a complete picture. In other words, it is not feasible to measure the efficiency using these tools when there are multiple inputs or outputs. Another well-known method used is the ratio. It computes the relative efficiencies of the outputs versus the inputs and is easily computed.

However, a problem with comparison via ratios is that when there are multiple inputs and outputs to be considered, many different ratios would be obtained and it is difficult to combine the entire set of ratios into a single judgment. The evaluation of supply chain efficiency needs to look into multidimensional construct as the results of the evaluation is important for fine-tuning an organization current operations and creating new strategies to keep up with competitions [23-32]. Single output to input financial ratios such as return on sales and return on investment may not be adequate for use as indices to characterize the overall supply chain efficiency.

Hence, the traditional tools discussed earlier, which do not take into account multiple constructs, would not be able to provide a good measure of supply chain efficiency [7]. Since, a company's supply chain efficiency is a complex phenomenon requiring more than a single criterion to be characterized, a number of studies have suggested a multi-factor performance measurement model may be applied for the evaluation of supply chain efficiency [11]. The development of a multi-factor performance measure, which reflects the efficiency of functional units and technologies in a supply chain, is important to policy makers by knowing how far a industry or company can be expected to increase its multiple outputs and decrease input level through the improvement of its efficiency.

Considering these reasons this research has selected data envelopment analysis and has shown much interest in the DEA. Functional products characterized by, e.g. a steady demand pattern and long product life cycles should be managed in a physically efficient supply chain that focuses on cost minimization and high utilization of resources, whereas innovative products with demand volatility and short life cycles should be transformed through a market- responsive supply chain that has extra capacity, capability of market demand information processing, and that is more flexible [9, 33]. Based on a categorization related to product characteristics in terms of volume, variety, variability, delivery window and stage in product life cycle, they show that each manufacturing pipeline may be classified as make-to-order, lean, agile and leagil [15,18]. Swafford et al.,(2008) [36] argue that one important driver for agile supply chain is "mass customization", where the company need to provide "customaries" products and services at a cost equal to or even close to the costs associated to mass production".

To aid in the selection of appropriate supply chain types for industries, Christopher et al. (2006) [9] develop a taxonomy typification based on lead-time and the predictability of demand. This could help this research to Categorize the index and sub indexes and conforming to food industry and the most structure of Iranians food industry entities. Kumar and Motwani (1995) [22] annotate Supply chain agility As a firm's ability to accelerate the activities on the critical path can be measured as a composite value of the strategic agility position of a firm on a percentage scale, based on the weighted sum of the firm's performance on each element of a matrix, that represents all combinations of time-segments and agility determinants (material and information flow, state of technology, specialized functions, human resource factors, quality, and flexibility).

Devor et al. (1997) [22]0 addressed the agility the ability of a producer of goods and services to operate profitability in a competitive environment of continuous and unpredictable change, Also, Quinn et al. (1997) [21] pursue the ability to accomplish rapid changeover from the assembly of one product to the assembly of a different product. Tan et al., [42] mentioned the specially agility index in supply chain such as Supply chain integration Information sharing, Supply chain characteristics, Customer service management, Geographical proximity. Ulusoy [1] reviews the specification of supply chain strategy in Logistics, Supplier relations, and Customer relations Production. Min and Mentzer [10], pursue that a supply chain can achieve the organization strategy through Vision and goals that are agreed upon: Information sharing, Cooperation, Process integration, Long term relationships and Supply chain leadership. Chen and Paulraj, [19] addressed the supply chain attributes (both lean and agile regions) Supplier-based reduction, Long-term relationships, Communication, Cross-functional teams, Supplier involvement. Li et al., [19] mentioned to Strategic supplier partnerships, Customer relationships, Information sharing, Information quality Internal lean practices, Postponement, Delivery dependability, Time to market as the high priority criteria in supply chain management. Burgess et al., [6] results conduce in sub index which have been used in this research as principles to design questionnaire and lead survey Leadership, Intra-organizational relationships, Inter-organizational relationships, Logistics, Process improvement orientation , Information systems , Business results and outcome. Zhou and Benton, [26] annotate the Supply chain planning in the JIT area Delivery practice Also Koh et al., [30] have hinted the rather efficiency of supply chain in Close partnership with suppliers, Close partnership with customers And JIT supply.

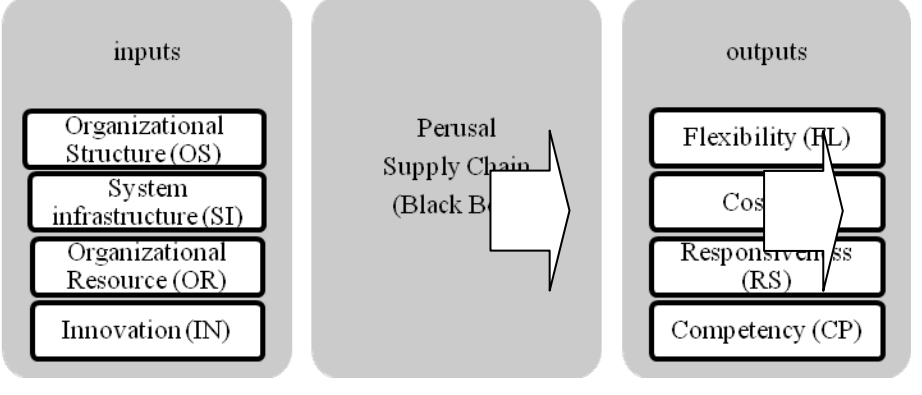


Fig. 1. Conceptual model of agility index in Supply Chain as a DMU

Of course to help managers to attain a sustainable competitive advantage, and reduce waste and cost numerous studies have attempted discuss agility in organizations. Our definition for agility is the ability of supply chain as a whole and its member rapidly align the network and its operation to dynamics and turbulent requirements and of customers. In [8], Christopher has identified that agile supply

chain (ASC) requires various distinguishing capabilities to respond changing environments. It has been categorized the attribute reviewed in introduction to create a structure of attributes to be helpful in designing a questionnaire.

The lean approach can be defined to the SC upstream of the decoupling point as the demand is smooth and standard products flow through a number of value streams. If a SC has a long lead-time it will not be able to respond quickly to demand. Lead-time needs to be minimized in lean manufacturing, because in lean definition time is waste and leanness tries to eliminate all waste. The difference between leanness and agility is that service is the critical factor for agility whilst cost and sales price, are crucial for leanness [21, 22].

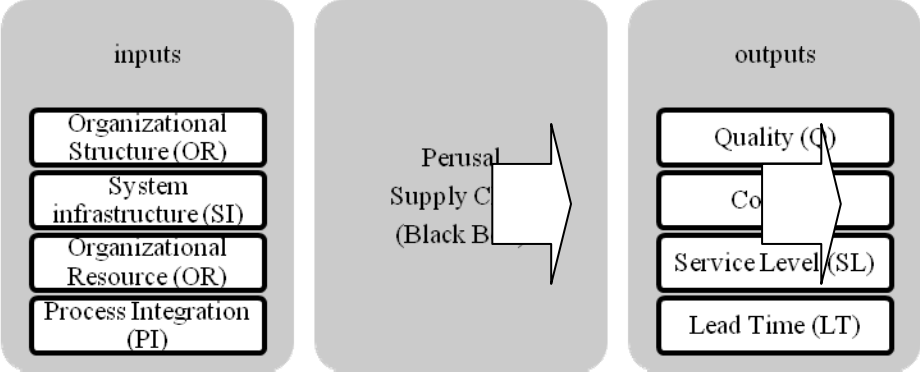


Fig. 2. Conceptual Model of Leanness index in Supply Chain as a DMU

3 Lean and Agile Supply Chain Conceptual Model

Different levels of agility and leanness can be measured in supply chains with the input and output oriented indices, agility and leanness retrieved through literature review. Some framework has been reported in literature for measuring the agility of organizations and supply chains. Metrics may consider common concepts of agility have been categorized and classified through different logics. Sharifi and Zhang (1999) [21] proposed a framework for agility measurement through drivers, providers, & capabilities in form of a framework.

Lin et al. (2006) [19] proposed a conceptual framework for agile enterprise based on agility drivers, agility capabilities, and agility enablers. We consider two types of factors affecting agility and leanness of supply chain as capabilities and providers of both through literature review. The capabilities are assumed to represent and measure the current and existing level of agility in supply chain. Also leanness providers are assumed to measure the potential level leanness in supply chain. Distinctly providers can be treated as input oriented factors of legality.

As mentioned after study indexes and sub-indexes and also under view lean and agile model structure, the results are shown and summarized in the Fig 1and Fig 2. and are ready to use in DEA-model and measuring step.

The quantity of each index can achieve from the questionnaire as explained and the result averages are utilized in DEA transformation measuring model. We have 1 input special for each model (innovation and process integration) and 3 common inputs and 3 special outputs for each one .There is 1 common output (cost) for each model which achieved from literature review.

4 Industry Application and Problem Context

As mentioned, measuring performance of agility and leanness is an essential and surviving factor for supply chains. agility and leanness and measuring efficiency of transformation process have a high priority in context of supply chain which produce low shelf life goods. Dairy industries are good testimonial of this situation in which the main processes of supply (i.e. new product development, sourcing, production, distribution & delivery) should be accomplished in agile lean context. More formally, as products of supply chains in dairy industries have a low shelf-life so, proper response to unexpected variations in customer demands through low cost, high speed, desirable quality which has been provable as the main aspect of agility and leanness, plays a substantial role. Iranian dairy industries with more than 200 active companies which supply their products through different chains provide a great context [6,10]. We applied the DEA model in assessment of efficiency of agility and leanness levels in fifteen food industry in Iran which supply products through different chains. Each dairy company has been assumed as an independent supply chain with aforementioned two levels of agility and leanness. So the proposed DEA model has properly been fitted.

4.1 Data Gathering and Survey

We used the experimental experiences of managers of selected diary supply chains to determine the values of agility and Leanness factors. The data collected from experts of Iranian diary supply chains through questionnaires. The experts who filled the questionnaires were experienced managers working for diary supply chains.

These managers had 10 years of rating experience on average. A set of 30 diary supply chains was selected and a manager of a given supply chain was requested to rate the affecting factors of agility and leanness (i.e. providers of agility and leanness, capabilities of agility and leanness) for all of their supply chains. These managers were left free to rate each question answer according to likert scale rating.

Since the data for this research was derived from questionnaire survey responses it was necessary to assess the reliability of the query. Having confirmed statistically that the questionnaire data are devoid of random effects, reliability tests were conducted as a measure of the internal consistency of instruments employed to measure concepts. For instruments measuring a concept to be reliable, they should be highly correlated. Cronbach's coefficient alpha, which computes an average of all possible split-half estimates, is the most widely used test of internal consistency.

Moreover, data reliability requires that instruments measuring the same concept should be sufficiently different from other instruments. Also mentioned predefinition was written in the beginning of the questionnaire and for each index so each manager could answer with enough knowledge. Between 10 to 20 question were designed for each index, (drafting 1 to 3 question for each sub-index), and the questionnaire were distributed between 20 purchasing and logistic manager directly related to supply chain. .

Reliability tests were conducted for the main elements of the research instruments, they are lean characteristics, agility attributes (inputs, outputs) as well as the entire questionnaire. This shows that the Cronbach's alpha for the overall scale of the survey leanness and agility consisting of 12 variables was found to be 0.849. Using results of earlier empirical studies, report that while Cronbach's alpha at 0.70 or higher is typically used to establish reliability of a construct, through there are situations in which values of 0.6 are acceptable, especially for broadly defined constructs like Agility and Lean attributes.

4.2 Review of DEA and Its Applications in Supply Chains

DEA is a mathematical programming technique that calculates the relative efficiencies of multiple DMUs based on multiple inputs and outputs. DEA measures the relative efficiency of each DMU in comparison to other DMUs. An efficiency score of a DMU is generally defined as the weighted sum of outputs divided by the weighted sum of inputs, while weights need to be assigned [7, 34]. DEA model computes weights that give the highest possible relative efficiency score to a DMU while keeping the efficiency scores of all DMUs less than or equal to 1 under the same set of weights. The fractional form of a DEA mathematical programming model is given as follows:

$$\text{Maximize } h_0 = \frac{\sum_{r=1}^{t} u_r y_{rj0}}{\sum_{i=1}^{m} v_i x_{ij0}} \tag{1}$$

$$\text{subject to: } \frac{\sum_{r=1}^{t} u_r y_{rj0}}{\sum_{i=1}^{m} v_i x_{ij0}} \leq 1, j = 1, \ldots . n$$

$$u_r \geq \varepsilon, \quad r = 1, \ldots, t \qquad v_i \geq \varepsilon, i = 1, \ldots,$$

where u r , the weight for output r; v i , the weight for input i; y rj , the amount of output r of DMU j; x ij , the amount of input i of DMU j; t, the number of outputs; m, the number of inputs; n, the number of DMUs; and, a small positive number. The objective function of equation (1) seeks to maximize the efficiency score of a DMU j0 by choosing a set of weights for all inputs and outputs. The first constraint set of equation (1) ensures that, under the set of chosen weights, the efficiency scores of all DMUs are not more than 1. The second and third constraint sets of equation (1) ensure that no weights are set to 0 in order to consider all inputs and outputs in the model. A DMUj0 is considered efficient if the objective function of the associated (equation (1)) results in efficiency score of 1, otherwise it is considered inefficient. By moving the denominator in the first constraint set in equation (1) to the right-hand

side and setting the denominator in the objective function to 1, (equation (1)) can be converted into a LP problem as follows:

$$\text{Maximize} \quad h_0 = \sum_{r=1}^{t} u_r y_{rj0} \tag{2}$$

$$\text{subject to:} \quad \sum_{i=1}^{m} v_i x_{ij0} = 1$$

$$\sum_{r=1}^{t} u_r y_{rj} - \sum_{i=1}^{m} v_i x_{ij0} \leq 0, j = 1, \ldots$$

$$u_r \leq \varepsilon, \quad r = 1, \ldots, t \qquad v_i \leq \varepsilon, i = 1, \ldots,$$

The dual model of equation (2) can be given as follows, which is equivalent to the envelopment form of the problem:

$$\text{Minimize} \quad \theta - \varepsilon\left(\sum_{i=1}^{m} s_i^- + \sum_{r=1}^{t} s_i\right) \tag{3}$$

$$\text{Subject to:} \quad \sum_{j=1}^{n} \lambda_j x_{ij} + s_i^- = \theta x_{ij0} i = 1, \ldots,$$

$$\sum_{j=1}^{n} \lambda_j y_{ij} + s_r^+ = y_{rj0}, r = 1, \ldots, t$$

$$\lambda_j, s_i^-, s_r^+ \geq 0$$

where θ, λ_j, S_i^-, S_r^+ are the dual variables. The variable θ is the technical efficiency score which we want to calculate, and S_i^-, S_r^+ are the input slacks and output slacks, respectively, Output slacks indicate how much shortages in the outputs, while input slacks indicates how much surpluses in the inputs. The slacks and efficiency are closely related, as the prior helps to decide on the later. Based on equation (3), a DMU_j0 is efficient if and only if, in the dual optimal solution, $\theta^* = 1$, and $s_i^{*-} = s_r^{*+} 0$ for all i and r, where an asterisk denotes a solution value in an optimal solution.

In this case, the optimal objective function value of equation (3) is 1,and the corresponding primal problem in equation (2) also has an optimal objective function value of 1. For an inefficient DMU_j0, appropriate adjustments to the inputs and outputs can be applied in order to improve its performance to become efficient. The following input/output adjustments (improvement targets) would render it efficient relative to other DMU_s:

$$x_{ij0}^{I} = \theta^* x_{ij0} - s_i^{-*}, i = 1, \ldots,. \tag{4}$$

$$y_{rj0}^{I} = y_{rj0} - s_i^{+*}, r = 1, \ldots \tag{5}$$

From the duality theory in LP, for an inefficient DMU_j0, $\lambda^* > 0$ in the optimal dual solution also implies that DMU i is a unit of the peer group. A peer group of an inefficient DMU j 0 is defined as the set of DMUs that reach the efficiency score of 1 using the same set of weights that result in the efficiency score of DMU_j0 .

The improvement targets given in equations (4) and (5) are obtained directly from the dual solutions. This is because the constraints in equation (3) relate the levels of

outputs and scaled inputs of DMU j 0 to the levels of the outputs and inputs of a composite DMU formed by the peer group. These improvement targets are considered "input-oriented" because the emphasis is on reducing the levels of inputs to become efficient. If outputs need to be increased during the efficiency improvement, output-oriented adjustments can be used. In this study, it is appropriate to use "input oriented" because we want to evaluate the technical efficiency based on a given set of inputs, while keeping at least the present output levels. In addition, managers also have more control over the inputs compared to the outputs. The dual model of the above formulation is also known as the envelopment model.

It has the benefit of solving the LP problem more efficiently than the primal model, when the number of DMUs is larger than the total number of inputs and outputs, which is normally the case in applying DEA. More importantly, the dual variables provide improvement target for an inefficient DMU to become efficient as compared to the efficient DMUs. An additional convexity constraint $\sum_{j=1}^{n} \lambda_j =$ can be added to equation (3) to yield a measure of the pure technical efficiency if the constant return-to-scale (Banker et al., 1984) assumption does not apply. The above model in equation (3) is used to calculate the technical efficiency. This model can also be referred to as the technical efficiency model and has been used to this research measuring.

4.3 Solving Model and Results

Tables 5 and 6 has shown The mean result of query has been used in model solving through MAX-DEA solver .after defining the DEA model in equation 3 the software has been used for entering data which the results is shown in Table 7 and 8 which the efficiency and slacks measured. The efficient units in lean SCs are 1,2,3,6,7,8,9,10,1,15,14 and in agile SCs are 1,2,3,4,6,7,9,10,14,15. It is shown that the standard deviation of the FL, RS are quite large. This may be due to the fact that FL and RS varies among different companies and environments the total number of DMUs which are being evaluated is 15. for all the DMUs. The technically efficient DMUs are DMU 1, 2, 3,6,7,9,10,14,15.

The selection of peers for each DMU is based on a linear combination of weights given to the nearest efficient. Minimal resource could for survey and company, DMU 7, ultimately is the most appropriate DMU because in final Evaluation in calculation of θ.

This represent the characteristics of food industry in Iran (Level of training, manpower utilization) and also lowest rate size in OS (organization Structure) we will achieve according to the questionnaire we will achieve an analysis of the current supply chain status, in conversion of its resources into key parameters of leanness (such as Service Level and cost) and agility (such as responsiveness and Flexibility) aided with needed tools.

5 Conclusion and Future Study

This research is a background to study contemporary and simultaneously the leanness and agility status of Iranian supply chains, the benefits of growth advantage and reducing weakness of each other ,hence the food industry of Iran can benefitly utilize the both tools and approach to gain both strategy.

This research and experiment indicate the status of supply chains in Iran to native Iranians industry managers Specially food industry to dissect current status of food industry supply chains. This study is initiated by the author because there is a lack of tools to measure supply chain leanness and Agility in Iran. DEA has been proven to be a reliable, flexible and efficient tool in measuring attributes of supply chain performance in lean and agile criteria. The study examines a model of efficiency which is the technical efficiency slack based model.

The technical efficiency model provides the measurement for efficiency while the slack base efficiency provides measurement for oversupply and deficiency of allocation efficiencies. the information obtained from these two models helps managers to identify the inefficient operations and take the right remedial actions for continuous improvement. In order to demonstrate the usefulness of these two models, data were collected from various supply chains and were asked from supply chains and logistic managers.

The results obtained from the analysis indicate that not all technically efficient companies are allocation efficient. This corresponds to the theoretical concepts of efficiencies which clearly distinguished the efficiencies between the technical efficiency and cost efficiency models. Also the results analysis support the validity of the models. Managers need to allocate their resources effectively and efficiently so that the best of outputs can be achieved.

In this research Cost minimization as well as infrastructure metrics maximization can be achieved if managers have the right tools for decision making and the correct information on hand. The opportunity cost derived from the model proves to be very useful information for managers. This piece of information used in combination with analysis on the input mix allocation will greatly help managers to make better decisions in resource planning and Process integration. The contribution of this study provides useful insights into the use of DEA as a modeling tool to aid managerial decision making in measuring supply chain Leanness and agility.

Future work of this study could look into the possibility of modeling DEA in a stochastic supply chain environment since lean and specially agile supply chain operates in a dynamic environment. In addition, it will also be interesting to look into evaluating the stochastic DEA model in multiple time period in order to examine whether there is any technological influence on the supply chain efficiency. Also we can use sensitivity analyzing for comparing the precision of enhancement and deduction quantity of effective resources in supply chain to increase the capabilities of supply chain in both field. Also, this research can continue in other context specially pharmaceutical industry.

References

1. Turner, K.: Modelling complexity in the automotive industry supply chain. Journal of Manufacturing Technology Management 16(4), 447–458 (2005)
2. Agarwal, A., Shankar, R., Tiwari, M.K.: Modelling the metrics of lean, agile and leagile supply chain: an ANP-based approach. European Journal of Operational Research 173, 221–225 (2006)
3. Banker, R.D., Charnes, A., Cooper, W.W.: Some models for estimating technical and scale inefficiencies in data envelopment analysis. Management Science 30(9), 1078–1092 (1984)
4. Baramichai, M., Zimmers Jr., E.W., Marangos, C.A.: Agile supply chain transformation matrix:an integrated tool for creating an agile enterprise. Supply Chain Management: An International Journal 12(5), 334–348 (2007)
5. Bruce, M., Daly, L., Towers, M.: Lean or agile. A solution for supply chain management in the textiles and clothing industry. International Journal of Operations and Production Management 24(2), 151–170 (2004)
6. Burgess, T., Hwarng, B., Shaw, N., De Mattos, C.: Enhancing value stream agility: the UK speciality chemical industry. European Management Journal 20(2), 199–212 (2002)
7. Chen, C.-M.: Evaluation and Design of Supply-Chain Operations using DEA, PhD Thesis, Erasmus Research Institute of Management – ERIM Rotterdam School of Management (RSM), Erasmus School of Economics (ESE) (June 2009), Print: Haveka `http://www.haveka.nl`
8. Christopher, M.: The agile supply chain. Industrial Marketing Management 29, 37–44 (2000)
9. Christopher, M., Peck, H., Towill, D.R.: A taxonomy for selecting global supply chain strategies. International Journal of Logistics Management 17(2), 277–287 (2006)
10. Cox, A., Chicksand, D.: The limits of lean management thinking: multipleretailers and food and farming supply chains. European Management Journal 23(6), 648–662 (2005)
11. de Treville, S., Shapiro, D., Hameri, A.-P.: From supply chain to demand chain: the role of lead time reduction in improving demand chain performance. Journal of Operations Management 21, 613–627 (2003)
12. Eyong Michael, E.: Creating A Competitive Supply Chain: Evaluating The Impact Of Lean & Agile Supply Chain, School Of Innovation. Design & Product Development (Idt) Se – 721 23, Västerås/Eskilstuna, Sweden (2010)
13. Behrouzi, F., Wong, K.Y.: Lean performance evaluation of manufacturing systems: A dynamic and innovative approach. Procedia Computer Science 3, 388–395 (2011)
14. Gil, N., Tommelein, I.D., Stout, A., Garrett, T.: Embodying product and process flexibility to cope with challenging project deliveries. Journal of Construction Engineering and Management 131(4), 439–448 (2005)
15. Goldsby, T.J., Griffis, S.E., Roath, A.S.: Modelling lean, agile, and leagile supply chain strategies. Journal of Business Logistics 27(1), 57–79 (2006)
16. Gunasekaran, A., Ngai, E.W.T.: Information systems in supply chain integration and management. European Journal of Operational Research 159, 269–295 (2004)
17. Khalili-Damghani, K., Taghavifard, M., Olfat, L., Feizi, K.: A hybrid approach based on fuzzy DEA and simulation to measure the effciency of agility in supply chain. International Journal of Management Science and Engineering Management 6(3), 163–172 (2011) ISSN 1750-9653

18. Krishnamurthy, R., Yauch, C.A.: Leagile manufacturing: a proposed corporate infrastructure. International Journal of Operations and Production Management 27(6), 588–604 (2007)
19. Li, G., Lin, Y., Wang, S., Yan, H.: Enhancing agility by timely sharing of supply information. Supply Chain Management: An International Journal 11(5), 425–435 (2006)
20. Mohammed, I.R., Shankar, R., Banwet, D.K.: Creating flex-lean-agile value chain by outsourcing, An ISM-based interventional roadmap. Business Process Management Journal 14(3), 338–389 (2008)
21. Mason-Jones, R., Naylor, J.B., Towill, D.R.: Lean, agile or leagile? Matching your supply chain to the market place. International Journal of Production Research 38(17), 4061–4070 (2000)
22. Mason-Jones, R., Towill, D.R.: Total cycle time compression and the agile supply chain. International Journal of Production Economics 62(1), 61–73 (1999)
23. Narasimhan, R., Swink, M., Kim, S.W.: Disentangling leanness and agility: an empirical investigation. Journal of Operations Management 24, 440–457 (2006)
24. Naylor, J.B., Naim, M.M., Berry, D.: Leagility: Integrating the lean and agile manufacturing paradigms in the total supply chain. International Journal of Production Economics 62, 107–118 (1999); formerly (1997) OccasionalPaper#47
25. Papadopoulou, T.C., Ozbayrak, M.: Leanness: experiences from the journey to date. Journal of Manufacturing Technology Management 16(7), 784–807 (2005)
26. Pettersen, J.: Defining lean production: some conceptual and practical issues. TQM Journal 21(2), 127–142 (2009)
27. Power, D.J., Sohal, A.S., Rahman, S.U.: Critical success factors in agile supply chain management: an empirical study. International Journal of Physical Distribution & Logistics Management 31(4), 247–265 (2001)
28. Prince, J., Kay, J.M.: Combining lean and agile characteristics: creation of virtual groups by enhanced production flow analysis. International Journal of Production Economics 85, 305–318 (2003)
29. Rickards, R.: Setting benchmarks and evaluating balanced scorecards with data envelopment analysis. Benchmarking: An International Journal 10(3), 226–245 (2003)
30. Sanchez, L.M., Nagi, R.: A review of agile manufacturing systems. International Journal of Production Research 39(16), 3561–3600 (2001)
31. Seydel, J.: Data envelopment analysis for decision support. Industrial Management & Data Systems 106(1), 81–95 (2006)
32. Shah, R., Ward, P.T.: Defining and developing measures of lean production. Journal of Operations Management 25, 785–805 (2007)
33. Shaw, N.E., Burgess, T.F., de Mattos, C., Stecy: Supply chain agility: the influence of industry culture on asset capabilities within capital intensive industries. International Journal of Production Research 43(15), 3497–3516 (2005)
34. Sherman, H.D., Ladino, G.: Managing bank productivity using data envelopment analysis (DEA). Interfaces 25(2), 60–73 (1995)
35. Slack, N.: The flexibility of manufacturing systems. International Journal of Operations & Production Management 7(4), 35–45 (1987)
36. Swafford, P.M., Ghosh, S., Murthy, N.N.: A framework for assessing value chain agility. International Journal of Operations and Production Management 26(2), 118–140 (2006)
37. van der Vaart, T., van Donk, D.P.: Buyer focus: evaluation of a new conceptfor supply chain integration. International Journal of Production Economics 92, 21–30 (2004)
38. van der Vorst, J.G.A.J., van Dijk, S.J., Beulens, A.J.M.: Supply chain design in the food industry. International Journal of Logistics Management 12(2), 73–85 (2001)

39. van Hoek, R.I., Harrison, A., Christopher, M.: Measuring agile capabilities in the supply chain. International Journal of Operations and Production Management 21(1/2), 126–147 (2001)
40. Wan, H.-D., Frank Chen, F.: A leanness measure of manufacturing systems for quantifying impacts of lean initiatives. International Journal of Production Research 46(23), 6567–6584 (2008), doi:10.1080/00207540802230058
41. Yang, B., Burns, N.: Implications of postponement for the supply chain. International Journal of Production Research 41(9), 2075–2090 (2003)
42. Yang, B., Burns, N.D., Backhouse, C.J.: Postponement:are view and an integrated framework. International Journal of Operations and Production Management 24(5), 468–487 (2004)
43. Zhu, J.: Multi-factor performance measure model with an application to Fortune 500 companies. European Journal of Operational Research 123(1), 105–124 (2000)

KPIs from Web Agents for Policies' Impact Analysis and Products' Brand Assessment*

Antonio Candiello and Agostino Cortesi

Dipartimento di Scienze Ambientali, Statistica ed Informatica,
Università Ca' Foscari, Venice, Italy
{candiello,cortesi}@unive.it

Abstract. Both Enterprises and Public Authorities (PAs) need a continuous and updated flux of reliable data in order to select the better choices. Classical Business Intelligence tools fed with internal data could be augmented with new tools able to extract KPIs of interest from the Raw Web made of unstructured HTML pages and from the Deep Web made of online DBs. The continuous growth of data made available on the web increases intrinsically, year by year, the reliability of this approach. A "Web Intelligence" agents-based framework supporting the evaluation of the effective impact of projects and initiatives has been designed and is currently being developed and tested; the system combines up-to-date indicators obtained via a systematic and high frequency staggered data scraping with lower-rate extraction of data from online data sources. The corresponding model for the management, monitoring and assessment of projects implemented by Enterprises and PAs is also presented.

Keywords: Public Authorities, Business Intelligence, Quality Management.

1 Introduction

It is yearly registered a continuous increase of the quantity of digital data produced, that is estimated by IDC [1] at 1.8 Zettabyte for the 2011. Even if only a fraction of this data is made available on the web (and even less data in text, HTML or XML form), the availability of up-to-date web data is improving the quality of indicators that can be extracted from this wide amount of structured and unstructured information. Properly designed web information agents could help both Enterprises and Public Authorities to gather updated and geographically referentiated data related to the impact of their projects and initiatives.

The popular Web 2.0 model is also making available increasing "User Generated Content" (UGC) data in forums, chats, blogs, wikis and Social Networks (SNs). UGC can be a precious source of information for Enterprises wishing to evaluate their brands' esteem in general. Enterprises can also search in UGC the specific attribute that consumers associate to the products, like reliability, cheapness, quality, luxury, etc. Marketing 2.0 [2] is the advanced form of marketing

* Work partially supported by Regione Veneto – Direzione Sistemi Informativi.

N. Chaki and A. Cortesi (Eds.): CISIM 2011, CCIS 245, pp. 192–201, 2011.

that makes use of Web 2.0 tools in order to facilitate and improve the relation between customers and Enterprises. Both *active* supporting initiatives (to suggest Enterprises' products and brands in the appropriate contexts) as well as *passive* analysis activities and tools (searching for relevant product/brand feedbacks) are conducted by the more advanced and customer-oriented Enterprises.

On the other hand, a relevant priority for Public Authorities aiming at promoting innovation and supporting social and economic developments, is the combination of eGovernment policies and domain-related support policies (frequently ICT). The effects of correlated innovation projects should be measurable via focused analysis of specific statistical indicators [3,4]. These can be either *direct* eGovernment or domain-related indicators (as is the case, for instance, of portal/online services access, wide band internet coverage of population, number of patents, etc) or *indirect* (impact) socio-economical indicators (as, for instance, average income, local GDP growth, availability of qualified engineers and so on).

Both consumer feedback trails related to products/brands brought by Enterprises (see [5]) and impact of innovation policies implemented by PAs [6] should be evaluated in an adequately wide time-span, month by month, quarter by quarter. The related feedback/impact measurements should follow this pattern, by registering at definite temporal intervals the state of the monitored indicators. A comprehensive strategy for medium-term branding assessment and/or impact measurement could help to outline the searched correlations between the *effects* (brand appreciation for Enterprises, social and economic improvements for PAs) and the *causes* (new products, better marketing, for Enterprises, or specific innovation policies/projects for PAs [7]). We searched, when possible, to assign to indicators a geographical dimension, where the smallest units considered are the municipalities.

The approach requires a thoughtful choice of specific statistical *Key Performance Indicators* (KPIs). These can be obtained via three main sources:

1. by harvesting the "Raw Web" with webbots, data scrapers, crawlers, spiders, searching for specific trails left by the consumers and the citizens;
2. by retrieving (via the so-called "Deep Web" of the online databases) the indicators available from official Institutions like Eurostat, National Governments, local Chambers of Commerce, or from companies specialized in market analysis, ratings and statistical analysis like Nielsen, IDC or Gartner;
3. by addressing consumers or citizens specific surveys.

As already said, the Raw Web webbot-based strategy is gaining relevance, due to the rapid growth of the quantity and the quality of the information that is available on the web. We expect an increasing trustworthiness of the measurements that will overcome the critical point of this strategy, i.e. the reliability of web-derived indicators data. The *webbots*, said also *data scrapers* when extracting data from a single web source, or *crawlers*, *spiders* when moving between links finding for the needed information, need to be updated when the source web sites change their content or even their appearance (see [8] for an introduction on the

theme; see also [9] for a performance comparison between crawler classes). The research focused on indicators connected to social community-related sources, like blogs, forums and social networks (the Web 2.0) that have the advantage to be continuously up-to-date – and that perhaps will never be closed to the web agents as could be for some internet content [10].

The second approach offers the highest quality data and represents the basic reference thanks to its officiality. The main problem with this approach is that the data officially monitored often do not cover the innovation context to enough detail for PAs, and domain-wide market statistics do not always match Enterprises' needs. Also, these indicators are not always provided in finer territorial detail than the regional one.

Focused surveys offer the opportunity to select specific product or innovation questions, generally not inspected by general statistics. Survey campaigns are generally limited in time (campaigns rarely last for more than a year), in space (provincial or municipal level) or in the context (market sector or consumers segment). Online tools for survey submissions could help in maintaining some efficiency trade-off in the use of this information channel.

The goal of our research is to explicit a comprehensive model for (a) the management of Enterprises' marketing initiatives or PA innovation projects, (b) their systematic monitoring and related impact analysis measurement and to support the model with (c) an integrated *Web Agents Intelligence* information system capable of registering and monitoring such policies, monitor the relative initiatives/projects against their goals, systematically evaluating their impact and finally reviewing the policies themselves on the basis of the resulting analysis.

We conducted the PA-related research in collaboration with our local government partner, Regione Veneto (in northern-east Italy) by extending the QoS monitoring strategy [11] to include impact analysis. The Enterprises-related research was conducted by analysing business-related KPIs on a regional scale and focusing on the products of the local district of sportswear.

The information system developed within this applied research provides public administrators capability to continuously improve their services via an objective evaluation of the resulting impact, keeping their citizens better informed and up-to-date regarding goals set in advance for the policies and the success rate of the local government funded innovation initiatives carried out for the public benefit. Enterprises, on the other hand, are able to evaluate the effective appreciation of their products/brands by the consumers; specific attributes that customers associate to the products (like *performance*, or *durability*, or *comfort*) can also be inspected.

The paper is organized as follows. In Section 2 a model for Enterprises' product initiatives and PAs' services innovation policy management is introduced, and in Section 3 the supporting Web Intelligence framework, its modules and its interaction with the model are presented. Then, in Section 4, some conclusions are drawn.

2 A Model for the Management of Policies for Enterprises and PAs

A comprehensive [12] model for monitoring innovation projects, validating the related policies and evaluating the effective direct and indirect impact on the areas affected could improve the success rate of the Public Authorities innovation initiatives, specifically for ICT [13,14]. Policies and related projects, which we consider mainly relating to eGovernment [4] as well as related to the wider context of ICT infrastructures [15] should be assessed also by the citizens themselves [16]. We adapted the classic Deming plan-do-check-act (PDCA) cycle to the Public Authorities requirements for innovation policies management [17]. This same model can be applied to internal quality management processes of Enterprises for the development of product- and brand-related initiatives. In this last case the impact analysis is substituted by the search for brand-appreciation evidences.

Each policy management PDCA phase is identified by a organizational process and is supported by specific subsystems of the Web Intelligence agents-based framework. The complete model is shown in Fig. 1.

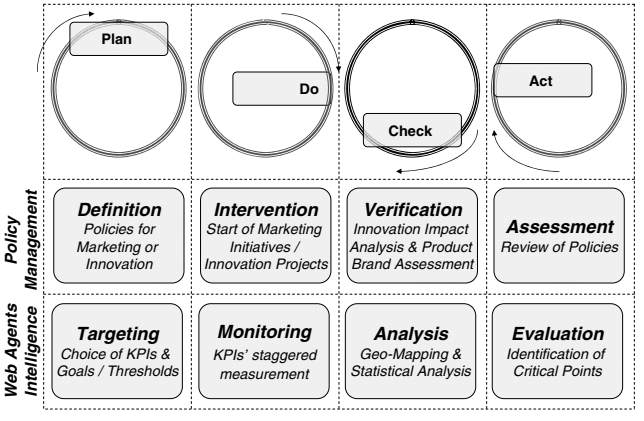


Fig. 1. The comprehensive model for Policy Management, supported by the tools of the Web Intelligence agents-based framework

The goals of this model are:

- finding an *objective validation* for the effectiveness of PAs / Enterprises policies,
- *qualifying and quantifying* the effectiveness through appropriate impact statistical indicators,
- *gathering the relevant indicators* via automatic (webbots/scrapers) and semi-automatic (extractors/wrappers), completing the data when needed with focused survey campaigns,

– *representing and mapping the indicators* showing the explicit relation with the affecting innovation projects and the areas involved.

The indicators are classified in three categories: (a) *direct* innovation indicators, mainly the ICT indicators enabling the Information Society, connected to the technology – examples of indicators in this category are population reached by the internet wide band, eGovernment services offered, eGovernment use of services, ratio of PCs/users, ICT education knowledgeability; (b) *indirect* socio-economical indicators related to the resultant impact of the innovation over the local communities, participation, sociality, economy and culture; (c) *specific* product- and brand-related indicators, able to report the evidence of consumers positive or negative opinions regarding specific products or lines of products.

Direct, innovation indicators are easier to manage, as they are strictly related to the innovation projects. For instance, internet wide band penetration ratio, or renewable energy end-user production could be directly related to infrastructure funding and incentives set up by the local governments; similarly, the growth of the number of accesses to eGovernment Portals depends on quality of the offered online services. These indicators require however the setup of specific measurement processes, as innovation evolution is not systematically monitored by the National or Regional Institutions dedicated to statistical data analysis.

Indirect, socio-economical indicators are more easily found in the periodic data reporting produced by National and International statistical or economical Institutions, but these are *second-level correlated* to innovation projects, i.e. there is the need to estimate and then evaluate their effective correlation with the intermediate innovation indicators which can then be put in direct correlation to the monitored projects. For instance, an investment for wide-area internet wide band could in the long-term sustain new business developments, widen social communities opportunities and in general improve the quality of life. The management of indirect socio-economical indicators requires however carefully staggered gathering of statistical data, and the estimation of the correlations between them and the "raw" innovation indicators. In the current phase of the research, we concentrated our efforts in extracting direct ICT innovation and eGovernment indicators and in selecting simple cases for the socio-economical impact indicators without modeling the effective correlations between the two classes of indicators – we are leaving this task for subsequent research phases.

Specific, product- and brand-related indicators are to be actively searched via appropriate techniques for the elaboration of the text found regarding consumers' opinion trails found on the general web. These indicators can be extracted directly (1) from the number of results of simple queries via online search engines as well as (2) from complex lexical semantic analysis of the pages where the brands and products are cited [5]. The first approach was used for the construction of KPIs, leaving the second approach at an experimental stage for the construction of maps similar to the Brand Association Map from Nielsen (see [18,19]), in an effort to deploy effective Decision Support Systems (DSS) for the Enterprises [20] built on the Web of Data.

3 The Web Agents Intelligence Technical Framework

We developed the Web Agents Intelligence framework around the following elements (see Fig. 2):

- the *Policy Manager*, a GUI panel for the management (and review) of policies, projects, the selection of impact indicators and the setting of targets,
- the *Events Scheduler* for the reliable planning of monitoring events with a minimal temporal unit of a day,
- the *Agents Manager* for the management of webbots/data scrapers, wrappers/adapters and for the launch of email/online survey campaigns,
- the *Map Viewer*, a geo-referentiated visualization engine built around the SpagoBI open source platform.

Let us discuss them in detail.

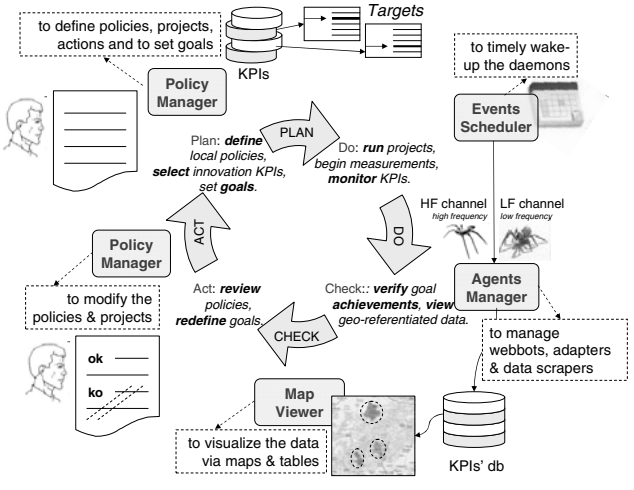


Fig. 2. The agent modules and their relationship with the PDCA cycle

Policies are defined by Enterprises or Public Authorities, and then developed into explicit projects/initiatives. Such projects have definite dates of deployment, they can frequently be geographically localized and they have associated milestones. A *policy manager* has been designed to provide the system all the available data regarding policies, projects/initiatives. The policies, their associated KPIs and targets are defined *ex ante* in the PLAN phase; then, in the ACT phase, the policies are reviewed on the basis of the *ex post* analysis of the associated KPIs.

The *Events Scheduler* manages then (in the DO phase) the execution of the daemons that scan the web, access the online repositories or launch the survey

campaigns via specific web portals. The scheduler uses as the smallest temporal unit the day, and offers the possibility to activate programmed monitoring for timed events ranging from daily intervals, passing through weekly, monthly, quarterly and other sized periodic intervals. The reliability of the scheduler is a central requirement for the framework, as the analysis phases need complete data sets in order to produce trustworthy and significative statistical elaborations. The scheduler fulfills a relevant requirement of our framework: the capability to monitor the trend of the KPIs inspected.

The *Agents Manager* contains the business logic to extract indicators from the web sources. We developed classes of webbots/data scrapers addressing unstructured Raw Web web pages and classes of wrappers/adapters addressing structured Deep Web online databases.

Webbots and data scrapers search for raw information found on the general web. Simple keyword-based searches via major search engines like Google and Yahoo! were experimented. Innovation and ICT-related words were combined to municipality names in order to create geo-referentiated maps; with the same technique products and brands were placed on geographical maps. General Web 2.0 production/consume indicators were also collected by querying Youtube, Flickr and other popular digital repositories. eGovernment indicators were also inspected (indirectly) via Yahoo! Sites, that counts the referring links to the selected web sites. We are currently experimenting the direct extraction of information from blogs- and forum-related web sites. These indicators are mainly used to estimate general innovation and ICT parameters and the brands' popularity via the analysis of the content produced.

The indicators extracted via this channel – webbots and data scrapers for web data harvesting – have a weaker reliability, due to the nature of raw information they are derived from. Also, the agents require a continuous maintenance activity because of the frequent change of the layout of the inspected web sites and/or of their respective APIs used to access the data. On the other way, there is the advantage that this data can be tuned both with respect to the time (as even daily updates can be considered), constituting a *high frequency* channel, and in space (a municipality level is reachable).

As online data source, Eurostat offers the widest option choices for the gathering of structured data, from CSV to XML to several web interfaces and API to manage data: as noted in [21], Eurostat offers more than 4,000 datasets containing roughly 350 million data values/items. Other data sources are more limited in the choices they offer, sometimes even limited to a fixed-format CSV or PDF file. We developed specific extractors for the data reported by the National Statistical Institute collected along with other socio-economical data by the regional Statistical Office managed by our local government partner. Specifically, income, number of inhabitants/families, age distribution data is available year by year at the required territorial resolution of a single municipality. We developed also specific wrappers for common online business directories.

The indicators extracted via this second channel – extractors for official / institutional data retrieval – mainly of socio-economical nature, have the highest

reliability. They also offer the advantage of being completely available at the municipality level. This is a *low frequency* channel, as the updates are typically released with an yearly periodicity. The scheduler has to be instructed to wake up the relative daemons at the right date when the updated data are available.

Webbots/scrapers for Raw Web data and webbots/wrappers for Deep Web data constitute the core of the Agents Manager. We are currently working on the (Semantic Web) extension of the Agents Manager to access the so-called Web of Data [21] or Linked Data [22] in order to extract from these sources higher quality KPIs.

As a third, complementary, input channel, we integrated in the agents manager the eGif engine [23]. The indicators obtained via this channel – mainly citizen feedback regarding eGovernment services and impact of innovation policies – are costly for the effort required in the survey campaign management, but can be useful to complete the statistical analysis for specific themes/areas. Other common online survey solutions can be integrated, as alternatives to eGif, for consumer surveys addressing the needs of Enterprises.

The impact analysis of gathered data is then (in the CHECK phase) managed by the *Map Viewer* with the support of the open source visualization engine SpagoBI engine [24]. The Map Viewer exposes the indicators data over the local territories. It allows to localize the impact of the Enterprises' and Public Authorities' policies. In order to be represented on maps, the KPIs have to be evaluated against a geographical dimension. As the smallest geographical units, we selected the municipalities (corresponding to LAU2 in the Eurostat regional classification).

We are currently experimenting extensions of the SpagoBI platform in order to be able to use also multi-dimensional geo-referentiated data patterns, as the *travelling time distance grid* research case that we tested for mountain-area municipalities and for more abstract patterns appropriate for attributes-brands correlation maps.

4 Conclusions

In this paper a comprehensive policy management model and its supporting Web Agents Intelligence framework has been presented; the model, drawn from quality management methodologies, offers the capability to measure the local impact of innovation policies brought forward by Public Authorities and to reveal the customer opinions regarding Enterprises' specific products and brands. The policy management model and the coupled Web Intelligence framework should help both in reviewing and improving their projects/initiatives by inspecting the resulting impact in detail.

The main features of the model are: (a) the qualification of the policies/projects and the definition of innovation targets, (b) a systematic and staggered measurement of the relevant innovation, economic, social and marketing indicators at the needed scale, (c) a detailed, geo-referentiated analysis of the evolution patterns of the indicators and the relation of products/brands with

specific attributes, (d) the re-assessment of policies and related projects/initiatives against the results obtained.

A first set of core indicators relevant for Public Authorities (ranging from socio-economical data like population, income, business presence, to ICT-related data related to education, eGovernment usage, user-produced content, wide band infrastructures), was extracted from official Institutions and raw web sources. A complete data set of the indicators has been created for all of the regional municipalities; the results, reported on the SpagoBI-powered maps, are currently discussed with regional government staff and the relations with the local ICT innovation initiatives were analyzed.

References

1. Gantz, J., Reinsel, D.: Extracting value from Chaos. Technical report, IDC (2011)
2. Consoli, D., Musso, F.: Marketing 2.0: A new marketing strategy. Journal of International Scientific Publications: Economy & Business 4(2), 315–325 (2010)
3. Janssen, M.: Measuring and Benchmarking the Back-End of E-Government: A Participative Self-Assessment Approach. In: Wimmer, M.A., Chappelet, J.-L., Janssen, M., Scholl, H.J. (eds.) EGOV 2010. LNCS, vol. 6228, pp. 156–167. Springer, Heidelberg (2010)
4. Neuroni, A., Rascon, A., Spichiger, A., Riedl, R.: Assessing and evaluating value and cost effectiveness of e-government initiatives: Focusing the step of the financial evaluation. In: Chun, S.A., Sandoval, R., Philpot, A. (eds.) dg.o 2010. ACM Digital Library, Digital Government Society (2010)
5. Aggarwal, P., Vaidyanathan, R., Venkatesh, A.: Using lexical semantic analysis to derive online brand positions: An application to retail marketing research. Journal of Retailing 85(2), 145–158 (2009)
6. Bernroider, E.W.N., Koch, S., Stix, V.: Elements of Comprehensive Assessments of IT Infrastructure Projects in the Austrian Ministry of Finance. In: Andersen, K.N., Francesconi, E., Grönlund, Å., van Engers, T.M. (eds.) EGOVIS 2010. LNCS, vol. 6267, pp. 84–91. Springer, Heidelberg (2010)
7. Misuraca, G., Ferro, E., Caroleo, B.: Assessing Emerging ICT-Enabled Governance Models in European Cities: Results from a Mapping Survey. In: Wimmer, M.A., Chappelet, J.-L., Janssen, M., Scholl, H.J. (eds.) EGOV 2010. LNCS, vol. 6228, pp. 168–179. Springer, Heidelberg (2010)
8. Schrenk, M.: Webbots, Spiders, and Screen Scrapers: A Guide to Developing Internet Agents with PHP/CURL. No Starch Press (2007)
9. Batsakis, S., Petrakis, E., Milios, E.: Improving the performance of focused web crawlers. Data and Knowledge Engineering 68(10), 1001–1013 (2009)
10. Jennings, F., Yates, J.: Scrapping over data: are the data scrapers days numbered? Journal of Intellectual Property Law & Practice 4(2), 120–129 (2009)
11. Candiello, A., Albarelli, A., Cortesi, A.: Three-layered qos for egovernment web services. In: Chun, S.A., Sandoval, R., Philpot, A. (eds.) dg.o 2010. ACM Digital Library, Digital Government Society (2010)
12. Ojo, A., Janowski, T.: A whole-of-government approach to information technology strategy management. In: Chun, S.A., Sandoval, R., Philpot, A. (eds.) dg.o 2010. ACM Digital Library, Digital Government Society (2010)

13. De', R., Sarkar, S.: Rituals in E-Government Implementation: An Analysis of Failure. In: Wimmer, M.A., Chappelet, J.-L., Janssen, M., Scholl, H.J. (eds.) EGOV 2010. LNCS, vol. 6228, pp. 226–237. Springer, Heidelberg (2010)
14. Janssen, M., Klievink, B.: Ict-project failure in public administration: The need to include risk management in enterprise architectures. In: Chun, S.A., Sandoval, R., Philpot, A. (eds.) dg.o 2010. ACM Digital Library, Digital Government Society (2010)
15. Lampathaki, F., Charalabidis, Y., Passas, S., Osimo, D., Bicking, M., Wimmer, M.A., Askounis, D.: Defining a Taxonomy for Research Areas on ICT for Governance and Policy Modelling. In: Wimmer, M.A., Chappelet, J.-L., Janssen, M., Scholl, H.J. (eds.) EGOV 2010. LNCS, vol. 6228, pp. 61–72. Springer, Heidelberg (2010)
16. Tomkins, A.J., PytlikZilllig, L.M., Herian, M.N., Abdel-Monem, T., Hamm, J.A.: Public input for municipal policymaking: Engagement methods and their impact on trust and confidence. In: Chun, S.A., Sandoval, R., Philpot, A. (eds.) dg.o 2010. ACM Digital Library, Digital Government Society (2010)
17. Candiello, A., Cortesi, A.: KPI-Supported PDCA Model for Innovation Policy Management in Local government. In: Janssen, M., Scholl, H.J., Wimmer, M.A., Tan, Y.-H. (eds.) EGOV 2011. LNCS, vol. 6846, pp. 320–331. Springer, Heidelberg (2011)
18. Akiva, N., Greitzer, E., Krichman, Y., Schler, J.: Mining and Visualizing Online Web Content Using BAM: Brand Association Map. In: Proceedings of the Second International Conference on Weblogs and Social Media, pp. 170–171. Association for the Advancement of Artificial Intelligence (2008)
19. Till, B.D., Baack, D., Waterman, B.: Strategic brand association maps: developing brand insight. Journal of Product & Brand Management 20(2), 92–100 (2009)
20. Dai, Y., Kakkonen, T., Sutinen, E.: MinEDec: a Decision-Support Model That Combines Text-Mining Technologies with Two Competitive Intelligence Analysis Methods. International Journal of Computer Information Systems and Industrial Management Applications 3, 165–173 (2011)
21. Hausenblas, M., Halb, W., Raimond, Y., Feigenbaum, L., Ayers, D.: SCOVO: Using Statistics on the Web of Data. In: Aroyo, L., Traverso, P., Ciravegna, F., Cimiano, P., Heath, T., Hyvönen, E., Mizoguchi, R., Oren, E., Sabou, M., Simperl, E.P.B. (eds.) ESWC 2009. LNCS, vol. 5554, pp. 708–722. Springer, Heidelberg (2009)
22. Bizer, C.: The Emerging Web of Linked Data. IEEE Intelligent Systems 24(5), 87–92 (2009)
23. Candiello, A., Albarelli, A., Cortesi, A.: An ontology-based inquiry framework. In: Gangemi, A., Keizer, J., Presutti, V., Stoermer, H. (eds.) DEGAS 2009. CEUR Workshop Proceedings, vol. 426 (2008)
24. Golfarelli, M.: Open Source BI Platforms: A Functional and Architectural Comparison. In: Pedersen, T.B., Mohania, M.K., Tjoa, A.M. (eds.) DaWaK 2009. LNCS, vol. 5691, pp. 287–297. Springer, Heidelberg (2009)

Stochastic Local Search Approaches in Solving the Nurse Scheduling Problem

Sudip Kundu[1] and Sriyankar Acharyya[2]

[1] Department of Computer Science and Engineering, B.P.Poddar Institute of Management and Technology, Kolkata, India
sudip.wbut@gmail.com
[2] Department of Computer Science and Engineering, West Bengal University of Technology, Kolkata, India
srikalpa8@yahoo.co.in

Abstract. Nurse Scheduling Problem (NSP) is very complex in nature and involves a large number of constraints to satisfy. In hospitals, nursing homes, and other health-care organizations, where the daily work is divided into shifts, the Nurse Scheduling Problem arises there to assign duties to all the nurses over a short-term planning period satisfying constraints as much as possible. It can be viewed as a Combinatorial Optimization Problem. The constraints are related to labor contract rules, preferences submitted by nurses, concerns of the employers and other real life situations. In this paper, apart from applying Simulated Annealing and Genetic Algorithms, we converted the cyclic NSP to Satisfiability Problem (SAT) and applied local search SAT techniques to solve it. In comparative assessment of the methods we show that USAT incorporated with a tabu list has outperformed other SAT methods in some instances. Otherwise, the SAT methods are almost equivalent in performance.

Keywords: Constraint Satisfaction, Meta-heuristics, Satisfiability, Tabu List, Simulated Annealing, Genetic Algorithms.

1 Introduction

The Nurse Scheduling Problem (NSP) is a special kind of Staff Scheduling Problem. It involves generating individual schedules for nurses consisting of working shifts and off days over a planning period. Duties of nurses are divided into shifts. In a schedule (roster), on each day, a nurse is assigned either a working shift or no duty. The assignment should primarily comply with nursing requirements in a ward. The other constraints are related to the rules laid down by the administration and the labor contract clauses.

Scheduling of nurses into shifts involves considerable time and resources, and it is often challenging to create schedules that satisfy all the requirements related to fluctuating service demand on different days and shifts. Nurses are assigned shifts according to the requirement, skill levels and their preferences. For example, they may submit their preferences for rest days [22], shift types, and work patterns. It is also important to evenly balance the workload among nurses satisfying their preferences.

N. Chaki and A. Cortesi (Eds.): CISIM 2011, CCIS 245, pp. 202–211, 2011.

There are three commonly used methods for solving NSPs, namely, Mathematical Programming (MP), Heuristics and Artificial Intelligence (AI). MP based algorithms of Warner [25] proposed a multi-choice programming model where objective function maximizes nurses' preferences. Millar and Kiragu [20] used a network model for cyclic and non-cyclic nurse scheduling. As there are conflicting constraints, multi-objective programming was applied by Arthur and Ravindran[6]. With a differing scheduling policy Musa and Saxena [21] proposed a single phase GP algorithm. Azaiez and AI Sharif [7] developed a GP model that considers different hospital objectives and nurses' preferences. Also there are heuristic methods [23] which can produce satisfactory results in reasonably short times. Dowsland [8], Dowsland and Thompson [9] used tabu search and its variant. Aickelin and Dowsland[5] developed a genetic algorithm approach that handles the conflict between objectives and constraints. Abdennadher and Schlenker [1] developed an interactive constraint based scheduler.

Actually, there are several approaches to the Nurse Scheduling Problem based on the framework of Constraint Satisfaction Problem (CSP) [13]. Earlier, we implemented Simulated Annealing and Genetic Algorithm to solve randomly generated problem instances [14]. In most of the cases, Simulated Annealing easily outperformed Genetic Algorithm. Afterwards, the NSP was converted to Satisfiability Problem (SAT) and we saw how GSAT and WalkSAT [15] [16] outperformed SA. In this paper we have applied other SAT techniques like USAT, Novelty, RNovelty, Novelty+, RNovelty+ to solve it.

The sections are arranged as follows: Section-2 describes the NSP, Constraints, and SAT formulation. Section-3 presents two heuristic methods (SA and GA) and some Local Search SAT techniques which we have used to solve this problem. In Section-4 we present our experimental results and the comparative assessment of different methods. Finally, Section-5 concludes the paper and suggests some further works.

2 Problem Description and Formulation

We have focused on the 3-shift system [12] where each day consists of three working shifts: a morning-shift, an evening-shift, and a night-shift. On any day, each nurse is assigned only one working shift, or else, given an off-day.

We divide the constraints into two categories: Hard and Soft. A roster is feasible if all the hard constraints are satisfied. In addition, the satisfaction of soft constraints determines the quality of a feasible roster.

Now, we present a mathematical model for the NSP. Here n and m represent the number of nurses and the number of days respectively. Index of nurse, day number, and shift value is given by i, j, and k respectively. So, $i \in \{1, \ldots,n\}$, $j \in \{1, \ldots,m\}$, and $k \in \{1, \ldots,4\}$. Shift value 4 represents off-day. Parameter *min_on_shift[k]* represents minimum number of nurses required for shift type k. Similarly, *max_on_shift[k]* represents maximum number of nurses allowed for shift type k. *min_shift[k]*gives the value for minimum number of k-type shift allowed for each nurse. Similarly, *max_shift[k]*gives the value for maximum allowed number of k-type shift for each

nurse. *shift_for_nurse[k][i]* stores total number of *k*-type shifts assigned to nurse *i*. And *nurses_on_day[k][j]*counts total number of nurses assigned *k*-type shift on day *j*. In our problem the Decision Variable, X (*i,j,k*) = 1 if nurse *i* is assigned to shift type *k* on day *j*, or 0 otherwise.

2.1 Hard Constraints

We have considered three different types of Hard Constraints:

HC1. On any day, one and only one shift is assigned to each nurse. The off-day is considered as shift 4. This means a nurse should either take a day off or be assigned to an available shift on each day.

For nurse no i, day no j, this is expressed as
$\sum X(i,j,k) = 1$, summation over k, where $1 \leq k \leq 4$;
The violation of this constraint contributes to the cost component cost_HC1.

HC2. It imposes the restriction on the number of assigned nurses in each working shift per day. i.e., in each shift, the number of nurses must be within the range of a maximum and a minimum values.

For day no j, shift no k it can be expressed as:
$\sum X(i,j,k) \leq$ max_on_shift[k], summation over i, where $1 \leq i \leq n$; and
$\sum X(i,j,k) \geq$ min_on_shift[k], summation over i, where $1 \leq i \leq n$;
The violation of this constraint contributes to the cost component cost_HC2.

HC3. The roster should be cyclic, i.e., the generated roster can be repeated indefinitely if no new constraint is introduced. In addition, it must avoid some prohibited working patterns which are predefined.

For i-th nurse, if we start checking from j-th day, the constraints corresponding to prohibited working patterns can be expressed as follows:

"Three consecutive night-shifts" : $X(i,j,3) + X(i,j+1,3) + X(i,j+2,3) \leq 2$
"Morning-shift after night-shift" : $X(i,j,3) + X(i,j+1,1) \leq 1$
"Evening-shift after night-shift" : $X(i,j,3) + X(i,j+1,2) \leq 1$
"Morning -shift after evening-shift" : $X(i,j,2) + X(i,j+1,1) \leq 1$
The violation of this constraint contributes to the cost component cost_HC3.

2.2 Soft Constraints

These are the secondary constraints. The quality of the schedule will depend on the degree to which these constraints are satisfied. In our approach, we have considered the following constraint type as soft constraints:

SC1. For each nurse, number of different shifts (morning, evening, night, or day-off) assigned should be within the range of a maximum and a minimum value.

For each *i* and each *k*
$\sum X(i,j,k) \leq max\_shift[k]$; and

$\sum X(i,j,k) \geq min\_shift[k]$, summation over j, where, $1 \leq j \leq m$;
The violation of this constraint contributes to the cost component *cost_SC1*.

2.3 Objective Function

The objective function is the total cost which is given by
Cost = cost_HC1 + cost_HC2 + cost_HC3 + cost_SC1.

2.4 SAT Formulation

In our problem a variable is represented as X(i,j,k) where 'i', 'j', 'k' represent nurse-id, day-no, and shift-value respectively. The Boolean variable X(i,j,k) is set TRUE if and only if i^{th} nurse on j^{th} day is assigned the shift k. Otherwise X(i,j,k) is set to FALSE. If we want to make a schedule of 'n' number of nurses for 'm' days, then we need to find out an appropriate assignment for all X(i,j,k) where $1 \leq i \leq n$, $1 \leq j \leq m$, and $1 \leq k \leq 4$.

Now, let us see how the clauses are generated for different types of constraints.

Clauses for constraint HC1. For i^{th} nurse, on j^{th} day, we need two set of clauses; one will ensure that at least one of the four shifts is set TRUE. This can be represented as follows:

$X(i,j,1) \cup X(i,j,2) \cup X(i,j,3) \cup X(i,j,4)$

Another set will ensure that no two shifts, k_1 and k_2, will be assigned to nurse no i on the same day j. This can be mathematically expressed as:
$\sim X(i,j,k_1) \cup \sim X(i,j,k_2)$ where k_1 and k_2 are combination of distinct shift values.

Clauses for constraint HC2. To generate clauses which will ensure that minimum l number of nurses are assigned shift-k on day j, we have to assign shift-k to at least one of every combination of $(n\text{-}l+1)$ nurses. This can be written as:

$X(i_1,j,k) \cup X(i_2,j,k) \cup X(i_3,j,k) \cup \ldots\ldots\ldots\ldots\ldots\ldots \cup X(i_{n-l+1}, j, k)$

For the upper bound, we have to generate clauses which will ensure that at most 'u' nurses are assigned shift-k on day no j. Now, if we choose any combination of $(u+1)$ nurses, then, for at least one value of i, we must have $X(i,j,k)$ = FALSE, as we cannot assign more than u nurses to shift-k at the same time. For shift-k, this can be mathematically expressed by the following logical proposition:

$$\sim X(i_1,j,k) \cup \sim X(i_2,j,k) \cup \sim X(i_3,j,k) \cup \ldots\ldots\ldots\ldots\ldots \cup \sim X(i_{u+1},j,k)$$

Clauses for constraint HC3. If the sequence $< k_1, k_2, k_3 >$ is prohibited, then for any nurse i, we cannot assign shift k_1 on day no j, followed by k_2 on day $(j+1)$, and k_3 on day (j+2), for any value of j, as this will violate the constraint. So, at least one value among $X(i,j,k_1)$, $X(i,j+1,k_2)$,and $X(i,j+2,k_3)$ must be FALSE. This can be expressed by the following logical proposition:

$\sim X(i,j,k_1) \cup \sim X(i,j+1,k_2) \cup \sim X(i,j+2,k_3)$

Clauses for constraint SC1. First we'll ensure that at least l number of k shift is assigned to nurse i. Now, if we choose any (m-l+1) shift values assigned to nurse i, at least one of them must be k. For nurse no i, this is represented as:

$X(i,j_1,k)$ U $X(i,j_2,k)$ U …… U $X(i,j_{m-l+1},k)$

Where all j's are distinct values representing different days.

Now, we'll consider the upper limit. If shift-k can be assigned maximum u times to nurse i, then if we choose any (u+1) shift values assigned to nurse i, all of them cannot be equal to 'k' at the same time. So, at least for one day, the assigned value of $X(i,j,k)$ will be FALSE. This can be written as:

$\sim X(i,j_1,k)$ U $\sim X(i, j_2,k)$ U ……U $\sim X(i, j_{u+1},k)$

Where $\{j_1,j_2,\ldots, j_{u+1}\}$ is combination of (u+1) different days.

Now, the NSP is fully converted to Satisfiability and we are left with a set of Boolean variables $X(i,j,k)$ and a set of clauses.

3 Solution Methods

The problem is NP-hard and therefore we are applying the stochastic local search methods to solve the problem instances. Here we will discuss different approaches and their implementation details.

3.1 Simulated Annealing (SA)

Simulated Annealing [14] starts with an initial trial solution which is obtained by randomly assigning each variable a value. When each X(i,j,k) is given truth values randomly, we can expect only a few of the constraints to be satisfied. The initial cost is calculated using the cost function. Then we choose a variable randomly from the solution and assign new value to that variable. This is how we move to a neighborhood solution. After the move, we calculate the cost again.

If the neighborhood solution is better, then it is always accepted. Otherwise, we generate a random number r between 0 and 1. Now we'll accept this inferior solution if and only if $r < \exp(-\Delta E/T)$, where ΔE is the increase in cost and T is the current temperature. This process is repeated.

To implement SA, we must tune the values of the parameters to their best. Otherwise, inferior solutions will be generated more frequently. The most important issue is the initialization of the temperature (T), and the rate at which it should decrease. We have chosen initial temperature as high as 2000. The process continues, and finally, the algorithm outputs the feasible solution of lowest cost.

3.2 Genetic Algorithm (GA)

Canonical GAs [19] were not intended for function optimization, but slightly modified versions proved successful [4] [14]. In our implementation GA starts with WP

number of randomly generated trial solutions called initial population, where WP was chosen to be around 10. Each complete trial solution is represented by a chromosome where the elements are the values of each X(i,j,k). First element of any chromosome represents value of X(1,1,1); second element represents value of X(1,1,2), and so on. Cross-over and Mutation operators are used to generate new chromosomes [4] which will form a new population.

We have implemented GA in two ways. In both cases, the best YG numbers of chromosomes get selected automatically for inclusion in the population of the next generation. R pairs of chromosomes take part in crossovers, generating 2*R new chromosomes (offspring) for the next generation. In GA1, solutions which are selected to form new chromosome are chosen according to their fitness values, wherein the more suitable they are, the more opportunities they have to reproduce. Mutation was done on the resultant chromosomes with a very small mutation probability.

In GA2, for each pair of chromosomes taking part in a crossover operation, one is chosen randomly from among the best YG, and the other is chosen randomly from the entire population. The remaining (WP - YG - 2*R) chromosomes of the next generation are obtained by mutation. The entire procedure is iterated over nGen generations typically having a high value around 10000. Among the chromosomes generated, the one that had the highest rating among all feasible solutions was finally outputted.

3.3 Algorithm GSAT(L)

GSAT is a greedy local search procedure. It assigns truth values to variables in an effort to satisfy all the clauses in a given set [24]. It finds out a variable *v* such that its truth-value when flipped causes the maximum net increase in the number of satisfied clauses. Then it flips the truth-value of *v* and again looks for a new variable with the same property. When there are two or more such variables, ties are resolved arbitrarily. This is repeated until a satisfying assignment for all X(i,j,k) is found. If no satisfying assignment is found within *maxflips* number of iterations, the procedure is restarted with a new initial truth assignment (*try*). This cycle is repeated for *maxtries* number of tries.

We have incorporated a tabu search strategy to improve the performance of GSAT. It prevents the same variable *v* from flipping [18] repeatedly. Tabu List is implemented as a FIFO queue and is initially kept empty. The most recently flipped variable is inserted into the list. Selection of next variable is done randomly [10] from among those variables not present in the list and causes the largest increase in the number of satisfied clauses. In this way, we prevent some variables from being flipped for L number of iterations, where L is the length of the tabu list. This is one effective way of avoiding local minima.

3.4 Algorithm WalkSAT(L,p)

The WalkSAT algorithm [16] with a tabu list of length L and random walk probability p, also starts with a random truth assignment. But the method of selecting variable

differs from GSAT. Here, we always select the variable from a randomly chosen unsatisfied clause. If variables in that clause can be flipped without unsatisfying other clauses, one of them is randomly selected. If no such variable exists, then with probability 'p' a variable is randomly chosen from the clause, otherwise, a variable is picked that minimizes the number of clauses that are currently satisfied but would become unsatisfied when the variable is flipped.

3.5 Algorithm USAT(L)

This algorithm differs from GSAT(L) [2] in the manner of selection of the variable to be flipped. Here an additional restriction is imposed on the selection of the variable, namely, we are required to select the variable from an unsatisfied clause. Given $L \geq 0$, the variables selected for flipping by GSAT(L) and USAT(L) are the same when *maxgain*, the maximum gain of a variable, exceeds 0. But when maxgain becomes $\leq$ 0, the selection of variables differs for the two algorithms. When maxgain = 0, in GSAT(L) any variable v with gain = 0 that is not in the tabu list can be selected. But in USAT(L), the same variable v can get selected only if it occurs in an unsatisfied clause.

3.6 Algorithm Novelty

In the Novelty algorithm [17], when selecting a variable from a randomly selected unsatisfied clause, we make use of the concept of the age of a variable, which is the flip number of the last flip of the variable. The variable having the maximum age is therefore the most recently flipped one. Consider the best and the second best variable. If the best variable is not the most recently flipped variable in the clause, then select it. Otherwise, with probability p select the second best variable, and with probability 1-p select the best variable. The Novelty algorithm is greedier than WalkSAT since only one of the two most improving variables from a clause is selected. This can lead either to a significantly improved performance or to getting stuck in local minima. Novelty+ [11] is a variant of Novelty that incorporates random walk.

3.7 Algorithm RNovelty

The RNovelty algorithm [17] is a variant of Novelty. It is identical to Novelty except when the best variable is the most recently flipped one. In this case, let n be the difference in the objective function between the best and the second best variable. Then there are four cases: i) When $p < 0.5$ and $n > 1$, pick the best. ii) When $p < 0.5$ and $n = 1$, then with probability $2*p$ pick the second best, otherwise pick the best. iii) When $p > 0.5$ and $n = 1$, pick the second best. iv) When $p > 0.5$ and $n > 1$, with probability $2*(p - 0.5)$ pick the second best, otherwise pick the best. RNovelty+ [11] is a variant of Novelty that incorporates random walk.

4 Experimental Results

Let us summarize our experimental observations on the Nurse Scheduling Problem. Methods were run on Intel 1.8 GHz PC with 4 GB RAM. For comparing the performance of different methods identical problem instances were considered. We computed the number of problems solved in a set of 100 problems of same instance and their average runtime in seconds. The average runtime was taken over the solved instances. The comparative performance of GA1, GA2, SA, GSAT(L), WalkSAT, USAT(L), Novelty, Novelty+, Rnovelty, and Rnovelty+ in solving eight different instances of NSP is shown in Figure-1. M and N here represent the duration (in days) and number of nurses respectively. In *remarks* Pr = probability of selection, rp = random walk probability.

M	N	Observation	GA1	GA2	SA	GSAT	WalkSAT	USAT	Novelty	Novelty+	Rnovelty	Rnovelty+
7	10	*Solved*	100	100	100	100	100	100	100	100	100	100
		Time	2.21	1.34	0.68	0.01	0.019	0.012	0.02	0.017	0.015	0.02
		Remarks							Pr = 85	Pr = 85, rp = 2	Pr = 60	Pr = 60, rp = 2
7	15	*Solved*	91	95	96	100	100	100	100	100	100	100
		Time	3.67	2.67	1.54	0.8138	1.02	0.8152	0.9645	0.9588	0.8115	0.9481
		Remarks							Pr = 85	Pr = 70, rp = 2	Pr = 60	Pr = 60, rp = 2
7	20	*Solved*	88	94	100	100	100	100	100	100	100	100
		Time	6.78	5.42	2.08	18.79	25.28	22.82	22.12	29.09	25.09	27.01
		Remarks							Pr = 85	Pr = 85, rp = 2	Pr = 60	Pr = 60, rp = 2
14	10	*Solved*	99	100	100	100	100	100	100	100	100	100
		Time	2.47	2.45	1.88	0.2693	0.2605	0.2502	0.9977	0.2901	1.78	27.01
		Remarks							Pr = 70	Pr = 85, rp = 2	Pr = 60	Pr = 60, rp = 2
14	15	*Solved*	41	95	100	100	100	100	100	100	100	100
		Time	10.91	5.54	2.98	1.77	2.15	1.549	16.64	8.14	26.75	29.01
		Remarks							Pr = 85	Pr = 85, rp = 2	Pr = 60	Pr = 85, rp = 2
14	20	*Solved*	40	94	100	100	100	100	100	100	100	100
		Time	17.69	11.02	3.99	24.81	25.21	24.71	175.18	40.32	196.18	54.68
		Remarks							Pr = 85	Pr = 85, rp = 2	Pr = 85	Pr = 60, rp = 2
21	10	*Solved*	96	98	100	100	100	100	100	100	100	100
		Time	8.35	7.85	3.64	29.80	28.33	25.02	29.09	26.79	28.45	65.74
		Remarks							Pr = 70	Pr = 85, rp = 2	Pr = 85	Pr = 85, rp = 2
21	15	*Solved*	39	90	100	100	100	100	100	100	100	100
		Time	19.11	18.51	5.98	57.68	42.80	44.28	125.38	51.18	90.69	54.21
		Remarks							Pr = 85	Pr = 85, rp = 2	Pr = 85	Pr = 85, rp = 2

Fig. 1. Comparative Performance of different methods in NSP instances

For each instance 100 problems were randomly generated. This means that for a particular instance, suppose, for the two-week roster of 15 nurses, the constraints will remain same; but 100 initial solutions will be generated randomly. We wanted to create instances that were realistic and indicative of real life situations. For this purpose we collected data from Peerless Hospital in Kolkata.

We have applied meta-heuristics like Simulated Annealing and Genetic Algorithm to solve it. We can see that overall performance of SA is very good. Genetic Algorithm has been implemented in two ways, GA1 and GA2, where GA1 is the basic

version and GA2 is a modification over it. The performance of GA2 is remarkably better than that of GA1.

We can see from Figure-1 that USAT (L) performs slightly better than other SAT methods in five instances: 7-days, 14-days and 21-days roster with 10 nurses; 14-days roster with 15 nurses and the same with 20 nurses. In other instances, the performance of USAT (L) is very competitive. The most important thing is that all of the SAT methods are able to satisfy all the hard and soft constraints. But as the size of the problem increases, the number of clauses increases rapidly. With our present machine we could solve up to the size given in Figure-1. Machine with larger RAM will be able to solve instances of larger size.

5 Conclusion

The Nurse Scheduling Problem (NSP) is modeled as a combinatorial optimization problem. To apply local search SAT techniques, NSP is converted to SAT instances. Our earlier experiments indicated that the greedy local search SAT techniques using tabu list solves this problem quite efficiently. In this paper we have compared the performance of different SAT techniques with SA and GA versions.

We also observe that SAT solvers are very close to each other in performance. The USAT (L) is slightly better than other methods. All the instances are solved satisfying all the constraints. But these SAT-based methods have a limitation. The number of clauses increases rapidly with increase in the size of the problem.

There can be situations where NSP may be over constrained. In that case no satisfying assignment exists. But we can get feasible solutions if the clauses corresponding to the hard constraints are satisfied. During the execution, whenever a truth assignment satisfying the hard constraints is found, it is immediately stored. This can only be replaced by another feasible solution with lesser cost. This implementation enables us to return the best feasible solution with minimum cost.

Some future works can be suggested. We can incorporate more constraints to the problem instances reflecting the real life situations. Nurses may submit their preferences for duties in terms of a combination <day no, shift >. If any conflict occurs it can be resolved by considering their seniority or experience or any other humanitarian ground.

References

1. Abdennadher, S., Schlenker, H.: Nurse scheduling using constraint logic programming. In: Proc. of the 11th Annual Conference on Innovative Applications of Artificial Intelligence (IAAI-1999), Orlando, Florida, United States, July 18-22, pp. 838–843 (1999)
2. Acharyya, S.: The Satisfiability Problem: A Constraint Satisfaction Approach, PhD Thesis, Computer Science & Engg., Calcutta University (2001)
3. Acharyya, S., Bagchi, A.: A SAT Approach for Solving The Staff Transfer Problem. In: Ao, S.I., Castillo, O., Douglas, C., Feng, D.D., Lee, J.-A. (eds.) IMECS-2008, Proceedings of International MultiConference of Engineers and Computer Scientists, Hong Kong. Lecture Notes in Engineering and Computer Science, pp. 64–68 (2008)

4. Aickelin, U., Dowsland: Exploiting problem structure in a genetic algorithm approach to a nurse rostering problem. Journal of Scheduling 3, 139–153 (2000)
5. Aickelin, U., Kathryn, A., Dowsland: An indirect genetic algorithm for a nurse-scheduling problem. Computers and Operations Research 31(5), 761–778 (2004)
6. Arthur, J.L., Ravindran, A.: A Multiple Objective Nurse Scheduling Model. AIIE Transactions 13, 55–60 (1981)
7. Azaiez, M.N., Al Sharif, S.S.: A 0-1 goal programming model for nurse scheduling. Computers and Operations Research 32(3), 491–507 (2005)
8. Dowsland: Nurse scheduling with tabu search and strategic oscillation. European Journal of Operational Research 106(2-3), 393–407 (1998)
9. Dowsland, Thompson: Solving a nurse scheduling problem with knapsacks, networks and tabu search. Journal of the Operational Research Society 51(7), 825–833 (2000)
10. Fukunaga, A.S.: Variable Selection Heuristics in Local Search for SAT. In: Proc. AAAI-1997, pp. 275–280 (1997)
11. Hoos, H.H.: On the Run-time Behavior of Stochastic Local Search Algorithms for SAT. In: Proc. AAAI-1999, pp. 661–666 (1999)
12. Ikegami, A., Niwa, A., Ohkura, M.: Nurse scheduling in Japan. Commun. Oper. Res. Society of Japan 41, 436–442 (1996) (in Japanese)
13. Kumar, V.: Algorithms for Constraint Satisfaction Problems: A Survey. A I Magazine 13(1), 32–44 (1992)
14. Kundu, S., Mahato, M., Mahanty, B., Acharyya, S.: Comparative Performance of Simulated Annealing and Genetic Algorithm in Solving Nurse Scheduling Problem. In: Ao, S.I., Castillo, O., Douglas, C., Feng, D.D., Lee, J.-A. (eds.) IMECS-2008, Proceedings of International MultiConference of Engineers and Computer Scientists, Hong Kong. Lecture Notes in Engineering and Computer Science, pp. 96–100 (2008)
15. Kundu, S., Acharyya, S.: A SAT Approach for Solving The Nurse Scheduling Problem. In: Proc. IEEE TENCON-2008, CD, Session P16 (Algorithms and Architecture), Hyderabad, India, November 18-21, 6 Pages (2008)
16. Kundu, S., Acharyya, S.: Performance of WalkSAT in Solving The Nurse Scheduling Problem. International Journal of Information Processing 4(1), 46–53 (2010)
17. MacAllester, D., Selman, B., Kautz, H.: Evidence for Invariants in Local Search. In: Proc. AAAI-1997, pp. 321–326 (1997)
18. Mazure, B., Sais, L., Gregoire, E.: Tabu Search for SAT. In: Proc. AAAI-1997, pp. 281–285 (1997)
19. Mitchell, M.: An Introduction to Genetic Algorithms. MIT Press (1996)
20. Millar, Kiragu: European Journal of Operational Research 104(3), 582–592 (February 1, 1998)
21. Musa, A., Saxena, U.: Scheduling Nurses Using Goal-Programming Techniques. IIE Transactions 16, 216–221 (1984)
22. Post, G., Veltman, B.: Harmonious Personnel Scheduling. In: Proceedings of the 5th International Conference on the Practice and Automated Timetabling, PATAT 2004, pp. 557–559 (2004)
23. Reeves, C.R.: Modern Heuristic Techniques for Combinatorial Problems. Orient Longman (1993)
24. Selman, B., Levesque, H.J., Mitchell, D.J.: A New Method for Solving Hard Satisfiability Problems. In: Proc. AAAI-1992, pp. 440–446 (1992)
25. Warner, H.W.: Scheduling Nursing Personnel According to Nursing Preference: A Mathematical Approach. Operations Research 24, 842–856 (1976)

GA Based Denoising of Impulses (GADI)

Jyotsna Kumar Mandal and Somnath Mukhopadhyay

Department of Computer Science and Engineering,
Kalyani University, Kalyani, West Bengal, India, 741235
jkm.cse@gmail.com, som.cse@live.com
http://jkmandal.com,http://www.klyuniv.ac.in

Abstract. In this paper we have proposed a novel method of removing random valued impulse noises from the digital images. A variable window such as 5 x 5 and 3 x 3 are utilized for such purpose. The proposed method is a switching median filter. The detection of noises in every 5 x 5 window of the noisy image is done using all neighbor directional weighted pixels. After detection of noisy pixel in the 5 x 5 window the proposed filtering scheme restored it to a pixel which is most suitable in the 3 x 3 and 5 x 5 window regions. This scheme is based on weighted median filtering on the 3 x 3 window regional pixels. Three user parameters of the proposed noise removal operator are searched in a 3D space using a randomized search and optimization technique i.e., Genetic Algorithm. Implementation of the scheme shows better noise removal performance and also preserves the image fine details well.

Keywords: All neighbor directional weighted pixels, de noising, genetic algorithm (GA), random valued impulse noise (RVIN), switching median filter, variable mask.

1 Introduction

Digital image gets corrupted by impulses during image acquisition or transmission because of perturbation in sensors and communication channels. Most common types of impulses are salt-and-pepper noise (SPN) and random valued impulse noise (RVIN). RVIN in the digital images is taken into account in this paper. Among the non linear filters, most popular method is standard median (SM) filter [2]. More effective schemes in noise suppression are center weighted median (CWM) filter [10] and adaptive center weighted median (ACWM) filter [4]. They give extra importance to some pixels of the filtering windows as a result these filters obtain better results compared to the standard median filter. The switching median filter has an impulse detector, used prior to filtering to classify the test pixel, such as iterative pixel-wise modification of MAD (PWMAD) (median of the absolute deviations from the median) filter [6] separates noisy pixels from the image details. The tri-state median (TSM) filter [3] and multi-state median (MSM) filter [5] are also switching median filters. The progressive

N. Chaki and A. Cortesi (Eds.): CISIM 2011, CCIS 245, pp. 212–220, 2011.

switching median filter (PSM) [16] performs noise detection and filtering iteratively. The signal-dependent rank ordered mean filter (SD-ROM) [1] uses rank order information for impulse detection and filtering. Some directional switching median filters, directional weighted median filter [8] and second order difference based impulse detection filter [15] have been developed to remove RVIN. Two consecutive switching median filters MWB[12] and MDWMF[13] are also surveyed in this paper.

Apart from the above mentioned filters such as median and mean based filters, several fuzzy based filter[14], neuro fuzzy based filter[11]., etc have also been devised. These type of filters performs better in terms of *PSNR(dB)* than other non soft computing based filters.

In this paper, a novel scheme has been proposed to suppress the random valued impulse noises which utilize all the neighborhood pixels in the 5 x 5 window. First we classify the center pixel in the test window as a noisy or noise free, by making the absolute differences of the center pixel with four directional pixels in the test window. We also check whether the center pixel lies within the maximum and minimum intensity range spread in the test window. To filter the noisy pixels both the 5 x 5 and 3 x 3 masks are used and a median based filter has been proposed. Three user parameters viz., number of iterations(I), threshold(T) and decreasing rate (R) of threshold in each iteration are varied and searched in 3 dimensional space to obtain optimal results. A genetic algorithm based optimization technique has been proposed for such purpose.

The performance of the proposed algorithm is experimented and compared with other methods under several noise densities on various bench mark images. Implementation of the proposed algorithm shows better noise suppressing quality and restoration of image details

Proposed impulse detection and filtering methods are given in sections 2 and 3 respectively. Proposed GA based optimization is described in section 4. Experimental results and discussions are demonstrated in section 5. Conclusions are given in section 6.

2 Impulse Detector

The proposed scheme uses 5 x 5 window in row major order of the noisy image to classify the center pixel as noisy or not, emphasizes on the pixels aligned in the four main directions along with two end pixels in each direction, shown in fig 1. The proposed impulse detection scheme is given in algorithm 1.

3 Impulse Filter

If any pixel is detected as noisy, the filtering scheme restores it to a pixel which is most suitable in the window. The technique has been depicted in algorithm 2.

Algorithm 1. Impulse detector

1: Let $y_{i,j}$ be the center pixel and W_{min} and W_{max} are the maximum and minimum intensity values respectively within the test window around $y_{i,j}$. Then the rule is given as:

$$y_{i,j} = \begin{cases} Undetected & : W_{min} < y_{i,j} < W_{max} \\ Noisy & : W_{min} \geq y_{i,j} \geq W_{max} \end{cases} \tag{1}$$

2: Let the set of seven pixels centered at (0, 0) in the k^{th} direction denoted by S_k (k=1 to 4), i.e.,

- S_1={(-1,-2),(-2,-2),(-1,-1),(0,0),(1,1),(2,2),(1,2)}.
- S_2={(1,-2),(0,-2),(0,-1),(0,0),(0,1),(0,2),(-1,2)}.
- S_3={(2,-1),(2,-2),(1,-1),(0,0),(-1,1),(-2,2),(-2,1)}.
- S_4={(-2,-1),(-2,0),(-1,0),(0,0),(1,0),(2,0),(2,1)}.

Then let $S_k^0 = S_k/(0,0)$, $\forall$ k =1 to 4.

3: Define $d_{i,j}^{(k)}$ as the sum of absolute differences of intensity values between $y_{i+s,j+t}$ and $y_{i,j}$ with (s,t)$\in S_k^0$ (k= 1 to 4), given in eqn 2.

4: Assign ω_m =2, ω_n=1 and ω_o=0.5.

$$d_{i,j}^{(k)} = (\sum_{(s,t)\in S_k^0} \omega_{s,t} |y_{i+s,j+t} - y_{i,j}|, 1 \leq k \leq 4) \tag{2}$$

where

$$\omega_{s,t} = \begin{cases} \omega_m & : (s,t) \in \Omega^3 \\ \omega_o & : (s,t) \in \Omega^2 \\ \omega_n & : \text{otherwise} \end{cases} \tag{3}$$

where

$$\Omega^3 = \{(s,t) : -1 \leq s,t \leq 1\}, and \tag{4}$$

where

$$\Omega^2 = \{(s,t) : (s,t) = \pm\{(-2,-1),(-1,-2),(1,-2),(2,-1)\}\}. \tag{5}$$

5: $d_{i,j}^{(k)}$ is known as direction index.

$$r_{i,j} = min\{d_{i,j}^{(k)} : 1 \leq k \leq 4\} \tag{6}$$

Three assumptions are made depending upon the values $r_{i,j}$.

1. $r_{i,j}$ is small when $y_{i,j}$ is on a noise free flat region.
2. $r_{i,j}$ is small when $y_{i,j}$ is on the edge.
3. $r_{i,j}$ is large when $y_{i,j}$ is noisy .

6: Form the complete decision rule to detect a noisy or noise free pixel depending upon the definition of $r_{i,j}$, by introducing a threshold (T) and which is given as

$$y_{i,j} = \begin{cases} Noisy\ Pixel & : W_{min} \geq y_{i,j} \geq W_{max} \\ Noise\ Free\ Pixel & : r_{i,j} \leq T\ and\ W_{min} < y_{i,j} < W_{max} \end{cases} \tag{7}$$

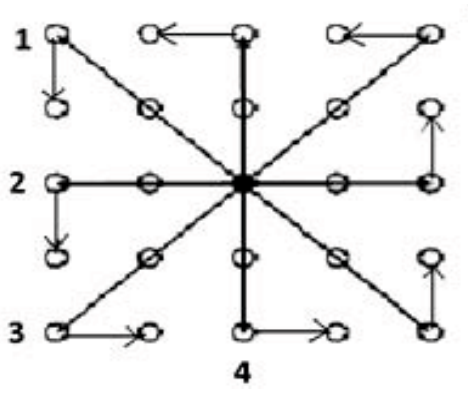


Fig. 1. All neighbor directional weighted pixels in the 5 x 5 window

Algorithm 2. Impulse filter

1: Calculate the standard deviations $\sigma_{i,j}^{(k)}$ of the intensity values of all $y_{i+s,j+t}$ with (s, t) $\in S_k^0$ (k= 1 to 4).

2:
$$l_{i,j} = \min_k \{\sigma_{i,j}^{(k)} : k = 1 \; to \; 4\} \tag{8}$$

3:
$$m_{i,j} = \max_k \{\sigma_{i,j}^{(k)} : k = 1 \; to \; 4\} \tag{9}$$

4: Select the directions within the 5 x 5 window where standard deviations are maximum or minimum. Use repetition operator $\diamond$ [2], for which standard deviation is minimum.

5: Calculate the median as
$$med = median\{\omega_{s,t} \diamond y_{i+s,j+t} : (s,t) \in \Omega^4\} \tag{10}$$
where
$$\Omega^4 = \{(s,t) : -1 \leq s,t \leq 1\} \; and \; (s,t) \neq (0,0). \tag{11}$$
and where
$$\omega_{s,t} = \begin{cases} \omega_m & : (s,t) \in s^0_{m_{i,j}} \\ \omega_l & : (s,t) \in s^0_{l_{i,j}} \\ \omega_n & : \text{otherwise} \end{cases} \tag{12}$$

6: Assign $\omega_m = 0, \; \omega_l = 2 \; and \; \omega_n = 1$.

7: Replace $y_{i,j}$ by med.

4 GA Based Optimization

In the proposed optimization process three user parameters are taken from the 3D search space using a randomized search and optimization technique, i.e.,

Genetic Algorithm (GA). It is a randomized search and optimization technique guided by the principle of natural genetic systems and is inspired by the biological evolution process. Initial population of individuals are encoded randomly followed by the fitness of all individuals are evaluated. Until a termination condition is obtained fittest individuals are selected for reproduction, crossover and recombination by generating new population in each cascading stage, is given in algorithm 3.

5 Simulations and Results

The performance of the operator is implemented under various noise densities and on several popular 8 bit gray scale images with dimensions of 512 x 512 like *Boats*, *Bridge* and *Lena* etc,. Experimental simulation is done repeatedly on an image and the best results obtained are presented. The algorithm have been executed on the machine configuration as ACPI uni-processor with Intel® Pentium® E2180 @ 2.00 Ghz CPU and 2.98 Gbyte RAM with MATLAB 8a environment.

5.1 Comparisons

Fig. 2 shows the comparative visual restoration effects between the existing algorithms and the proposed filter when the *Lena* image is 60% noisy. Considering very high noise ratio and fine details/textures of the images, the the proposed filter can enhance the noisy image effectively.

Table 1. Comparison of restoration results in terms of *PSNR (dB)* for *Lena* image

Filter	40% Noisy	50% Noisy	60% Noisy
SM[2]	27.64	24.28	21.58
PSM[16]	28.92	26.12	22.06
ACWM[4]	28.79	25.19	21.19
MSM[5]	29.26	26.11	22.14
SD-ROM[1]	29.85	26.80	23.41
Iterative Median[9]	30.25	24.76	22.96
Second Order[15]	30.90	28.22	24.84
PWMAD[6]	31.41	28.50	24.30
DWM Filter[8]	32.62	30.26	26.74
GADI(Proposed)	32.90	30.87	28.49

Results obtained in the proposed GADI has been compared with different existing techniques on *Lena* image corrupted with 40% to 60% noise densities and given in table 1. Proposed filter performs significantly better than existing filters under the specified noise densities. In table 2 restoration results are compared

Algorithm 3. GA based Optimization Algorithm

1: **Chromosome Encoding and Initial Population:** Binary chromosomes **P** is used to encode the three parameters I, T and R as : Iteration I (3bits), Threshold T (10bits) and Decreasing Rate of Threshold (7bits). The size of the encoded chromosome(P) is of 20 bits. First 3 bits represent the number of iterations(I) to a maximum of 8, intermediate 10 bits used to encode the maximum threshold 1024 and last 7 bits to represent a value between 0.6 to 0.95. Let the decimal value of the first 3 bits is i, then the mapping is done as, if $(i < 3)$ then i=3 and if $(i > 6)$ then i=6. Let the decimal value of the intermediate 10 bits is t if $(t < 300)$ then t=300 and if $(t > 1000)$ then t=1000. Let the decimal value of the last 7 bits is d, and then the mapping as:

$$R(d) \ = 0.6 + \frac{0.95 - 0.6}{127 - 0} * d; \tag{13}$$

2: **Fitness Evaluation:** The objective is to maximize the value of *PSNR* using eq. 14, as a result this equation is used as fitness function/criterion f for the chromosomes in GA based optimization technique.

$$f(I_1, I_2) = PSNR(dB) = 10 * \log_{10}(\frac{255^2}{\frac{1}{M*N}\sum_{m,n}(I_1 m, n - I_2 m, n)^2}) \tag{14}$$

where M and N are the dimensions of the input images respectively. I_1 and I_2 are the original and enhanced images respectively.

3: **Elitism:** The best chromosome of the current generation is copied to the next generation without being involving it by the crossover and mutation stage. In proposed scheme, we have kept one copy of the best chromosome of each generation outside the population. But it goes through the cross over and mutation stage. If the worst chromosome of the current generation is better than the best chromosome of the previous generation then it survives otherwise is replaced by the best one saved.

4: **Selection:** We select better chromosomes to make the mating pool with same size of population using the Binary Tournament Selection(BTS) method [7]. We select two chromosomes randomly and then the best one is copied to the pool until the mating pool is full. Tie is resolved by selecting the chromosome which requires less number of iterations and shows better fitness.

5: **Crossover:** It takes place between randomly selected parent chromosomes from the mating pool. Uniform crossover [17] method is followed in proposed scheme. It's a probabilistic operation. It is repeated n/2 times for a mating pool of size n. Randomly two parent chromosomes are selected from the mating pool followed by generation of a binary mask of length 20 bits. If the mask bit is 1 then we swapped bitwise the parent chromosomes. Otherwise crossover does not occur for that bit position.

6: **Mutation:** Mutating a binary gene involves simple negation of the bit of the parent chromosome. It's also a probabilistic operation. It occurs with very low probability (μ_m) in each generation. If a randomly generated number is lies within the range specified for (μ_m), then mutation occurs, otherwise mutation does not occur for the particular bit position.

7: **Parameters of Genetic Algorithm:**

- Population size (P) - We have chosen P in the range [5, 10]
- Chromosome length (L)- fixed (20 bits)
- μ_c = [0.8-0.9] and μ_m = [0.01-0.1]
- The simulation scheme executed the program for [30, 40] generations.

Fig. 2. Results of different filters in restoring 60% corrupted image Lena, (a) Original image (b)Noisy Image (c) (SD-ROM)[1] (d)(MSM)[5] (e)(PWMAD)[6] (f)(DWM)[8] (g)**GADI(Proposed)**

Table 2. Comparison of restoration results in terms of *PSNR (dB)* for *Bridge* image

Filter	40% Noisy	50% Noisy	60% Noisy
ACWM[4]	23.23	21.32	19.17
MSM[5]	23.55	22.03	20.07
SD-ROM[1]	23.80	22.42	20.66
Second Order[15]	23.73	22.14	20.04
PWMAD[6]	23.83	22.20	20.83
DWM Filter[8]	24.28	23.04	21.56
GADI(Proposed)	24.73	24.34	23.88

Table 3. Comparison of restoration results in terms of *PSNR (dB)* for *Boat* image

Filter	40% Noisy	50% Noisy	60% Noisy
ACWM[4]	26.17	23.92	21.37
MSM[5]	25.56	24.27	22.21
SD-ROM[1]	26.45	24.83	22.59
PWMAD[6]	26.56	24.85	22.32
DWM Filter[8]	27.03	25.75	24.01
GADI(Proposed)	28.21	27.75	26.31

with existing algorithms for *Bridge* image. Proposed (GADI) performs better than any existing filter considered. From table 3 it is seen that in restoring 40% to 60% noisy *Boat* images the DWM[8] filter performs better in terms *PSNR(dB)* than other existing filters. But proposed (GADI) performs better than DWM[8].

6 Conclusion

In this paper, a novel method to suppress random valued impulse noises in digital images has been presented in which tuning parameters are searched in supervised way through GA-based optimization technique. The detection operator classifies the center pixel by making absolute differences of the pixels aligned in the four main directions with the center pixel in the 5 x 5 window along with a simple decision rule based on comparison of intensity values. The filtering operator uses modified and advanced median filtering technique on some particular pixels in the 3 x 3 window. Prior to compute median operation on the pixels are selected by calculating standard deviation on the four main directional pixels in the 5 x 5 window. The proposed filter is very accurate and easy-to-implement for denoising random valued impulses. GA based operator for such purpose in digital images performs well in terms of PSNR (dB) and restoration results in visual form.

Acknowledgments. Authors express deep sense of gratitude towards the Dept of CSE, University of Kalyani and the IIPC Project, AICTE, (Govt. of India), of the department where the computational recourses are used for the work.

References

1. Abreu, E., Lightstone, M., Mitra, S.K., Arakawa, K.: A new efficient approach for the removal of impulse noise from highly corrupted images. IEEE Transactions on Image Processing 5(6), 1012–1025 (1996)
2. Brownrigg, D.R.K.: The weighted median filter. Communications of the ACM 27(8), 807–818 (1984)
3. Chen, T., Ma, K., Chen, L.: Tri-state median filter for image de noising. IEEE Transaction Image Processing 8(12), 1834–1838 (1999)
4. Chen, T., Wu, H.R.: Adaptive impulse detection using center weighted median filters. IEEE Signal Processing Letters 8(1), 1–3 (2001)
5. Chen, T., Wu, H.R.: Space variant median filters for the restoration of impulse noise corrupted images. IEEE Transactions on Circuits and Systems-II: Analog and Digital Signal Processing 48(8), 784–789 (2001)
6. Crnojevic, V., Senk, V., Trpovski, Z.: Advanced impulse detection based on pixel-wise mad. IEEE Signal Processing Letters 11(7), 589–592 (2004)
7. Goldberg, D.E.: Genetic algorithm in search, optimization and machine learning. Addison- Wesley (1989)
8. Dong, Y., Xu, S.: A new directional weighted median filter for removal of random - valued impulse noise. IEEE Signal Processing Letters 14(3), 193–196 (2007)
9. Forouzan, A.R., Araabi, B.: Iterative median filtering for restoration of images with impulsive noise. Electronics, Circuits and Systems 1, 232–235 (2003)
10. Ko, S.J., Lee, Y.H.: Center weighted median filters and their applications to image enhancement. IEEE Transactions on Circuits and Systems 38(9), 984–993 (2001)
11. Kong, H., Guan, L.: A neural network adaptive filter for the removal of impulse noise in digital images. Neural Networks Letters 9(3), 373–378 (1996)
12. Mandal, J.K., Sarkar, A.: A novel modified directional weighted median based filter for removal of random valued impulse noise. In: International Symposium on Electronic System Design, pp. 230–234 (December 2010)
13. Mandal, J.K., Sarkar, A.: A modified weighted based filter for removal of random impulse noise. In: Second International Conference on Emerging Applications of Information Technology, pp. 173–176 (February 2011)
14. Russo, F., Ramponi, G.: A fuzzy filter for images corrupted by impulse noise. IEEE Signal Processing Letter 3, 168–170 (1996)
15. Sa, P.K., Dash, R., Majhi, B.: Second order difference based detection and directional weighted median filter for removal of random valued impulsive noise. IEEE Signal Processing Letters, 362–364 (December 2009)
16. Wang, Z., Zhang, D.: Progressive switching median filter for the removal of impulse noise from highly corrupted images. IEEE Transactions on Circuits and Systems 46(1), 78–80 (1999)
17. Michalewicz, Z.: Genetic algorithms +data structures = evolution programms. Springer, Heidelberg (1996)

Handling Write Lock Assignment in Cloud Computing Environment

Sangeeta Sen and Rituparna Chaki

West Bengal University of Technology
BF-142, Salt Lake, Kolkata, India
{sangeetaaec,rituchaki}@gmail.com

Abstract. Cloud Computing is a newly developed technology for complex system with large scale resource, information, software sharing with the help of agents. The persisting problem is in the assignment of data locks to the users in case of conflicts. This paper deals with a technique of handling data locks by providing a novel architecture lest a conflict occurs between the users. The architecture uses multiple agents including mobile agents and fixed agent. The proposed scheme enhances the speed of write lock assignment in the SaaAS (Software as an Agent Service) Environment quite efficiently with the help of passing few messages. It also takes a glance into the portability and interoperability of agents in a Cloud computing environment.

Keywords: Cloud computing, fixed and Mobile Agent, Write Lock access algorithm.

1 Introduction

Cloud Computing can be defined as accessing the shared resources, software and information on web and paying as per usage. Scalability is one of the main features of cloud computing and is always transparent to the users. Cloud computing system logically divided into three different layers. The lowest layer, Infrastructure as a Service (*IaaS)* provides basic infrastructure component. Above it Platform as a Service (*PaaS)* provide more platform oriented services, allows the use of hosting environment as per specific need. Finally at the top, Software as a Service (*SaaS)* features a complete application offered as service on demand [3]. Cloud Computing services are achieved not only by intra cloud communication, but also by inter cloud communication. Intra cloud communication are done by high speed LAN and inter cloud communication by low speed WAN [1]. Mobile agents migrate through the cloud and communicate with different agents and cloud service providers to provide inter cloud communication. Thus it can be assumed that they maintain interoperability and portability in a cloud computing environment [4].

In this paper, a cloud computing model using mobile agents is used to reduce the inter-operability problems. The rest of the paper is organized as follows: section 2 gives a brief review of current state of the art, section 3 gives the overview of proposed architecture to solve the problem of access lock. Finally a conclusion of the work discussed in section 4 along with the future scope.

N. Chaki and A. Cortesi (Eds.): CISIM 2011, CCIS 245, pp. 221–230, 2011.

2 Previous Works

Realization of Open Cloud Computing Federation Based on Mobile Agent [4] highlights the process of providing a uniform resource interface to the user by incorporating the services of multiple Cloud Computing Service Providers. It tries to realize portability and interoperability between different kinds of Cloud Computing platform through Mobile Agent. Agent-based Online Quality Measurement Approach in Cloud Computing Environment [2] shows the requirement of online evolution during the phase of running services to maintain the Quality of Service by Cloud Computing environment. The fixed Agent with a defined architecture could provide the potential ability to fulfill the requirement. SaaAS – the Mobile Agent based Services for Cloud Computing in Internet Environment [1] gives an idea of Divided Cloud and Convergence coherence mechanism of SaaAS where every data has the "Read Lock" and "Write Lock". The cooperation between Main Consistence Server (MCS) and Domain Consistence Server (DCS) guarantees to achieve the cache coherence in SaaAS.DCS converge the lock requests of all users in its domain.

After analyzing the paper on SaaAS [1] it is found to lack information regarding the process of assigning write lock to user in case a conflict arises. The paper does not provide any method of prioritization of requests for write lock from users. The existing methodologies [1, 2 and 4] do not offer any solution in case the MCS is damaged and the control over the cloud environment is lost. Although agent based techniques are commonly used for cloud computing, the present review work finds no solution in case the agent is lost in the middle of a transaction creating a communication gap.

3 The Proposed Architecture

Here we propose a new methodology for assigning data locks while taking care of the above problems. This architecture uses multiple agents to provide the solution of the interoperability problems as discussed so far. Mobile agents are able to migrate bringing their own code and execution state. Fixed agents are not able to migrate but able to do the specific job as describe by the cloud developer.

3.1 Cloud Computing Based on Agent

The SaaAS architecture is made up of certain Working groups containing the domains. Each domain or cloud is made up of users, fixed agent and server which are connected through high bandwidth LAN. These domains are interlinked through low bandwidth WAN. Some domains are major domains located in low populated working groups and others have servers called DCS (Domain Consistence Server). The major domains contains Database Consistence Server (DBCS) which maintain the database of all the information flowing through the network and MCS (Main Consistence Server) which serves to provide the users with data lock. Mobile agents

are acting as the key driver for the user in the process of acquiring data lock from the MCS and the fixed agents are used for scheduling of request.

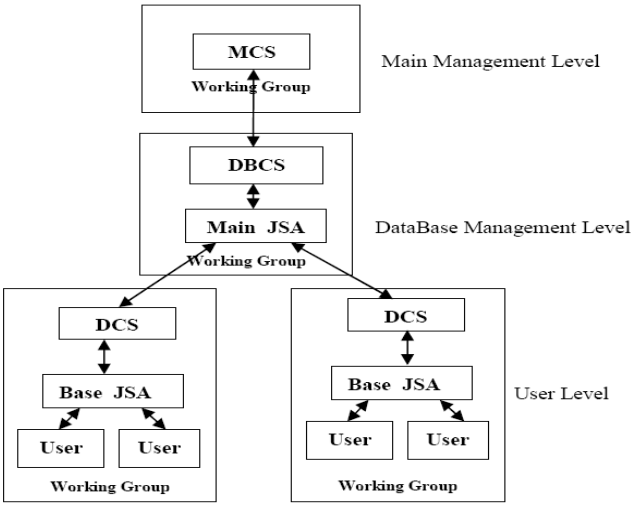


Fig. 1. The Proposed Architecture

The proposed architecture consists of three levels as follows:
User level: This level manages all the users in different Working Groups.
Database management level: This level actually manages the main Database Server where records of all the transaction are maintained.
Main management level: Main Consistent Server present in this level provides lock to the users. The agents can be classified as Mobile agents, Fixed agents and Group agents.

Description of Mobile Agents:

Data Structure	Description
IA (Interface Agent) (Source_Id, Termination_Time, forward/receive, Status, message)	This Agent takes the information from the user to the DCS.
WA (Working Agent) (Source_Id, Termination_Time, Forward/receive, Status, message)	This Agent takes the information from DCS to the DBCS.
JSA (Job Scheduling Agent) (Source_Id, Termination_Time, Forward/receive, Status, message)	This Agent communicates with DBCS and MCS.
RA (Reply Agent) (Source_Id, Termination_Time, Forward/receive, Status, message)	This Agent is used to send the reply from MCS or DBCS to user with a Lock_Grant message.
QA (Queuing Agent) (Source_Id, Termination_Time, Forward/receive, Status, message)	This Agent is used to send the waiting message from DBCS to user.
UA (Update Agent) (Source_Id, Termination_Time, Forward/receive, Status, message)	This Agent is used to update the MCS database. This message goes to MCS from DBCS.

Description of Fixed Agents:

Data Structure	Description
Base_JSA (Base_Job Scheduling Agent)	This Agent schedule the request came from users
Main_JSA (Main_Job Scheduling Agent)	This Agent schedule the request came from DCSs.

Description of Group Agents:

Data Structure	Description
DMA (Domain Management Agent)	Manage all the mobile Agents presents in a domain.
FDMA (Fixed Domain Management Agent)	Manage all the fixed Agents in a domain.
GMA (Group Management Agent)	Manage DMA and FDMA in a working group.
MMA (Main Management Agent)	Manage all the GMA.

3.2 Data Structure

Detail Description of the Messages:

Data Structure	**Description**
Req_msg (Data_File_Id, Lock_Req, Completion_Time, Local_Time_Stamp, Authentication_Code)	This message carries the information related to request of lock from USER to DCS.
MUReq_msg (Data_File_Id, Lock_Req, Completion_Time, Lock_Status, User_Id)	This message carries the information from DCS to DBCS if the pre-existing lock is found in DCS.
UReq_msg (Data_File_Id, Lock_Req, Completion_Time, User_Id)	This message carries the information from DCS to DBCS if the pre-existing lock is not found in DCS.
URep_msg (Destination_Id, Lock_acquire_time, Lock_Grant, User_Id)	This message carries the information from DBCS to USER if the lock is already acquired by another user.
Lock_msg (Data_File_Id, Lock_Req, User_Id)	This message carries the information from DBCS to MCS if the lock is not acquired by another user.
Rep_msg (Destination_Id, Data_File_Id, Lock_Grant, User_Id)	This message carries the information from MCS to USER when a lock is first time granted.
WRep_msg (Destination_Id, Data_File_Id, Lock_Grant , User_Id)	This message carries the information from DBCS to USER when DBCS grant the lock request to waited user.
Update_msg (Data_File_Id)	This message updates the data base of MCS. So it carries the information from DBCS to MCS. MCS update its database repeatedly.

Next, the detailed descriptions of the three types of Servers are presented.

Domain Consistence Server (DCS)

Data Structure	Description
DCS_Main(Data_File_Id, Lock_Present, Completion_Time)	This table maintains the information of Data_File_Id, Lock_Present and Completion_Time of the user.
DCS_Back (Data_File_Id, User_Id, Authentication_Code)	This table maintains the information of Data_File_Id, User_Id and Authentication_Code of the user.

DataBase Consistence Server (DBCS)

Data Structure	Description
DBCS_Main (Data_File_Id, Lock_Present,Completion_Time)	This table maintains the information of Data_File_Id, Lock_Present and Completion_Time of the user coming from DCS.
DBCS_Back (Data_File_Id, User_Id, Lock_acquire_time, DCS_Id)	This table maintains the information of Data_File_Id, User_Id Lock_acquire_time and DCS_Id of the user coming from DCS.
DBCS_Wait (Data_File_Id, Completion_Time, User_Id, Lock_Req, DCS_Id)	This table maintains the information of Data_File_Id, Completion_Time, User_Id, Lock_Req and DCS_Id of the user coming from DCS when a conflict occurs between two requests.

Main Consistence Server (MCS)

Data Structure	Description
MCS_Main (Data_File_Id, Lock_Present, User_Id)	This table maintains the information of Data_File_Id, Lock_Present and User_Id of the user coming from DBCS.

Detail Description of Acknowledgement

Data Structure	Description
Ack (Src_Id, Dest_Id, Report)	This message assures the delivery of agents in the middle of transaction.

Descriptions of the field used in algorithm and flow chart:

Name	Description	Name	Description
a	Data_File_Id	b	Lock_Req
c	Completion_Time	d	Local_Time_Stamp
e	Authentication_Code	f	Lock_Status
g	User_Id	m	Lock_acquire_time
n	Lock_Grant	x	Destination_Id
y	Lock_Present	i	Source_Id
j	Termination_Time	k	Forward/receive
s	Status	z	DCS_Id

3.3 Proposed Logic

The user requests for a write lock from the server by generating an interface agent and loading it with a Req_msg. The Base_JSA schedules this package on its way to DCS. DCS checks its own tables for pre-existing write lock and creates UReq_msg/MUReq_msg. IA is duplicated to WA and sent along with the message to Main_JSA by DCS, besides an Ack Message to User. After getting scheduled by Main_JSA, the request is received by DBCS. DBCS checks its own table for pre-existing write lock and creates Lock_msg/ URep_msg. WA duplicated to JSA or QA. DBCS sends the request (Lock_msg and JSA) to MCS; also DBCS sends the package (QA and URep_msg) to user via Main_JSA, DCS, and Base_JSA. An Ack message is sends to DCS by DBCS via Main_JSA. MCS assign the lock and conveys the reply (RA and Rep_msg) to the user through the channel (DBCS, Main_JSA, DCS, Base_JSA) and updates its own table. In case users present in DBCS_Back table are next to receive the access DBCS sends package (WRep_msg and RA) to notify user (Scheduled for receiving access priority wise) about their status.

3.4 Case Study

Case1: No User is author and/or no pre-existing Write Lock

Let us consider that User1 and User8 want the Write lock on a Data file D simultaneously. For that User1 creates an Agent 'IA1' and User8 creates an Agent 'IA8', and Req_msg.

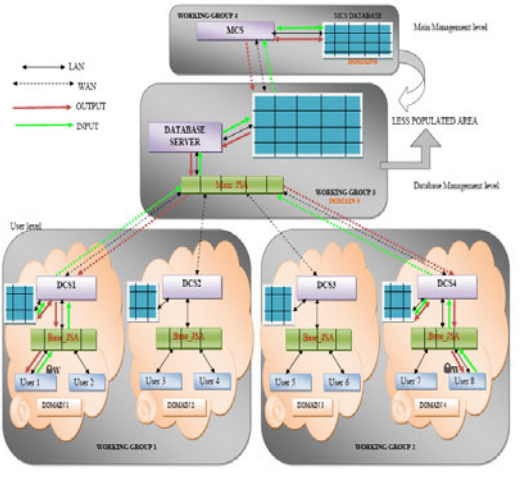


Fig. 2. Case 1- None of the Users is Author

Agent 'IA1' is dispatched to Base_JSA1 and 'IA8' is dispatched to Base_JSA4 for scheduling the requests. Base_JSA8 dispatch agents to DSC1 & DCS2. After scheduling based on Status and Timestamp Base_JSA1 and DCS1 and DCS2 check their DCS_Main tables for the presence of any pre-existing Write lock on the file 'D'.

If no such Write Lock is found in their respective tables, then an UReq_msg and WA1 & WA4 are created by duplicating IA and send the request by Both DCSs to Main_JSA. On receiving the request Main_JSA schedule the request based on Status and Completion_Time, send it to DBCS_Wait table. DBCS check its DBCS_Main table for pre-existing Write Lock for the first request. In case no pre-existing write lock is found then creates Lock_msg and JSA1 (duplicate of WA) and send the request to MCS.MCS grant the request and send it to DBCS with the help of RA and Rep_msg, also update its MCS_Main table. DBCS update its table and check that Data_File_Id of Rep_msg is present on DBCS_Wait table, if yes then updates DBCS_Back table after calculating the Lock_acquire_time. URep_msg is created and send it to User through QA (duplicate of WA). If no match found in DBCS_Wait table then repeat the same process done for the first request and forward the Rep_msg to the respective user. After completion the task of first request of a user, second user gets the Write Lock and this information goes to User using WRep_msg and duplicating WA. If pre-existing write lock exists, both requests goes to DBCS_Back table and DBCS calculates the Lock_acquire_time for both the users. DBCS creates URep_msg, QA1 and QA2 after duplicating WA. DBCS send the msg to the users. After completion the task of User3, User1 and User8 get the Write Lock based on Lock_acquire_time and this information goes to Users with the help of WRep_msg and duplicating RA.

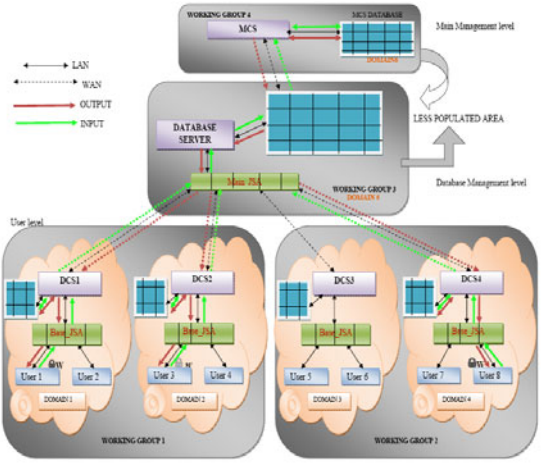


Fig. 3. Case 2 – User 1 in Author

Case 2: One of the users is author and there is a preexisting write lock

Say, User 1 is the author. DBCS check its DBCS_Main table for pre-existing Write Lock. Here DBCS found that User 3 already have the Write lock on Data File 'D', In the DBCS Completion_Time of author may be checked with a fixed value. If Completion_Time is greater than the fixed value no other changes takes place in the database, on the other if it is lesser then the fixed value the author get the priority and get the next execution position and the author's Completion_Time is added to all the other user's execution time. DBCS calculates the Lock_acquire_time for both User1

and User8. DBCS creates URep_msg and an agent QA1 and QA2. DBCS sends the message to the User1 and User8 through QA1 and QA2 respectively. On completion of the task of User3, first User1 and then User8 gets the Write Lock and updates DBCS_Main table.

3 Algorithm

The algorithm is divided into 6 segments.

a. User_lock_request algorithm

($IA_{i,j,k,s,p}$,$Req\_msg_{a,b,c,d,e}$,$RA_{i,j,k,s,p}$,$URep\_msg_{x,m,n,g}$/$Rep\_msg_{x,a,n,g,}$/$WRep\_msg_{x,a,n,g}$)

```
/* process of generating user lock request */
create IA and required Req_msg and concatenate with IA;
ssnd IA to Base_JSA for getting the lock;
read RA with Rep_msg /URep_msg /WRep_msg from Base_JSA for lock access;
if Lock_Grant=0 (URep_msg) then{
   wait for Duration= Lock_acquire_time (m); read WRep_msg from Base_JSA;
}
else Lock_Grant =1 (Rep_msg/WRep_msg){
   user will retrieve data of D from remote server and save the data in local  buffer;
}
```

b. Base_JSA scheduling algorithm

($IA_{i,j,k,s,p}$,$Req\_msg_{a,b,c,d,e}$,$RA_{i,j,k,s,p}$,$URep\_msg_{x,m,n,g}$/$Rep\_msg_{x,a,n,g}$,/$WRep\_msg_{x,a,n,g}$)

```
/* process showing Base_JSA schedule the request received by it  */
get the IA with Req_msg from Users;
if Forward/receive =1 (Request forward) then{
   compare Status with previous request Status;
   while Status same then
      compare Local_Time_Stamp; assign priority; send IA to DCS;
}
else Forward/receive =0 (Request receive){
   read RA from DCS; check Status for assigning priority; send RA to User;
}
```

c. Domain_server algorithm

($IA_{i,j,k,s,p}$,$Req\_msg_{a,b,c,d,e}$,$RA_{i,j,k,s,p}$,$URep\_msg_{x,m,n,g}$/$Rep\_msg_{x,a,n,g}$,/$WRep\_msg_{x,a,n,g}$, $WA_{i,j,k,s,p}$,$MUReq\_msg_{a,b,c,f,g}$,/$UReq\_msg_{a,b,c,g}$)

```
  /* process showing how DCS modified its table and genarate requests */
receive the IA with Req_msg from Base_JSA;
if Status of IA=1{
   check Authentication Code;
   if match not found  then
      assign Status=0; check DCS_Main table for pre-existing Write lock;
      create duplicate of IA to WA;
   check Lock_Present (y) status;
   switch(y){
      Case 0: create UReq_msg and concatenate with WA.
      Case 1: create MUReq_msg and concatenate with WA.
   }
```

```
    send MUReq_msg/ UReq_msg to Main_JSA
    (update);
}
receive Rep_msg/URep_msg/WRep_msg from Main_JSA;
(update);
update (a,c,g,y){
    update DCS_Main and DCS_Back table.
}
send RA to Base_JSA.
```

d. Main_JSA scheduling algorithm

($WA_{i,j,k,s,p}$,$MUReq\_msg_{a,b,c,f,g}$,/$UReq\_msg_{a,b,c,g}$,$RA_{i,j,k,s,p}$,$URep\_msg_{x,m,n,g}$/$Rep\_msg_{x,a,n,g}$,/$WRep\_msg_{x,a,n,g}$)

```
/* process showing Main_JSA schedule the request received by it  */
get the WA with MUReq_msg/ UReq_msg from DCS;
if Forward/receive =1 (Request forward) then{
    compare Status with previous request Status;
    while Status same then compare Completion_Time; assign priority;
    send WA to DBCS;
}
else if Forward/receive =0 (Request receive){
    read RA or QA with URep_msg from DBCS;
    check Status for assigning priority; send RA or QA to DCS.
}
```

e. DataBase_server algorithm

($WA_{i,j,k,s,p}$,$MUReq\_msg_{a,b,c,f,g}$,/$UReq\_msg_{a,b,c,g}$,$RA_{i,j,k,s,p}$,$URep\_msg_{x,m,n,g}$/$Rep\_msg_{x,a,n,g}$,/$WRep\_msg_{x,a,n,g}$,$JSA_{i,j,k,s,p}$,$Lock\_msg_{a,b,g}$,$UA_{i,j,k,s,p}$,$Update\_msg_{a}$)

```
/* process showing how DBCS modified its table and genarate requests */
get the WA from Main_JSA;
check();
check(x,a,n,g){
   if Lock Status not present then{
      search DBCS_Main table for pre-existing write  lock.
      Switch(y){
      Case 1: Update ();
      Case 0: create JSA duplicating WA and Lock_msg. Concatenate each other
      and send to MCS.
    }
    else Lock Status  present
    Update ();
}
update (a,g,m,z){
    calculate Lock_acquire_time and update DBCS_Back;
    create URep_msg and QA duplicating WA;
    send QA with URep_msg to Main_JSA;
 }
get RA from MCS;
if DFID of DBCS_Wait =DFID of Rep_msg then set status_indicator=0; Update ();
```

```
}
else{
   send RA with Rep_msg to Main_JSA; set status_indicator=1 and Check ();
}
grant the lock to user in DBCS_Back. Create WRep_msg and send to Main_JSA;.
check DBCS_Main table for Data_File_Id and corresponding Lock_Present;
if not present{
  create both UA and Update_msg. Concatenate each other and send to MCS
}
```

f. Main_server algorithm
($JSA_{i,j,k,s,p}$,$Lock\_msg_{a,b,g}$,$UA_{i,j,k,s,p}$,$Update\_msg_{a}$,$RA_{i,j,k,s,p}$,$Rep\_msg_{x,a,n,g}$)

```
 /* process showing how MCS grant the lock */
get Lock_msg from DBCS;
assign the Lock and update MCS_Main table;
create RA duplicating JSA and Rep_msg concatenate each other and send to DBCS;
receive UA. Update MCS_Main table.
```

4 Conclusion and Future Work

Cloud computing is the latest buzz in the computing industry. While many solutions have been proposed for cloud computation problems, the interoperability of cloud services still remain a challenge. This paper presents a novel architecture for cloud computing, using multiple agent services. As all requests are not directed, but are rerouted before MCS, the pressure or load on the MCS is substantially reduced. The use of multiple number of fixed and mobile Agents enables the jobs to be completed in a more efficient and effective manner. The idea of acquiring a lock on a Data File, as used in this architecture, can be extended to the concept of acquiring lock on resources. In this architecture message size can be controlled level wise depending upon the bandwidth available, thereby reducing the overall communication cost. This Architecture can be modified to reduce the number of message passing and extend the algorithm for read lock access with better conflict management.

References

1. Chen, G., Lu, J., Huang, J., Wu, Z.: SaaAS - The Mobile Agent based Service for Cloud Computing in Internet Environment. In: 2010 Sixth International Conference on Natural Computation (2010)
2. Liu, Z., Liu, T., Lu, T., Cai, L., Yang, G.: Agent-based Online Quality Measurement Approach in Cloud Computing Environment. In: 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (2010)
3. Jensen, M., Schwenk, J., Gruschka, N., Iacon, L.L.: On technical Security Issues in Cloud Computing. In: 2009 IEEE International Conference on Cloud Computing (2009)
4. Zhang, Z., Zhang, X.: Realization of Open Cloud Computing Federation Based on Mobile Agent. Intelligent Computing and Intelligent Systems (2009)

The Design of Active Feedback Controllers for the Generalized Projective Synchronization of Hyperchaotic Qi and Hyperchaotic Lorenz Systems

Sundarapandian Vaidyanathan[1] and Sarasu Pakiriswamy[2]

[1] Research and Development Centre, Vel Tech Dr. RR & Dr. SR Technical University
Avadi-Alamathi Road, Avadi, Chennai-600 062, India
sundarvtu@gmail.com
http://www.vel-tech.org/

[2] Dept. of Computer Science and Engineering, Vel Tech Dr. RR & Dr. SR Technical University
Avadi-Alamathi Road, Avadi, Chennai-600 062, India
sarasujivat@gmail.com
http://www.vel-tech.org/

Abstract. This paper investigates the design of active feedback controllers for achieving generalized projective synchronization (GPS) of hyperchaotic systems *viz.* identical hyperchaotic Qi systems (Chen, Yang, Qi and Yuan, 2007), and non-identical hyperchaotic Lorenz system (Jia, 2007) and hyperchaotic Qi system. Lyapunov stability theory has been used to establish the synchronization results (GPS) derived in this paper using active feedback control. Since the Lyapunov exponents are not required for these calculations, the active feedback control method is very effective and convenient for achieving the general projective synchronization (GPS) of hyperchaotic Qi and hyperchaotic Lorenz systems. Numerical simulations are presented to demonstrate the effectiveness of the synchronization results derived in this paper.

Keywords: Active control, hyperchaos, feedback control, generalized projective synchronization, hyperchaotic Qi system, hyperchaotic Lorenz system.

1 Introduction

Chaotic systems are nonlinear dynamical systems, which are highly sensitive to initial conditions. The hyperchaotic systems have at least two positive Lyapunov exponents and exhibit more complex behavior than common chaotic systems. Studies of synchronization of hyperchaotic systems are distinctively interesting and challenging works.

In most of the chaos synchronization approaches, the *master-slave* or *drive-response* formalism is used. If a particular chaotic system is called the *master* or *drive system* and another chaotic system is called the *slave* or *response system*, then the idea of synchronization is to use the output of the master system to control the slave system so that the output of the slave system tracks the output of the master system asymptotically.

The seminal work by Pecora and Carroll ([2], 1990) is followed by a variety of impressive approaches for chaos synchronization such as the sampled-data feedback

N. Chaki and A. Cortesi (Eds.): CISIM 2011, CCIS 245, pp. 231–238, 2011.

synchronization method [3], OGY method [4], time-delay feedback method [5], back-stepping method [6], active control method [7], adaptive control method [8], sliding control method [9], etc.

In generalized projective synchronization [10], the chaotic systems can synchronize up to a constant scaling matrix. Complete synchronization [2], anti-synchronization [11], hybrid synchronization [12] and projective synchronization [13] are special cases of generalized projective synchronization.

This paper addresses the design of active feedback controllers for the generalized projective synchronization (GPS) of identical hyperchaotic Qi systems (Chen, Yang, Qi and Yuan, [14], 2007) and non-identical hyperchaotic Lorenz system (Jia, [15], 2007) and hyperchaotic Qi system (2007).

This paper is organized as follows. In Section 2, we provide a description of the hyperchaotic systems studied in this paper. In Section 3, we derive results for the GPS between identical hyperchaotic Qi systems (2007). In Section 4, we derive results for the GPS between non-identical hyperchaotic Lorenz system (2007) and hyperchaotic Qi system (2007). In Section 5, we summarize the main results obtained in this paper.

2 Systems Description

The hyperchaotic Qi system ([14], 2007) is described by the dynamics

$$\begin{aligned} \dot{x}_1 &= a(x_2 - x_1) + \epsilon x_2 x_3 \\ \dot{x}_2 &= c x_1 - d x_1 x_3 + x_2 + x_4 \\ \dot{x}_3 &= x_1 x_2 - b x_3 \\ \dot{x}_4 &= -f x_2 \end{aligned} \tag{1}$$

where x_1, x_2, x_3, x_4 are the *state* variables and a, b, c, d, ϵ, f are constant, positive parameters of the system.

The system (1) is hyperchaotic when the system parameter values are chosen as $a = 35,\ b = 4.9,\ c = 25,\ d = 5,\ \epsilon = 35$ and $f = 22$.

Figure 1 depicts the state orbits of the hyperchaotic Qi system (1).

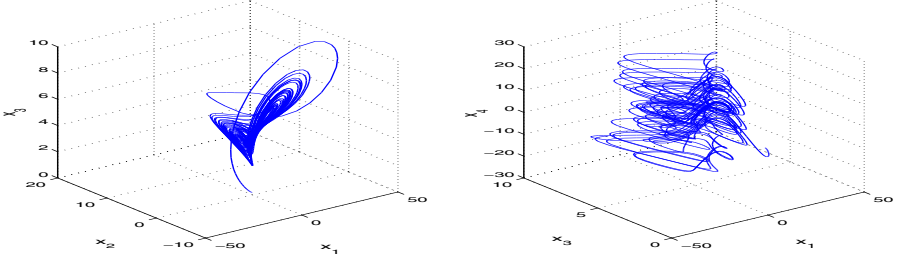


Fig. 1. State Orbits of the hyperchaotic Qi System

The hyperchaotic Lorenz system ([15], 2007) is described by the dynamics

$$\begin{aligned}\dot{x}_1 &= p(x_2 - x_1) + x_4 \\ \dot{x}_2 &= -x_1 x_3 + r x_1 - x_2 \\ \dot{x}_3 &= x_1 x_2 - q x_3 \\ \dot{x}_4 &= -x_1 x_3 + s x_4\end{aligned} \tag{2}$$

where x_1, x_2, x_3, x_4 are the *state* variables and p, q, r, s are constant, positive parameters of the system.

The system (2) is hyperchaotic when the system parameter values are chosen as $p = 10,\ q = 8/3,\ r = 28$ and $s = 1.3$.

Figure 2 depicts the state orbits of the hyperchaotic Lorenz system (2).

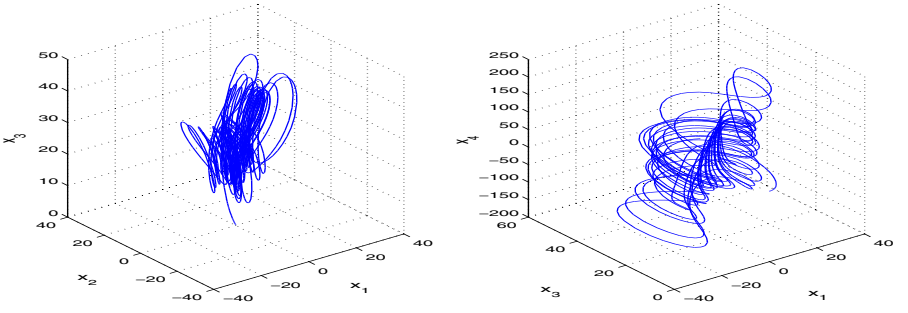


Fig. 2. State Orbits of the hyperchaotic Lorenz System

3 Generalized Projective Synchronization of Identical Hyperchaotic Qi Systems

3.1 Main Results

In this section, we derive results for the active feedback control design for achieving generalized projective synchronization (GPS) of identical hyperchaotic Qi systems ([14], 2007).

Thus, the master system is described by the hyperchaotic Qi dynamics

$$\begin{aligned}\dot{x}_1 &= a(x_2 - x_1) + \epsilon x_2 x_3 \\ \dot{x}_2 &= c x_1 - d x_1 x_3 + x_2 + x_4 \\ \dot{x}_3 &= x_1 x_2 - b x_3 \\ \dot{x}_4 &= -f x_2\end{aligned} \tag{3}$$

where x_1, x_2, x_3, x_4 are the *state* variables and a, b, c, d, ϵ, f are constant, positive parameters of the system.

Also, the slave system is described by the controlled hyperchaotic Qi dynamics

$$\begin{aligned}\dot{y}_1 &= a(y_2 - y_1) + \epsilon y_2 y_3 + u_1\\ \dot{y}_2 &= cy_1 - dy_1 y_3 + y_2 + y_4 + u_2\\ \dot{y}_3 &= y_1 y_2 - by_3 + u_3\\ \dot{y}_4 &= -fy_2 + u_4\end{aligned} \tag{4}$$

where y_1, y_2, y_3, y_4 are the *state* variables and u_1, u_2, u_3, u_4 are the active controls.

For the GPS of (3) and (4), the synchronization errors are defined as

$$e_i = y_i - \alpha_i x_i, \quad (i = 1, 2, 3, 4) \tag{5}$$

where the scales $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ are real numbers.

A simple calculation yields the error dynamics

$$\begin{aligned}\dot{e}_1 &= a(y_2 - y_1) + \epsilon y_2 y_3 - \alpha_1[a(x_2 - x_1) + \epsilon x_2 x_3] + u_1\\ \dot{e}_2 &= cy_1 - dy_1 y_3 + y_2 + y_4 - \alpha_2[cx_1 - dx_1 x_3 + x_2 + x_4] + u_2\\ \dot{e}_3 &= y_1 y_2 - by_3 - \alpha_3[x_1 x_2 - bx_3] + u_3\\ \dot{e}_4 &= -fy_2 - \alpha_4[-fx_2] + u_4\end{aligned} \tag{6}$$

We consider the active nonlinear controller defined by

$$\begin{aligned}u_1 &= -a(y_2 - y_1) - \epsilon y_2 y_3 + \alpha_1[a(x_2 - x_1) + \epsilon x_2 x_3] - k_1 e_1\\ u_2 &= -cy_1 + dy_1 y_3 - y_2 - y_4 + \alpha_2[cx_1 - dx_1 x_3 + x_2 + x_4] - k_2 e_2\\ u_3 &= -y_1 y_2 + by_3 + \alpha_3[x_1 x_2 - bx_3] - k_3 e_3\\ u_4 &= fy_2 + \alpha_4[-fx_2] - k_4 e_4\end{aligned} \tag{7}$$

where the gains k_1, k_2, k_3, k_4 are positive constants.

Substitution of (7) into (6) yields the closed-loop error dynamics

$$\dot{e}_i = -k_i e_i, \quad (i = 1, 2, 3, 4) \tag{8}$$

We consider the quadratic Lyapunov function defined by

$$V(e) = \frac{1}{2} e^T e = \frac{1}{2} \left(e_1^2 + e_2^2 + e_3^2 + e_4^2\right) \tag{9}$$

Differentiating (9) along the trajectories of the system (8), we get

$$\dot{V}(e) = -k_1 e_1^2 - k_2 e_2^2 - k_3 e_3^2 - k_4 e_4^2 \tag{10}$$

which is a negative definite function on $\mathbb{R}^4$, since k_1, k_2, k_3, k_4 are positive constants.

Thus, by Lyapunov stability theory [16], the error dynamics (8) is globally exponentially stable. Hence, we obtain the following result.

Theorem 1. *The active feedback controller (7) achieves global chaos generalized projective synchronization (GPS) between the identical hyperchaotic Qi systems (3) and (4).* ■

3.2 Numerical Results

For the numerical simulations, the fourth order Runge-Kutta method is used to solve the two systems of differential equations (3) and (4) with the active controller (7).

The parameters of the identical hyperchaotic Qi systems are chosen as

$$a = 35,\ b = 4.9,\ c = 25,\ d = 5,\ \epsilon = 35,\ \ f = 22.$$

The initial values for the master system (3) are taken as

$$x_1(0) = 12,\ \ x_2(0) = 3,\ \ x_3(0) = 17,\ \ x_4(0) = 20$$

The initial values for the slave system (4) are taken as

$$y_1(0) = 5,\ \ y_2(0) = 28,\ \ y_3(0) = 9,\ \ y_4(0) = 6$$

The GPS scales α_i are taken as

$$\alpha_1 = -3.18,\ \ \alpha_2 = 2.65,\ \ \alpha_3 = 3.42,\ \ \alpha_4 = -6.29$$

We take the state feedback gains as $k_1 = 5,\ \ k_2 = 5,\ \ k_3 = 5$ and $k_4 = 5$.

Figure 3 shows the time response of the error states e_1, e_2, e_3, e_4 of the error dynamical system (6) when the active nonlinear controller (7) is deployed. From this figure, it is clear that all the error states decay to zero exponentially in 1.5 sec.

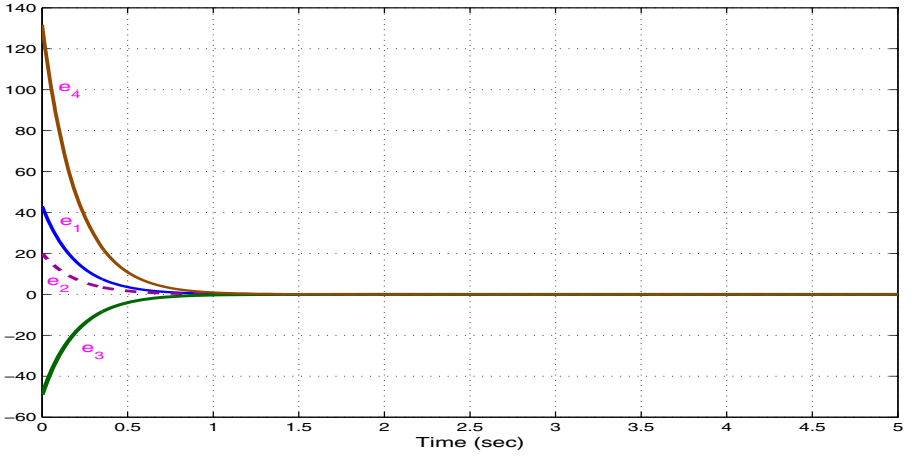


Fig. 3. Time Responses of the Error States of (6)

4 Generalized Projective Synchronization of Non-identical Hyperchaotic Lorenz and Hyperchaotic Qi Systems

4.1 Main Results

In this section, we derive results for the generalized projective synchronization (GPS) of the hyperchaotic Lorenz system ([15], 2007) and hyperchaotic Qi system ([14], 2007).

Thus, the master system is described by the hyperchaotic Lorenz dynamics

$$\begin{aligned}\dot{x}_1 &= p(x_2 - x_1) + x_4\\ \dot{x}_2 &= -x_1x_3 + rx_1 - x_2\\ \dot{x}_3 &= x_1x_2 - qx_3\\ \dot{x}_4 &= -x_1x_3 + sx_4\end{aligned} \tag{11}$$

where x_1, x_2, x_3, x_4 are the *state* variables and p, q, r, s are constant, positive parameters of the system.

Also, the slave system is described by the controlled hyperchaotic Qi dynamics

$$\begin{aligned}\dot{y}_1 &= a(y_2 - y_1) + \epsilon y_2y_3 + u_1\\ \dot{y}_2 &= cy_1 - dy_1y_3 + y_2 + y_4 + u_2\\ \dot{y}_3 &= y_1y_2 - by_3 + u_3\\ \dot{y}_4 &= -fy_2 + u_4\end{aligned} \tag{12}$$

where y_1, y_2, y_3, y_4 are the *state* variables and u_1, u_2, u_3, u_4 are the active controls.

For the GPS of (11) and (12), the synchronization errors are defined as

$$e_i = y_i - \alpha_i x_i, \quad (i = 1, 2, 3, 4) \tag{13}$$

where the scales $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ are real numbers.

A simple calculation yields the error dynamics

$$\begin{aligned}\dot{e}_1 &= a(y_2 - y_1) + \epsilon y_2y_3 - \alpha_1\left[p(x_2 - x_1) + x_4\right] + u_1\\ \dot{e}_2 &= cy_1 - dy_1y_3 + y_2 + y_4 - \alpha_2\left[-x_1x_3 + rx_1 - x_2\right] + u_2\\ \dot{e}_3 &= y_1y_2 - by_3 - \alpha_3\left[x_1x_2 - qx_3\right] + u_3\\ \dot{e}_4 &= -fy_2 - \alpha_4\left[-x_1x_3 + sx_4\right] + u_4\end{aligned} \tag{14}$$

We consider the active nonlinear controller defined by

$$\begin{aligned}u_1 &= -a(y_2 - y_1) - \epsilon y_2y_3 + \alpha_1\left[p(x_2 - x_1) + x_4\right] - k_1e_1\\ u_2 &= -cy_1 + dy_1y_3 + \alpha_2\left[-x_1x_3 + rx_1 - x_2\right] - k_2e_2\\ u_3 &= -y_1y_2 + by_3 + \alpha_3\left[x_1x_2 - qx_3\right] - k_3e_3\\ u_4 &= fy_2 + \alpha_4\left[-x_1x_3 + sx_4\right] - k_4e_4\end{aligned} \tag{15}$$

where the gains k_1, k_2, k_3, k_4 are positive constants.

Substitution of (15) into (14) yields the closed-loop error dynamics

$$\dot{e}_i = -k_i e_i, \quad (i = 1, 2, 3, 4) \tag{16}$$

We consider the quadratic Lyapunov function defined by

$$V(e) = \frac{1}{2} e^T e = \frac{1}{2} \left(e_1^2 + e_2^2 + e_3^2 + e_4^2\right) \tag{17}$$

Differentiating (17) along the trajectories of the system (16), we get

$$\dot{V}(e) = -k_1 e_1^2 - k_2 e_2^2 - k_3 e_3^2 - k_4 e_4^2 \tag{18}$$

which is a negative definite function on $\mathbb{R}^4$, since k_1, k_2, k_3, k_4 are positive constants.

Thus, by Lyapunov stability theory [16], the error dynamics (16) is globally exponentially stable. Hence, we obtain the following result.

Theorem 2. *The active feedback controller (15) achieves global chaos generalized projective synchronization (GPS) between the non-identical hyperchaotic Lorenz system (11) and the hyperchaotic Qi system (12).* ■

4.2 Numerical Results

For the numerical simulations, the fourth order Runge-Kutta method is used to solve the two systems of differential equations (11) and (12) with the active controller (15).

The parameters of the hyperchaotic Lorenz system (11) and hyperchaotic Qi system (12) are taken so that the systems are hyperchaotic (see Section 2).

The initial values for the master system (11) are taken as

$$x_1(0) = 23, \; x_2(0) = 12, \; x_3(0) = 8, \; x_4(0) = 5$$

The initial values for the slave system (12) are taken as

$$y_1(0) = 10, \; y_2(0) = 4, \; y_3(0) = 16, \; y_4(0) = 15$$

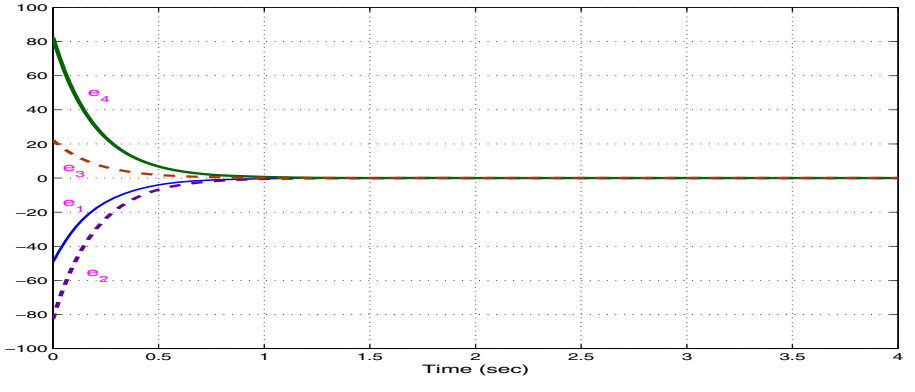


Fig. 4. Time Responses of the Error States of (14)

The GPS scales α_i are taken as

$$\alpha_1 = 2.58, \ \ \alpha_2 = 7.23, \ \ \alpha_3 = -8.32, \ \ \alpha_4 = -1.46$$

We take the state feedback gains as $k_1 = 5, \ k_2 = 5, \ k_3 = 5$ and $k_4 = 5$.

Figure 4 shows the time response of the error states e_1, e_2, e_3, e_4 of the error dynamical system (14) when the active nonlinear controller (15) is deployed. From this figure, it is clear that all the error states decay to zero exponentially in 1.5 sec.

5 Conclusions

In this paper, active feedback controllers have been derived so as to achieve generalized projective synchronization (GPS) of identical hyperchaotic Qi systems (2007), and non-identical hyperchaotic Lorenz system (2007) and hyperchaotic Qi system (2007). The synchronization results derived in this paper have been proved using Lyapunov stability theory. Numerical simulations are presented to demonstrate the effectiveness of the synchronization results derived in this paper.

References

1. Liao, T.I., Tsai: Adaptive synchronization of chaotic systems and its application to secure communications. Chaos, Solitons and Fractals 11, 1387–1396 (2000)
2. Pecora, L.M., Carroll, T.L.: Synchronization in chaotic systems. Phys. Rev. Lett. 64, 821–824 (1990)
3. Yang, T., Chua, L.O.: Control of chaos using sampled-data feedback control. Internat. J. Bifurcat. Chaos 9, 215–219 (1999)
4. Ott, E., Grebogi, C., Yorke, J.A.: Controlling chaos. Phys. Rev. Lett. 64, 1196–1199 (1990)
5. Park, J.H., Kwon, O.M.: A novel criterion for delayed feedback control of time-delay chaotic systems. Chaos, Solit. Fract. 17, 709–716 (2003)
6. Yu, Y.G., Zhang, S.C.: Adaptive backstepping synchronization of uncertain chaotic systems. Chaos, Solit. Fract. 27, 1369–1375 (2006)
7. Ho, M.C., Hung, Y.C.: Synchronization of two different chaotic systems using generalized active control. Physics Letters A 301, 424–428 (2002)
8. Chen, S.H., Lü, J.: Synchronization of an uncertain unified system via adaptive control. Chaos, Solitons and Fractals 14, 643–647 (2002)
9. Konishi, K., Hirai, M., Kokame, H.: Sliding mode control for a class of chaotic systems. Phys. Lett. A 245, 511–517 (1998)
10. Zhou, P., Kuang, F., Cheng, Y.M.: Generalized projective synchronization for fractional order chaotic systems. Chinese Journal of Physics 48(1), 49–56 (2010)
11. Sundarapandian, V.: Anti-synchronization of Lorenz and T chaotic systems by active nonlinear control. Internat. J. Computer Information Systems 2(4), 6–10 (2011)
12. Sundarapandian, V.: Hybrid synchronization of hyperchaotic Rössler and hyperchaotic Lorenz systems by active control. Internat. J. Advances in Science and Technology 2(4), 1–10 (2011)
13. Mainieri, R., Rehacek, J.: Projective synchronization in three-dimensioned chaotic systems. Phys. Rev. Lett. 82, 3042–3045 (1999)
14. Chen, Z., Yang, Y., Qi, G., Yuan, Z.: A novel hyperchaos system with only one equilibrium. Phy. Letters A 360, 696–701 (2007)
15. Jia, Q.: Hyperchaos generated from the Lorenz chaotic system and its control. Phys. Letters A 366, 217–222 (2007)
16. Hahn, W.: The Stability of Motion. Springer, New York (1967)

Parallel Hybrid SOM Learning on High Dimensional Sparse Data

Lukáš Vojáček[1], Jan Martinovič[1], Jiří Dvorský[1],
Kateřina Slaninová[1,2], and Ivo Vondrák[1]

[1] Department of Computer Science,
VŠB – Technical University of Ostrava, 17. listopadu 15,
708 33 Ostrava, Czech Republic
{lukas.vojacek,jan.martinovic,jiri.dvorsky,ivo.vondrak}@vsb.cz
[2] Department of Informatics
SBA, Silesian University in Opava
Karviná, Czech Republic
slaninova@opf.slu.cz

Abstract. Self organizing maps (also called Kohonen maps) are known for their capability of projecting high-dimensional space into lower dimensions. There are commonly discussed problems like rapidly increased computational complexity or specific similarity representation in the high-dimensional space. In the paper there is proposed the effective clustering algorithm based on self organizing map with the main purpose to reduce high dimension of the input dataset. The problem of computational complexity is solved using parallelization; the speed of proposed algorithm is accelerated using the algorithm version suitable for data collections with certain level of sparsity.

1 Introduction

Recently, the problem of high-dimensional data clustering is arising together with the development of the information and communication technologies which supports growing opportunities to process the large data collections. High-dimensional data sets are commonly available in the areas like medicine, biology, informational retrieval, web analysis, social network analysis, image processing, financial transaction analysis and others.

With increasing data dimensionality, there are commonly discussed two main challenges. The first is the dimensionality, which rapidly increases the computational complexity with respect to the number of dimension. Therefore, this problem makes some common algorithms computationally impracticable in many real applications. The second challenge is the specific similarity representation in the high-dimensional space. In [2] Beyer et al. presents, that for any point in the high-dimensional space, the expected distance computed by Euclidean measure to the closest and to the farthest point shrinks with the growing dimensionality. This can be the reason for the decreasing effectiveness of common clustering algorithms in many datamining tasks.

N. Chaki and A. Cortesi (Eds.): CISIM 2011, CCIS 245, pp. 239–246, 2011.

The authors propose the effective clustering algorithm exploiting the features of neural networks, especially Self organizing maps (SOM), for the reduction of data dimensionality. The problem of computational complexity is solved using the parallelization of the basic SOM algorithm; the speed of proposed algorithm is accelerated using the algorithm version suitable for data collections with certain level of sparsity.

2 Self Organizing Maps

Self organizing map (SOM), also called Kohonen map, is a type of artificial neural network invented by professor Teuvo Kohonen in 1982 [9]. The input space of the training samples is represented in a low dimensional (often two-dimensional) space, called map. The model is capable of projecting a high-dimensional space to a lower-dimensional space [12] and is efficient in structure visualization due to its feature of topological preservation using a neighborhood function. Obtained low-dimensional map is often used for pattern detection, clustering, or for characterization and analysis of the input space. SOM technique has been applied in many spheres like speech recognition [15,4], image classification [8,1], document clustering [7,5] etc. Detailed description of SOM application is provided in [4].

There are known several variants of the SOM algorithm interpretations [10,13]. Depending up to the implementation we can use serial, or parallel algorithms. As conventional variant of the serial algorithm interpretations can be considered standard On-line SOM.

On-line SOM Algorithm is the conventional method, where the weight vectors $\boldsymbol{w}_k(t)$ are updated during the training phase recursively for each input vector $\boldsymbol{x}(t)$. The winning neuron d_c is commonly selected by calculating similarity using Euclidean distance:

$$d_k(t) = \|\boldsymbol{x}(t) - \boldsymbol{w}_k(t)\|, \tag{1}$$

$$d_c(t) \equiv \min_k d_k(t). \tag{2}$$

The weight vectors are then updated using learning-rate factor $\sigma(t)$ and the neighborhood function $h_{ck}(t)$:

$$\boldsymbol{w}_k(t+1) = \boldsymbol{w}_k(t) + \sigma(t) h_{ck}(t)[\boldsymbol{x}(t) - \boldsymbol{w}_k(t)], \tag{3}$$

or using cosine measure, which can be more suitable for large sparse data sets with high dimensions [16].

The learning-rate factor $\sigma(t)$ is used for correction of the weight vectors; during the learning phase its value is reduced. The concrete updated weight vectors $\boldsymbol{w}_k(t)$ are set by the neighbor function $h_{ck}(t)$, which determines the distance between nodes c and k. The distance is typically decreasing during the learning phase, from an initial value (often comparable to the dimension/or the half of dimension of the lattice) to the value equal to one neuron (one node in the lattice

of neurons forming the SOM), commonly is used the standard Gaussian neighborhood function. For the serial online SOM algorithm were published several variants to improve its computational efficiency; as an example we can mention WEBSOM [11].

2.1 Modified Calculation of Euclidean Distance

During the competitive learning phase, where for each input vector $\boldsymbol{x}(t)$ is computed similarity to all weight vectors $\boldsymbol{w}_k(t)$, and where the best matching unit (BMU) is founded, we have notified the problem with high computation complexity while working with higher dimensions. Instead of standard method for calculation of similarity with Euclidean distance, see Eq. (4), we have used its modified version using multiplication, see Eq. (5).

$$d_k(t) = \sqrt{\sum_{i=1}^{n}(x_i - w_i)^2}, \tag{4}$$

where n is the dimension of input vector $\boldsymbol{x}$, and $\boldsymbol{w}$ is the neuron's weight.

$$d_k(t) = \sqrt{\sum_{i=1}^{n} x_i^2 - 2x_i w_i + w_i^2} \tag{5}$$

Due to this modification we can calculate x_i^2 and w_i^2 at the beginning, and then during the computation of Euclidean distance we can compute only $2x_i w_i$. Finally, in the actualization phase, we must recalculate only the part w_i^2, because only this value is changing during the weight actualization.

3 Parallel SOM Algorithms

Till lately, most of the conventional algorithms were designed as sequential. The sequential algorithms were well suited to the past generation of computers, which basically performed the operations in the sequential fashion. With the development of the parallel computation, where the operations were performed simultaneously, there is growing the necessity to redesign the serial algorithms to their parallel implementations.

The parallelization of SOM learning algorithms can be implemented by the network partitioning. The *network partitioning* is the implementation, where the neural network is partitioned among the processors. Then, each input data sample is processed by its assigned processor or the parallel task. The network partitioning was implemented by several authors [6,18].

The main principle of our parallel implementation is based on division of the neural network into the parts. This division is shown in the Fig. 1, where we have the map of 3×4 nodes. Neurons in this map are gradually assigned to

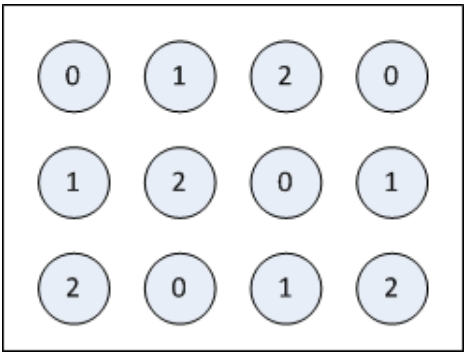


Fig. 1. SOM Map Division

three processes – cycle. There is not always the possibility to divide the map into identical parts i.e. parts with the same number of neurons. In these cases, the neural network is divided the principle, where the parts of the network differ in at the most one neuron. During the parallel learning process, each part is matched with the individual process, and at best, a one process corresponds to one physical processor.

Generally, we have two main phases of the SOM algorithm:

Finding BMU - this phase is very fast due to the map division into multiple processes and due to the fact that we are working with sparse data. The modified computation of Euclid distance Eq. (5) is used during learning the SOM. Each process finds its own BMU in its part of the map; this node is then compared with other BMU obtained by other processes. The information about the BMU of the whole network is then transmitted to all the processes to perform updates of the BMU neighborhood.

Weight actualization - Weight vector of neurons in the BMU neighborhood are updated in this phase. The updating is done also in parallel manner. Each process can effectively detect if some of its neurons belong to BMU neighborhood. If so, the selected neurons are updated.

4 Hybrid Learning

During the learning process, the standard approach is, that there are changed all the neuron weights. In our case, where we are working with sparse data collection, there can be changed only the weight values which are different from zero[14]; this is performed using cosine measure. However, this method has two main problems: (1) there can occur a case, where multiple different input patterns can be matched with the same neuron, even though they are dissimilar. This can be caused by the actualization of only the appropriate part of the weight. (2) Another problem occurs with making too favorable only the part of weights, which are actualized. Described problems bring certain level of error, which can be eliminated using two approaches:

– Standard actualization in periodic epochs, while one epoch is passing through the all training set.
– Standard actualization only if any of weight component exceeds the limit value (threshold).

In our experiments we have used sparse data collection of 8707 input vector dimension; the data collection consisted on 1033 records. We have set the SOM map of 25×25 neurons, for learning phase 50 epochs were processed. The threshold for weight actualization to 0.3 is chosen in hybrid learning algorithm. Several methods of weight actualization were tested e.g. Euclidean distance in the first and the last epoch, or every even epoch etc.

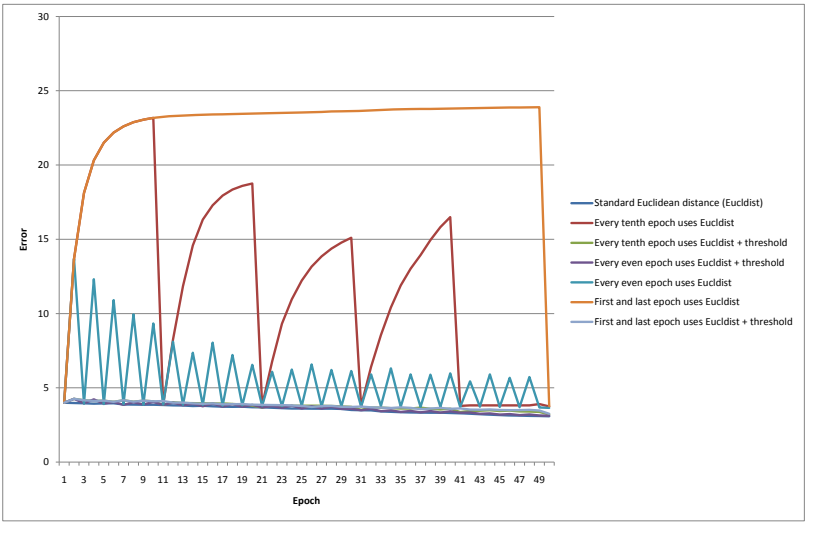


Fig. 2. Error Evolution for Different Approaches with Different Weight Actualization

Trends of error through all epochs in different SOM learning experiments are shown in Fig. 2. The main goal of these optimizations was to accelerate the SOM learning phase, while not making the final error worse than it can be achieved with standard approach.

It is evident that the weight actualization in the SOM using only cosine measure entails expected increasing inaccuracy of the result.

But if we set specific threshold, for which is processed typical weight actualization using Euclidean distance, we can reach the acceleration and the error improvement at the same time. The best results were obtained while the threshold was set to 0.3 level, but this value depends on the dimension of the training set.

The computational times are shown in the Table 1. As we can see, our approach leads to the time acceleration while the worsening of the error is relatively small with comparison to the typical SOM algorithm. The time values were reached by the parallel calculations executed on 48 processors.

Table 1. Final Error and Total Time for Different Experiments

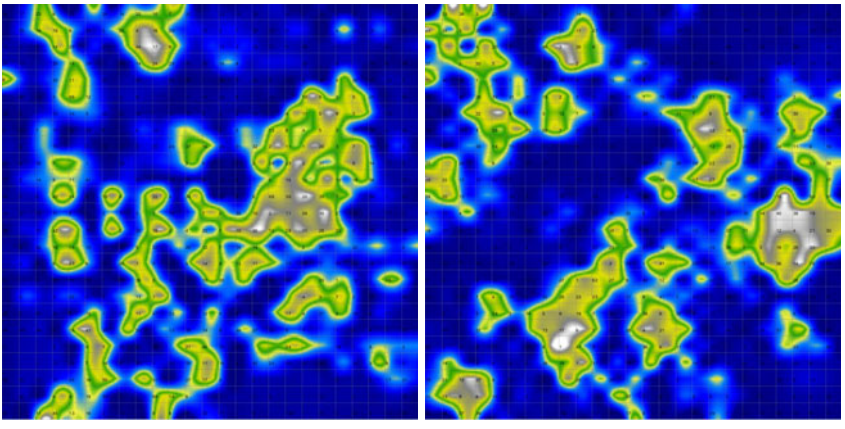
Usage of Euclidean distance	Threshold	Error	Time (sec)
All epochs (standard algorithm)		3.0875	185.649
First and last epoch		3.7578	7.351
First and last epoch	yes	3.2512	18.002
Every tenth epoch		3.7523	7.835
Every tenth epoch	yes	3.2124	23.128
Every even epoch		3.6535	14.962
Every even epoch	yes	3.1200	53.747

The last part of our experiment was oriented to scalability of mentioned approach. The results of used algorithm are shown in Table 2. As we can see, the acceleration is dependent not only on the optimization for the data sparsity, but neither on the selected map dimension and on the dimension of the input data set. Our proposed solution is more suitable for the high-dimensional sparse datasets width large amount of input records. Another important factor is map dimension. In our approach it is better utilized the processor power while working with higher map dimensions. This finding is very positive for us; it enables us to generate more detailed results. With more processors we are able to select higher dimensions of SOM map for more detailed computations.

Table 2. Computational Time of Parallel Solution with Respect to Number of Processors and Map Size

Map Size	Computational Time (sec)				
	#1	#8	#16	#32	#48
25×25	95	29	9	8	25
50×50	338	70	30	23	16
100×100	1194	284	143	66	46

Examples of final SOM maps from the standard 3(a) and hybrid 3(b) algorithm, which represents obtained clusters after the learning phase, are presented in the Fig. 3 For the visualization was used the Smoothed Data Histograms[3] of SOM maps.



(a) Standard approach (b) Hybrid Approach

Fig. 3. Smoothed Data Histograms of SOM Maps

All the experiments were performed on Windows HPC server 2008 with 6 computing nodes, where each node has 8 processors with 12 GB of memory.

5 Conclusion

In this paper we presented the parallel implementation of the SOM neural network algorithm. Parallel implementation was tested on HPC cluster containing 6 nodes and 40 processor cores. The achieved speed-up was very good.

The training phase was speed up using several methods namely: utilization of cosine measure for sparse data, division of the SOM map into several smaller parts and consequent parallel computation of BMU in each part, and modification of Euclidean distance computation. The final errors achieved by our method were almost identical to standard SOM learning algorithm, which computation time is several times greater.

In the future work we intend to focus on hybrid learning methods, where division of neurons are controlled by Fiedler vector[17].

We intent to use the application of this method in various scopes of our research, like finding behavioral patterns of users, business intelligence sphere, or analysis of e-learning systems.

Acknowledgment. This work is supported by the grant of Grant Agency of Czech Republic No. 205/09/1079, and by the grant of Silesian University, Czech Republic No. SGS/24/2010 - The Usage of BI and BPM Systems to Efficiency Management Support.

References

1. Bekel, H., Heidemann, G., Ritter, H.: Interactive image data labeling using self-organizing maps in an augmented reality scenario. Neural Networks 18(5-6), 566–574 (2005)
2. Beyer, K., Goldstein, J., Ramakrishnan, R., Shaft, U.: When Is "Nearest Neighbor" Meaningful? In: Beeri, C., Bruneman, P. (eds.) ICDT 1999. LNCS, vol. 1540, pp. 217–235. Springer, Heidelberg (1998)
3. Pampalk, E., Rauber, A., Merkl, D.: Using Smoothed Data Histograms for Cluster Visualization in Self-Organizing Maps. In: Dorronsoro, J.R. (ed.) ICANN 2002. LNCS, vol. 2415, pp. 871–876. Springer, Heidelberg (2002)
4. Gas, B., Chetouani, M., Zarader, J.-L., Charbuillet, C.: Predictive Kohonen Map for Speech Features Extraction. In: Duch, W., Kacprzyk, J., Oja, E., Zadrożny, S. (eds.) ICANN 2005. LNCS, vol. 3697, pp. 793–798. Springer, Heidelberg (2005)
5. Georgakis, A., Li, H.: An ensemble of som networks for document organization and retrieval. In: Proceedings of AKRR 2005, International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning, pp. 141–147 (2005)
6. Gropp, W., Lusk, E., Skjellum, A.: Using MPI: portable parallel programming with the message-passing inferace. MIT Press (1999)
7. Kishida, K.: Techniques of document clustering: A review. Library and Information Science 35(1), 106–120 (2005)
8. Kohonen, O., Jaaskelainen, T., Hauta-Kasari, M., Parkkinen, J., Miyazawa, K.: Organizing spectral image database using self-organizing maps. Journal of Imaging Science and Technology 49(4), 431–441 (2005)
9. Kohonen, T.: Self-Organization and Associative Memory, 3rd edn. Springer Series in Information Sciences, vol. 8. Springer, Heidelberg (1984) (1989)
10. Kohonen, T.: Things you haven't heard about the self-organizing map. In: Proc. ICNN 1993, International Conference on Neural Networks, Piscataway, NJ, pp. 1147–1156. IEEE, IEEE Service Center (1993)
11. Kohonen, T.: Exploration of very large databases by self-organizing maps. In: Proceedings of ICNN 1997, International Conference on Neural Networks, PL1–PL6. IEEE Service Center, Piscataway (1997)
12. Kohonen, T.: Self Organizing Maps, 3rd edn. Springer, Heidelberg (2001)
13. Lawrence, R.D., Almasi, G.S., Rushmeier, H.E.: A scalable parallel algorithm for self-organizing maps with applications to sparse data mining problems. Data Mining and Knowledge Discovery 3, 171–195 (1999)
14. Machon-Gonzalez, I., Lopez-Garcia, H., Calvo-Rolle, J.L.: A hybrid batch SOM-NG algorithm. In: The 2010 International Joint Conference on Neural Networks, IJCNN (2010)
15. Meenakshisundaram, S., Woo, W.L., Dlay, S.S.: Generalization issues in multiclass classification - new framework using mixture of experts. Wseas Transactions on Information-Science and Applications 4, 1676–1681 (2004)
16. Pullwitt, D.: Integrating contextual information to enhance som-based text document clustering. Neural Networks 15, 1099–1106 (2002)
17. Vojáček, L., Martinovič, J., Slaninová, K., Dráždilová, P., Dvorský, J.: Combined method for effective clustering based on parallel som and spectral clustering. In: Proceedings of Dateso 2011. CEUR Workshop Proceedings, vol. 706, pp. 120–131 (2011)
18. Wu, C.-H., Hodges, R.E., Wang, C.J.: Parallelizing the self-organizing feature map on multiprocessor systems. Parallel Computing 17(6-7), 821–832 (1991)

Operational Algebra for a Graph Based Semantic Web Data Model

Abhijit Sanyal[1] and Sankhayan Choudhury[2]

[1] IBM India Pvt. Ltd, Kolkata, India
[2] University of Calcutta, Kolkata, India
abhijit.sanyal@in.ibm.com, sankhayan@gmail.com

Abstract. An operational algebra with formal definition is necessary for any conceptual data model for smooth data retrieval. We have already proposed a graph-based aspect-oriented data model for a web based application, called 'Semantic Graph Data Model' (SGDM) [7]. The proposed model aims to represent the semantic within data in a web based application in an efficient way. SGDM should be supported by an operational algebra for retrieval of data. In this paper, we have proposed a graph based algebra that consists of semantic and non-semantic operators for the evaluation of a semantic query within SGDM. The proposed operators of SGDM are being implemented in a way such that the corresponding object relational operators are being executed on the equivalent object relational data model to offer a doable solution for the end users.

Keywords: Operational Algebra, Semantic Web Data Model, Graph Data Model, Query Processing.

1 Introduction

Recently the information published in Internet has become increasingly user and context-dependent to meet the demand of tailored content access, presentation and functionality based on user's location, device, personal preferences and needs. In order to realize such adaptive behavior, appropriate conceptual data model is needed for efficient capturing of inherent semantic within an application. Moreover, the conceptual level design must be supported by an operational algebra that offers formal mathematical basis of data retrieval. In last one decade, few semantic Web data models have been proposed with their own formalization techniques to cater this functionality.

In 2002 [1], an RDF algebra called RAL, is for defining (and comparing) the semantics of different RDF query languages and for performing algebraic optimizations. The operators defined in RAL are of mainly three types: extraction operators that retrieve the needed resources from the input RDF model, loop operators that support repetition and construction operators that build the resulting RDF model.

In 2005 [2], a conceptual model called Abstraction Layer Model (ALM), proposes a query algebra for large scale data-intensive web information system (WIS). This provides a general algebraic approach consisting of underlying type system and

N. Chaki and A. Cortesi (Eds.): CISIM 2011, CCIS 245, pp. 247–256, 2011.

join-operation. The query algebra is extended to rational tree types that can handle Uniform Resource Locator (URL). This model applies Kleene algebras with tests for showing the personalization mechanism of a WIS to preferences and intension of users.

In 2006 [3], an algebra is proposed for querying over Extensible Markup Language (XML) documents present in Web. In this conceptual model an XML document is represented as a rooted, ordered, labeled tree. A tree is composed by a set of vertices, or elements, connected with arcs. The operators of this algebra can be unary or binary. Unary operator takes in a forest and return a forest, their general form is $\alpha_P(F)$, where α is the operator, P is a predicate and F is the input forest. Binary operators take in two forests and return a forest; their general form is $F\alpha_P G$. The algebra is *closed*: all the operators take in forest(s) and return a forest. Consequently the operators can be composed with each other. This algebra is equipped with union and difference operators, which are quite similar to their relational counterparts. Union takes in two forests and returns a new forest composed by the trees contained in the two input forests; difference takes in two forests and returns a sub forest of the first input forest, composed by those trees which are not included in the second input forest. This algebra is inspired by relational algebra but it doesn't provide any mapping to relational algebra.

In 2007 [4], a concept algebra based Web knowledge discovery engine is proposed for on-line knowledge discovery. In this work, a visualized concept network explorer and a semantic analyzer is developed to locate, capture, and refine queries based on concept algebra. A graphical interface is built using concept and semantic models to refine users' query structures. This tool kit described here can generate a structured XML query package that accurately express users' information needs for on-line searching and knowledge acquisition.

In 2008 [5], a formal descriptive mechanism is proposed for RDF. It demonstrates the formalization of relational calculus and logic inference for RDF data model with two examples. This paper shows that RDF is a formal specification method which is a binary relation. The formal architecture of RDF includes two primary parts, i.e., formal language and inference mechanism. The properties of this formalized system is demonstrated both in the level of syntax and semantics to ascertain the correctness of the formal architecture.

Traditional search engines for shallow web cannot index and query hidden web databases since the pages are dynamically created in response to query forms. The lack of formal foundations and query support made it difficult to develop search engines for hidden web. In 2009 [6], an algebraic language, called Integra, as a foundation for another SQL-like query language called BioFlow, for the integration of Life Sciences data on the hidden Web. The algebra presented here adopts the view that the web forms can be treated as user defined functions and the response they generate from the back end databases can be considered as traditional relations or tables. It can extend the traditional relational algebra to include integration primitives such as schema matching, wrappers, form submission, and object identification as a family of database functions. These functions are then incorporated into the traditional relational algebra operators to extend them in the direction of semantic data integration. To support the well known concepts of horizontal and vertical integration, two new operators, called link and combine are proposed.

We have proposed a graph-based aspect-oriented conceptual web data model, called 'Semantic Graph Data Model' (SGDM) [7] for capturing and representing the semantic data in a web based application. It can handle concerns of the semantic relationship present in applications with effective domain knowledge reuse in the modeling phase itself. But there is no formalization technique is defined for SGDM to develop semantic query. So, in this paper we propose a novel Operational Algebra that will be executed on SGDM for querying data. But to provide a doable solution to the end-users, in the back end the equivalent queries in object relational algebra are formed and being executed on the corresponding object relational equivalent of SGDM representation for generating intended output. This entire process is being implemented in form of software.

2 Overview of SGDM Model with a Real World Example

Semantic Graph Data Model (SGDM) [7] is an aspect-oriented conceptual graph data model for handling semantic data in web application. The web data model can also handle hypertext and semi-structured data [8] efficiently through multi-layered structure with minimum two-tier architecture. The first tier (T1) provides the object-oriented conceptual model for semi-structured data allowing the complex user defined data type and functions. T1 is not layered and the second tier (T2) is built on top of T1. The layers are present from T2 onwards and they are nothing but higher level abstractions made on the constructs present in the lower layer. T2 depicts the top most view of the model in the aspect-oriented semantic data along with hypertext. The semantic relationship amongst them is also described here. The model constructs present in T1 and T2 in SGDM are listed in Table1 and Table 2.

The aspect-orientation can be implemented in any domain model using SGDM by identifying major crosscutting concerns present in business. We propose a construct called 'Aspect Node' (AN) in the Tier-2 (T2) or its higher level abstractions in SGDM to capture the different concerns involved in a business decision. The relevant concerns - the aspect nodes involved in a business decision can be identified separately and the resultant sub graph can be used for efficient query processing through the removal of the irrelevant items from the search space. This enormously reduces the searching effort for semantic based information from Web based Information System (WIS) resulting time-cost advantage. The most important feature of semantic web is the process of reusing the domain knowledge. This can be applied for developing the domain model for another specific business problem in the similar domain. Web Ontology Language (OWL) explicitly represents the meaning of terms in vocabularies and further specifies the relationships between them in any Semantic Data Model. But here we do not need a query language as ANs are powerful enough to hold the semantics with a particular concern and semantic queries will be formed using drag-and-drop feature, without taking help from developer. So SGDM is sufficient enough to represent semantics and retrieve information based on semantic queries without using any specific OWL kind of language rather using simple graph operation.

Table 1. SGDM constructs for T1

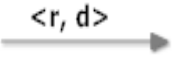
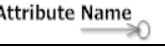
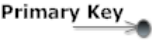
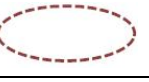
SGDM T1 Constructs	Graphical Notation
Class Node (CN)	
Object Node (ON)	
Broken Directed Edge without Label (Instance creation of CN)	
Directed Edge with Label (Association between two CNs)	
Directed Edge without Label (Inheritance between two CNs)	
Attributes	
Determinant	

Table 2. SGDM constructs for T2 or higher level

SGDM T2 Constructs	Graphical Notation
Aspect Link (connection between related CN and AN)	
Aspect Node (AN)	
Broken Directed Thick Edge with Label (Association between ANs)	Label
Broken Thick Edge without Label (Connection between AN and its associated semantic Group)	
Semantic Group (SG) (Group of individual elements for AN, based on a particular concern of CN)	
Constraints	
Page	
Page Link (Connection between CN and Page)	

Let us consider a semantic web based information system "e-retail system" for illustrating the data model SGDM. Tier 1 and Tier2 representation are depicted here through figure 1 and figure 2 respectively. The creation of aspect nodes considering concerns of a customer are being done through proper analysis of the submitted queries in requirement analysis phase. Let us consider few semantic queries in Table 3 that are considered as resources in our proposed Resource description framework [7].

Table 3. Sample Semantic Queries based on e-retail System

No	Semantic Query Statement
1	Choice of products for mediocre customers
2	Trend of purchasing of the young customers
3	Choice of Stores by rich customers
4	Product satisfaction of mediocre customers
5	Age group of maximum Byers

In Table3, Query1, 'mediocre customers' is subject, 'Choice of products' is object and the order placed by the customer is considered as predicate. Accordingly three ANs are created in T2 of e-retail system – $A_{mediocre\_customers}$, A_{orders} and $A_{choice\_of\_products}$. The broken directed edges with labels – 'Orders_placed_by_Mediocre_Customers' and 'Choice_of_Products_in_Orders_placed' represent the relationship between the ANs $A_{mediocre\_customers}$, A_{orders} and A_{orders}, $A_{choice\_of\_products}$ respectively. The sub-graph formed by these three ANs, helps to find and represent the preferable products of the mediocre customers efficiently without complex graph-query operations in less time as search space is enormously reduced.

In Table3, Query2 "Trend of purchasing of the young customers" can be formed in a similar way by using the ANs $A_{young\_customers}$, A_{orders} and $A_{choice\_of\_products}$. In this case 'trend of purchasing' is semantically similar to 'choice_of_products' of previous query1, so the same AN $A_{choice\_of\_products}$ is used for handling aspect 'trend of purchasing' also.

All the different types of card holders have been considered under the AN $A_{young\_customers}$ for finding the customers based on their age as mentioned in the membership database of e-retail system. In this case same ANs A_{orders} and $A_{choice\_of\_products}$ have been effectively re-used for the creation of the sub-graph which provides the example of domain knowledge re-use. It reduces the development cost for preparing different sales reports as per business requirement.

3 Operational Algebra for SGDM

In this section, a set of operators are proposed for SGDM that enables the user to retrieve and manipulate the semantic data. These operators and their functions constitute the Operational Algebra for SGDM. SGDM domain model is essentially a graph consisting of various types of nodes and edges.

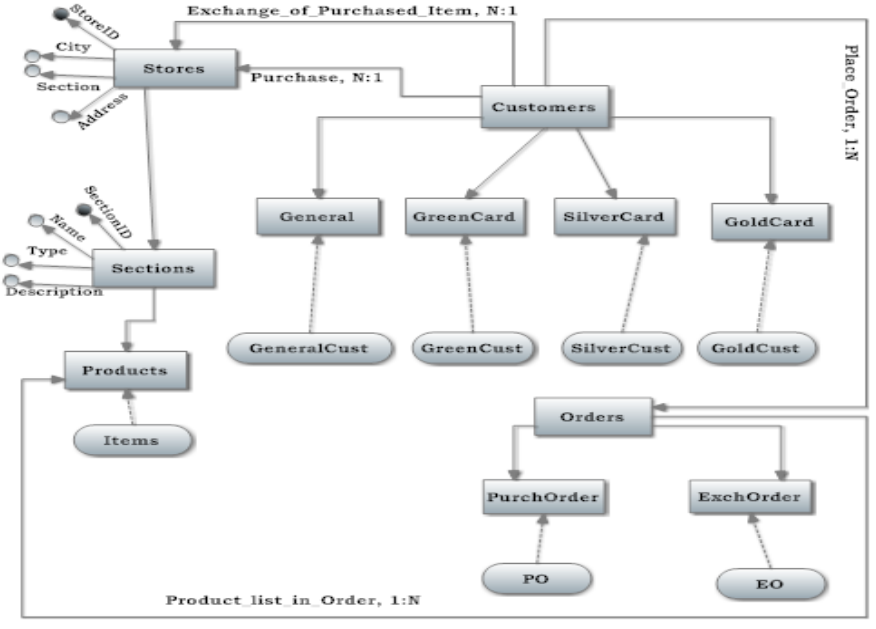


Fig. 1. SGDM diagram for "e-retail system" domain model: Tier-1 representation

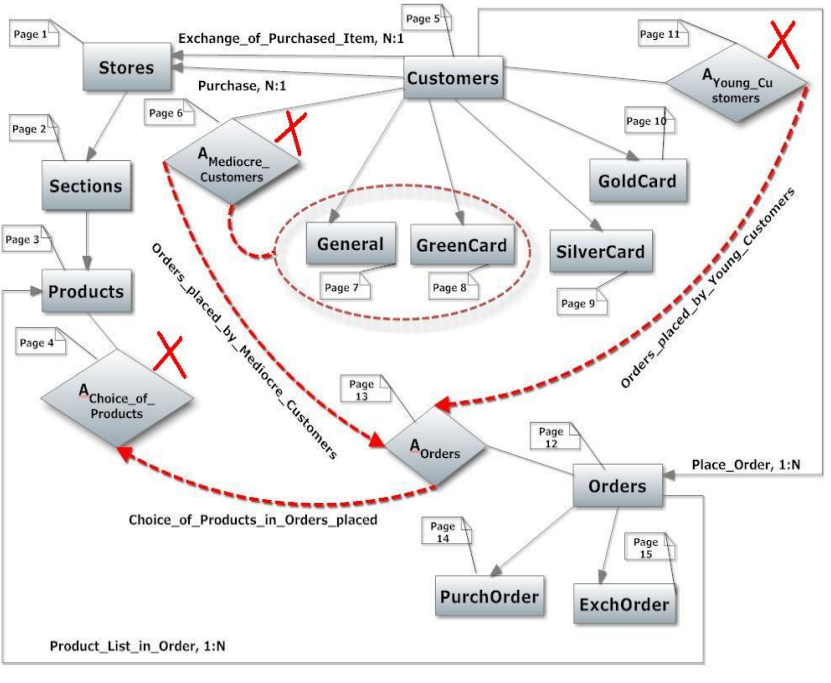


Fig. 2. Figure2. SGDM diagram for "e-retail system" domain model: Tier-2 over Tier-1 Representation

Mainly there are two types of operators in SGDM - Semantic and Basic. Semantic operators work on top of the basic operators. Semantic operators are applied on AN where as basic operators are applied on CNs only. Any semantic query is designed using semantic operators but the general operations are performed by the Basic Operators in the underlying layer to get the non-semantic granular information. The final result of any semantic query is a sub graph of SGDM domain model according to the closure property of graph theory.

3.1 Definition of SGDM Operators

Semantic Operators

Retrieve operator (Ω): The Retrieve Operator (Ω) is a unary semantic operator applied on AN, that extracts the CN with which the AN is linked in SGDM domain model. The algebraic notation of this operator is $\Omega\ (A_S) = C_o$, where C_o is the CN, linked with the operand AN, A_S in SGDM domain model. There could be multiple number of AN $A_{S1} \ldots A_{Sn}$ based on the different concerns that results same CN.

e.g. To find the CN for 'young customers' and 'mediocre customers' the retrieve operator can be used as,

$\Omega\ (A_{young\_customers}) = C_{customers}$ and $\Omega\ (A_{mediocre\_customers}) = C_{customers}$

Association Operator (μ): The Association Operator (μ) is a binary semantic operator that performs association between two ANs. The algebraic notation of this operator is $\mu_X (A_1, A_2) \Rightarrow G_s$, where x is label of the edge between two ANs - A_1, A_2 and G_s is the resultant graph with nodes and edges that is generated from the association of the ANs.

e.g. To find all the orders placed by mediocre customers, the association operator is applied on ANs $A_{mediocre\_customers}$ and A_{Orders} to get the resultant sub-graph G_{s1}, as

$\mu_{orders\_placed\_by\_mediocre\_customers} (A_{mediocre\_customers}, A_{Orders}) \Rightarrow G_{s1}$

Again, the details of the products purchased in every order can be found from the association of two relevant adjacent ANs as –

$\mu_{choice\_of\_products\_in\_orders\_placed} (A_{choice\_of\_products\_in\_orders\_placed}, A_{Orders}) \Rightarrow G_{s2}$

Semantic Join Operator (ψ): The Semantic Join Operator (ψ) is n-nary operator, used to combine more than one sub-graphs generated after applying semantic association operator on the concerned ANs. There should be a common AN between two operand graphs based on which the semantic join is being done. The result of the operation is also a graph generated through graph union of the participating graphs i.e. consisting of all the nodes and edges of the sub graphs on which the operator ψ is applied. The algebraic notation of this operator is –

$G_{s1}\ \psi_{\Theta 1}\ G_{s2}\ \psi_{\Theta 2}\ G_{s3} \ldots \psi_{\Theta n}\ G_{sn} = G_Z$

where $G_{s1\ldots n}$ are semantic graphs generated as a result of the semantic association operator, Θ_1, $\Theta_2 \ldots \Theta_n$ are the respective common vertices or ANs of the operand graphs and G_Z is the ultimate resultant graph that generated after the applying join operator ψ.

e.g. The result of query1 "Choice of products for mediocre customers" can be found joining two sub-graphs G_{s1} and G_{s2} by using semantic join operator as –

$G_{choice\_of\_products\_for\_mediocre\_customers} = G_{s1}\ \psi A_{Orders}\ G_{s2}$

where A_{Orders} is the common AN between two sub-graphs G_{s1} and G_{s2}.

Non-Semantic Basic Operators

gSelect Operator (Є): The gSelect Operator (Є) is used to select the lower level CNs from their higher level CN, based on a particular concern present in the AN linked with that operand CN. Generally, the gSelect operator (Є) is preceded by retrieve operator. The algebraic notation of this operator is $Є_{CON}\ (C_h) = C_{li}$ where C_{li} (i=1…n) is the set of CNs that are extracted based on the concern CON hold by the AN linked with higher level CN C_h. If $C_h = C_l$, i.e. AN is attached with lowest level CN then gSelect operator is not used after retrieve operator.

e.g. To select the mediocre customers from a retail store, gSelect operator will be applied as –

$Є_{purchasing\_type\ =\ medium}\ (CN_{customers}) = \{C_{General}, C_{GreenCard}\}$

gProject Operator (Δ): The gProject Operator (Δ) is a unary operator that is applied only on CN in SGDM, to get the required granular data based on a specific attribute or list of attributes of that CN. If $attr_1, attr_2, \ldots, attr_n$ be the specified attribute list to be projected from a CN C_m , the algebraic expression will be like –

$\Delta\ (attr_1, attr_2, \ldots, attr_n)\ C_m \Rightarrow O_{mi}$,

where Δ is the gProject Operator and O_{mi} could be a list of ONs under the CN C_m (where, i = 1...n) for the attributes $attr_1, attr_2, \ldots, attr_n$.

e.g. If we are interested to find the all individual young customers (i.e. $O_{young\_customers}$) from $CN_{customers}$, we need to execute the projection operation on as –

$\Delta\ (age = 18\text{-}28)\ (CN_{customers})$.

4 Illustration of Semantic Query Evaluation through SGDM Operators

In this section the processing of a semantic query is being explained through an example. As discussed in section2, the result of query2 in Table3 will be a sub-graph consisting of three ANs $A_{young\_customers}$, A_{orders} , $A_{choice\_of\_products}$ and broken directed edges with labels – 'Orders_placed_by_Young_Customers' and 'Choice_of_Products_in_Orders_placed' as shown in Fig2. This semantic query can be evaluated in the following way using SGDM Operational Algebra –

Step1. Semantic Retrieve (Ω) Operator is applied on ANs $A_{young\_customers}$, A_{orders} and $A_{choice\_of\_products}$ to get their linked CNs in SGDM –

$\Omega\ (A_{young\_customers}) = CN_{customers}$

$\Omega\ (A_{choice\_of\_products}) = CN_{Products}$

$\Omega\ (A_{orders}) = CN_{orders}$

Step2. gProject Operator is applied on $CN_{customers}$ to find all individual young customers where 'young' is considered as the age group 18 to 28 years. The condition for treating a customer as young is known through requirement analysis phase.

$$\Delta_{(age = 18\text{-}28)} (CN_{customers}) = O_{young\_customer}$$

Step3. The semantic activity between ANs $A_{choice\_of\_products}$, $A_{young\_customers}$ i.e. the orders placed by young customers is constituted by the third AN A_{orders}. The details of the orders placed by young customers and the product details of the already ordered transaction (by all customers) are found by applying Association (μ) Operator in the following way:

$$\mu_{orders\_placed\_by\_young\_customers} (A_{young\_customers}, A_{Orders}) \Rightarrow G_{s1}$$

Here the G_{S1} sub graph is supposed to hold the information about the ordered details by young customer group.

$$\mu_{choice\_of\_products\_in\_orders\_placed} (A_{choice\_of\_products}, A_{Orders}) \Rightarrow G_{s2}$$

Here the sub graph G_{S2} holds the information of the product details present in already processed order.

Step4. The semantic join operation is executed on the resultant graphs G_{s1} and G_{s2} based on common AN A_{Orders} to get the final graph that provides the result of the semantic query2,

$$G_{trend\_of\_purchasing\_of\_young\_customers} = G_{s1} \psi A_{Orders} G_{s2}$$

The semantic operators are being executed on the SGDM, but to provide a doable solution the relational algebra based operations are performed on the equivalent object relational data model representation of e-retail system. The equivalent data model presentation rules are described in detail in [8]. As a result the graph reflecting the e-retail scenario can be transformed to the collection of some tables like order, product, customer etc. This concept is crystallized in form of software through an interpreter development in GME tool. The equivalent relational algebra expressions for the above mentioned query (query 2 in Table 3) will be like this:

$$Young\_cust = \sigma_{Cust\text{-}type = Young} (Customer);$$

$$Temp\_table1 = young\_cust \bowtie Order; \text{ (based on cust\_id)}$$

$$Temp\_table2 = product \bowtie order; \text{ (based on product\_id)}$$

$$Result\_table = Temp\_table1 \bowtie Temp\_table2; \text{ (based on order\_id)}$$

5 Conclusion

In this paper an attempt has been made to propose a novel Operational Algebra on SGDM for execution of semantic query. Set of semantic and non-semantic operators are proposed, illustrated and implemented. Thus our proposal offers an integrated

solution that consists of efficient data modeling of a semantic web based application through SGDM and the semantic query evaluation through graph based algebra. A systematic approach is also proposed here to map the proposed SGDM Operational Algebra to standard Relational Algebra for rendering a doable solution. Our future work will be proposal of a Visual Query Language that can be used by an end-user and in the back end graph algebra will be responsible for query processing.

References

1. Frasincar, F., Houben, G.-J., Vdovjak, R., Barna, P.: RAL: an Algebra for Querying RDF. In: Proceedings of the 3rd International Conference on Web Information Systems Engineering (WISE 2002). IEEE (2002)
2. Schewe, K.-D., Thalheim, B.: Conceptual modelling of web information systems. Journal of Data & Knowledge Engineering 54(2) (August 2005)
3. Buratti, G., Montesi, D.: A Data Model and an Algebra for Querying XML Documents. In: Proceedings of the 17th International Conference on Database and Expert Systems Applications (DEXA 2006). IEEE (2006)
4. Hu, K., Wang, Y.: A Web Knowledge Discovery Engine Based on Concept Algebra. In: Proceeding of the IEEE 2007 International Conference on Electrical and Computer Engineering, Canada (2007)
5. Qian, Z.: Formal Descriptive Mechanism of the Ontology Language RDF. In: Proceedings of the 2008 International Conference on Management of e-Commerce and e-Government. IEEE Computer Society, Washington, DC (2008)
6. Hosain, S., Jamil, H., Hosain: An Algebraic Language for Semantic Data Integration on the Hidden Web. In: Proceeding of the 2009 IEEE International Conference on Semantic Computing, pp. 237–244 (2009)
7. Sanyal, A., Choudhury, S.: An Aspect-oriented Conceptual Level Design for Semantic Web based Application. In: Accepted in 2nd International IEEE Conference on Computer & Communication Technology (ICCCT-2011), MNNIT, Allahabad, India, September 15-17 (2011)
8. Sanyal, A., Sanyal, S., Choudhury, S.: Physical Level Implementation of a Web Data Model. In: ACM International Conference on Communication, Computing & Security (ICCCS 2011), NIT Rourkela, India, February 12-14 (2011) Published in ACM with ISBN-978-1-4503-0464-1

A Survey on the Semi-Structured Data Models

Supriya Chakraborty[1] and Nabendu Chaki[2]

[1] Department of Computer Application, JIS College of Engineering, Paschimbanga, India
supriya.k6@gmail.com
[2] Department of Computer Science, Calcutta University, Paschimbanga, India
nabendu@ieee.org

Abstract. Semi-structured Databases (SSD) are becoming extremely popular in versatile applications including interactive web application, protein structure analysis, 3D object representation, personal lifetime information management. The list is endless. In order to meet the challenges of today's complex applications, a generic SSD model is in demand. Many works have been reported on this. In this paper, expectations from a generic SSD model are studied by a critical survey among existing models.

Keywords: Semi structured data, Generic data model, OEM, BDFS, ORA-SS.

1 Introduction

In the twentieth century, organizations are overwhelmed by Structured (SD), Semi-Structured (SSD) and Unstructured Data (UD). Organizations are suffered integrated information retrieval from SD, SSD and UD. Lack of uniformity, different standards/formats for representation/visualization, dissimilar models, diversified interfaces of users, and human interaction retrieve missing, vague, and incorrect information. Thus a uniform view of different data is necessary with minimum human interaction. As existing data models are dedicated for types of data. A generic model that could organize all types of data is very necessary for uniform view. Due to the varied organization, expressive power and flexibility of SSD, it is assumed that different types of data models are congregated into semi-structured data model. As logical integration is possible among many types of data but it restricts flexibility, adaptability and robustness of applications. Thus generic model needs to belong on the conceptual level as such semantics of the data could be exhausted as per functionality, operations or applications. Many environmental operations, coercion on data type, format etc also need to be supported on the generic model. The access of data in the generic model need to be involved with linguistic supports (LS) and artificial intelligence (AI). As fifth generation language would be comprised of LS and AI but desired orientation of data is not in the literature to support both. A generic model of SSD with desired orientation for LS and AI may meet challenges of today's complex applications. The orientation of data is the major challenge for proposing generic model of SSD. The orientation of SSD for generic model is primarily depends on the expectations from that model.

N. Chaki and A. Cortesi (Eds.): CISIM 2011, CCIS 245, pp. 257–266, 2011.

Our investigation is revealed with the fact that many models and integration techniques have been proposed for representing SSD; but each one has its own purpose. Thus generic nature of SSD is in vain to be represented. Identification of expectations of features of the generic model is very challenging task. The expectations are exposed by the comparative study among SSD models.

This paper is organized as follows. It commence with the brief summary of the work, following the introduction section. The comparative study among models is shown in section 2 and its subsections that include tables, charts, graphs etc. Finally expectations from generic model are summarized following conclusion and future work. This paper ends with the reference.

2 Data Models of Semi-Structured Data

A brief summary of each model is presented mentioning its novel idea, advantages, limitations and points on which we do not agree in successive paragraphs.

OEM (Object Exchanges Data Model) is used to represent SSD described in [1] and [2]. Two primary features of OEM are i) Data is self-describing and ii) Flexible enough to accommodate loosely typed schema. Few limitations of the OEM are i) how multiple objects from the root are maintained in the run time? This problem is always associated in the data structure. A general tree might have been used in this regard as per our intuition ii) new property which is not applicable for existing records or objects can easily be incorporated because of the self describing feature or no schema definition. But existence of the new property can not be confirmed before run time of the query. That is costly to accommodate both in logical and physical level at run time. iii) Specialty (special attribute for few objects) cannot be retrieved except brute force search. iv) and redundancy is not controlled.

ORA-SS (An Object Relationship-Attribute Model) proposed the formal semantic model of SSD in [3] and [4]. Designing concepts of SD was expressed in ORA-SS for SSD. The concept of reference to reduce the redundancy was introduced that automatically alleviates the anomaly problems. Mapping between ORA-SS and other model has also been worked out to show the flexibility, acceptability and transformability of the model. Both attribute which "refers to" and "referenced attribute" whether will be resided in the same page or same block that lead to reduced performance for retrieval of correlated data are not discussed. ORA-SS is simply designing model that also included Instance Diagram, but still no formal repository structure of the model were really discussed.

Wrapper (TSIMMIS Mediator and Wrapper) provides access to heterogeneous information sources by converting application queries into source specific queries and also does all necessities in [5], [6]. Mediator creates a view from different data sources. The wrapper data is fed into the Mediator using Mediator Specific Language (MSL). Limitations or points in which we do not agree on the above two concepts are 1) The details of Wrapper are very necessary when the data is unstructured. 2) Different system has its own constraints like memory may start either from top or down. No emphasis is given on this point. 3) Why object oriented logic based language MSL and LOREL are chosen? 4) This scheme demands additional hardware, and separate view management.

CO (Content Oriented) Data Model ([7]) separates content and type of SSD. Type of data is mentioned only once. It improves access time and reduces storage requirement for SSD. This model does not propose any inbuilt facility for collection. Discussion on structure of collection (hierarchical, dictionary or linear linked list) was completely ignored. A complete example that represents a situation and steps of query processing are expected from [7]. The point on which we do not agree is that how entities will form the database. Clarification need on "Is SQL or Xquery work on CO data model or how query processing will work on CO data model?"

As BDFS (Basic Data Model for Semi-structured) Data Model ([8], [9]) is a formal and elegant data model, where labels of edges in the schemas are formulae of a certain theory τ. Few limitations of BDFS model are such that 1) there is no possibility of specifying additional constraints, such as existence of edges or bounds number of nodes emanating from the node, 2) the possibility to deal with incomplete and vague information was ruled out, and 3) no semantic information was embedded. Point in which we do not agree or need further clarification is that conformance algorithm of database to schema.

The extension model of BDFS contributes to extend BDFS schema with constraints, assumption of incomplete theory on τ, and proposed very powerful language, called ALCQ. Till date technical gap exist to implement features of ALCQ.

DGDM (Data Graph Data Model) ([10]) is a simple logical model flexible enough, parametric, and very expressive to manipulate internal and external SD and SSD. Frequently used operators have been clearly defined and expressive power over Xquery was compared with examples. Few limitations of this model is that it does not deal with constraints in SSD, could not show explicit relationship between XML schema standards, elimination of duplicate data, speed up data manipulation, relationship with XQuery and other languages. Dimension of ignorance on Data Graph Data Model is discussed in [11]. Different type of ignorance similar to Incompleteness, Imprecision, Non-Specificity, Vagueness, Uncertainty, and Absence along with their inter dependencies was defined and exemplified on SSD.

XML (Extended Markup Language) is a markup language designed for data exchange, discussed in [12]. DTD and XML schema have been proposed for defining schema. XML also provide a mechanism to create link between elements described in [13]. There are several ways to extract information from XML documents. None of them are considered as universal. One of the well known methodologies is SAX discussed in [14]. SAX provides a read-only and forward-only event driven interface. Another interface is XmlReader which support pull model. More sophisticated navigation is XPathNavigator that enables cursor type navigation powered by XPathexpression [15]. But semi-defined nature of schema of SSD is not properly defined into XML.

As iMeMx represent uniform view, called resource view of heterogeneous data like text, image, audio, and video described in ([16][17]). This advance model is specially designed for personal information management system in a personal computer. This model also bridges the gap between internal and external information of the file system. No distributed feature, semantic search or incompleteness of data was discussed.

The focus of OZONE, discussed in [18] is simply integration of structured object database model (ODMG) and its query language OQL with the semi-structured database model OEM and its query language LORE ([4]). In this approach, SD is the entry point for SSD and vice versa. This scheme was not for storing data. Unified access of SD and SSD was only ensured.

2.1 Abstract Level Comparison

The purpose, persistent storage (PS), allocation of data in different locations (ADL), existence of semantic data, and query languages (QL) are the evaluation indexes for abstract level comparison among all the models shown in the Table 1.

Table 1. Abstract comparison among SSD models

Data Model	No	Purpose	Reference	PS	ADL	Semantic Data	QL
OEM	A	Self describing nature can easily model loosely typed schema.	[1,2]	Y	N	Each entry	LORE
ORA-SS	B	Designing concepts of SD are proposed on SSD. Redundancy is reduced.	[3,4]	N	Y	Label data	xQuery
M & W*	C	Uniform view of heterogenious data sources by logical integration	[5,6]	N	N	Datatype coercion	MSL & LORE
CO Data Model	D	It also separates schema from content but the redundancy is much reduced in this model.	[7]	Y	Y	Linked with RDF & ECMA	Natural language
BDFS	E	Incomplete theories of schema are expressed.	[8,9]	Y	Y	Formula on path	Dl & CL
DGDM**	F	Structures of type tree, graph, hierarchy, relation are expresed and operators are defined.	[10,11]	N	Y	Edge label & order	Extended Relation Algebra.
XML	G	It is tree structure, attribute-value pair tagged within document.	[12,13, 14,15]	Y	Y	Schema	xQuery
iMeMX	H	Unified view of heterogeneous data.	[16,17]	N	N	No	Natural Language
OZONE	I	Integrate ODMG with query OQL to OEM with query LORE	[18]	N	N	No	OQL, LORE

Legend: Dl -Description logic, CL - constraint language, QL – Query language, MC – Model Compatibility

* M&W implies Mediator and Wrapper. ** DGDM implies Data Graph Data Model

2.2 Comparison in Conceptual or Logical Level

Most of the evaluation indexes of conceptual level are self explanatory; often used indexes are explained for easy reference. Disjunction (row 3 of Table 2) means few properties of SSD are mutually exclusive. Coercion of data type (row 5 of Table 2) implies matching of different formats of same data type. For example date value of format dd/mm/yyyy and mm/dd/yyyy yield true if both dates are same. Easy to integrate heterogeneous data (row 12 of Table 2) implies integration among SD, SSD, and UD. Easy to query without knowing data type (row 13 of Table 2) signifies whether the model at the conceptual model proposes any special view that eliminates the outside and inside file information. Restructuring or reordering of data (row 15 of Table 2) implies whether any instance of the model is converted into another structure like balanced binary search tree, or heap for achieving any desired performance goal. Logically remove an element (row 16 of table 2) stands for deletion of data not physically occurred but only a marking to distinguish the data is deleted. It has significance in performance because of very often data deletion need reordering of data that consumes so much overhead.

Table 2. Checklist for Evaluation Index at Conceptual Level

	S	**S**	**D**		**M**	**O**	**D**	**E**	**L**	**S**
Evaluation indexes in Conceptual Level	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	
Repetition of schema data?	Y	Y	N	N	N	N	Y	N	N	
Is it possible to insert a record if schema not matched ?	Y	N	N	Y	Y	Y	N	Y	N	
Is disjunction supported?	N	Y	N	N	Y	N	N	N	N	
Is constraint allowed on the attribute level?	N	Y	N	Y	Y	N	Y	N	N	
Is Coercion of Datatype supported?	N	N	N	N	Y	Y	N	N	N	
Is the model integrated with RDBMS?	Y	Y	Y	N	N	Y	Y	Y	Y	
Is conversion of the model data done into RDBMS?	Y	N	N	N	N	Y	Y	N	Y	
Is conversion of RDBMS data done into model?	Y	Y	Y	Y	Y	Y	Y	Y	Y	
Is redundancy reduced?	N	Y	N	Y	Y	N	N	N	N	
Is the order of the record maintained in the model?	N	N	N	Y	Y	N	N	Y	N	
Is inheritance applicable in the model?	N	Y	N	N	Y	N	N	N	N	
Easy to integrate heterogeneous data?	N	N	N	N	Y	Y	N	Y	N	
Easy to query without knowing data types?	N	N	N	N	Y	N	N	Y	Y	
Makes optimization harder?	Y	N	Y	N	N	N	N	Y	N	
Is Restructuring or reordering of data possible?	N	N	Y	Y	Y	Y	Y	Y	Y	
Is logically remove an element possible?	N	N	Y	N	Y	N	N	N	N	

Y and N imply Yes and No in table 2. A, B, C etc are the name o the model which is numbered in column No of Table 1. For example A stands for OEM, B stands for ORA-SS etc.

2.3 Comparison in Application or Query Level

Query language is always an integrated part of data model for storage and retrieval of data. Computational advantages are expected from organized structure of data by query languages. Here envision is whether expectations are supported from the application level or not. Our understanding summaries on already proposed models on table 3. As above few indexes are briefly explained. Allow query that combine different part of the document (row 9 of Table 3) implies similar to the group by function in relational database. Support for Unit Coercion (row 17 of Table 3) refers to comparing values with different unit of same type of data for example comparing currency values between rupees and dollar. Support ID Creation (row 20 of Table 3) implies creation of unique identification of object. Navigation Support (row 22 of table 3) implies whether data value is navigated by exploring from the application level to physical level without query.

In table 3, Y, N and D stand for Yes, No and Not discussed correspondingly. A, B etc imply same as in table 2.

Table 3. Checklist for Evaluation Indexes in Application or query level

	SSD MODELS							
Evaluation Indexes in Application Level	A	B	C	D	E	F	G	H	I
Query data types and collections of multiple documents/file	Y	Y	Y	Y	Y	Y	Y	Y	Y
Allow data-oriented, document-oriented and mixed queries	N	N	N	Y	Y	Y	N	Y	N
Accept streaming data	N	N	N	Y	Y	Y	Y	Y	N
Support operations of various data models	N	Y	Y	Y	Y	Y	Y	N	Y
Allow conditions/constraints on text elements.	N	N	Y	Y	Y	Y	Y	N	Y
Support for hierarchical and sequence queries.	Y	Y	Y	Y	Y	Y	Y	Y	Y
Manipulate NULL values.	N	N	N	N	Y	Y	Y	N	Y
Support quantifiers (,, and ~)	N	N	N	N	Y	Y	N	N	N
Allow queries that combine different parts of document(s).	N	N	N	N	Y	Y	N	N	N
Support for aggregation	Y	Y	Y	N	Y	Y	Y	Y	Y
Able to generate sorted results.	N	N	Y	N	Y	Y	Y	Y	Y
Support composition of operations.	Y	Y	Y	Y	Y	Y	Y	Y	Y
Reference traversal	N	N	Y	N	N	Y	Y	Y	Y
Able to use environment information (date/time) as part of queries	Y	Y	N	Y	Y	Y	Y	Y	Y
Able to support run time updates	Y	Y	Y	Y	Y	Y	Y	N	Y
Support for type coercion.	N	Y	N	Y	Y	Y	Y	Y	Y
Support for unit coercion	N	N	N	N	Y	N	N	N	N
Preserve the structure of the documents.	N	N	N	N	Y	N	N	N	N
Transform and create XML structures	Y	Y	N	Y	Y	Y	Y	N	Y
Support ID creation.	Y	Y	N	N	N	Y	Y	N	N
Structural recursion.	N	Y	N	N	Y	Y	Y	Y	N
Navigation support	N	N	N	N	N	N	N	N	Y
Allow queries that combine different parts of document(s).	Y	D	N	D	D	D	Y	Y	Y

2.4 Experimental Result

In the implementation level, the ratio of actual and overhead data size, run time resource requirement influences the efficiency of the model. The experiment is done with the schema R (Faculty name, Department, Designation, NCP, ICP, NJ, IJ); where NCP, ICP, NJ and IJ imply type of publications. The dataset is taken from the repository of "JIS College of Engineering". The size of raw data is specified in table 4. Redundancies in paper title exist in the experimental data. Instance sizes conforming to R of each model are specified in Table 5. The implementation is done using jdk 1.6 with IDE netBeans. All implemented models are only prototype implementation for investigation as per our understanding. The result may differ from the original work.

Table 4. Raw Data Size

R	No. of Characters	Kilo Bytes
Faculty	4131	32.27
Publication	23001	179.69
		211.96 (Total)

Table 5. Storage Space and Program Load of Models

Model Name	Instance Size in KB	Program Load
OEM	322.66	Medium
ORA-SS	155.98	Medium
CO Data Model	282.04	Heavy
BDFS with extension	3592.68	Heavy
Data Graph Data Model	3235.34	Heavy
XML	229.69	Simple
iMeMX	447.32	Medium
Wrapper, Mediator	2746.67	Medium
Ozone	1235.32	Heavy

The instance size of ORA-SS is lower comparative to instance sizes of other models as per experimental result. Reduces redundancies and non inclusion of control objects are the reason for this result. The instance size of BDFS and DGDM are very close to each other because both need so many control objects. The instance size of the OZONE is high because both size of source and materialized data are calculated. The M & W instance size gets much higher because of source data instance size, mediator data size and wrapper data size are included.

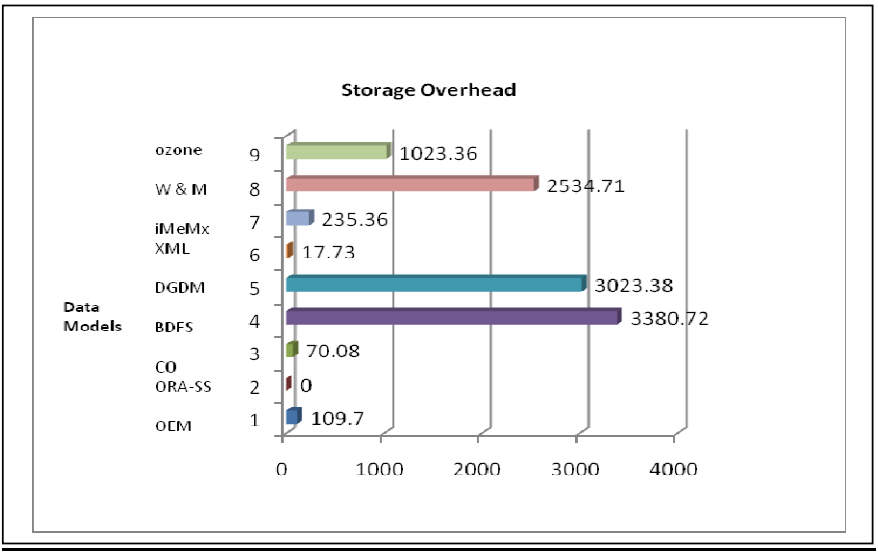


Fig. 1. Overhead comparison of instance space of experimental data

Considering facts and phenomenon our investigation is inferred with the followings:

- A generic semi-structured data model is in demand to meet the challenges of modern complex applications.
- Implementation of domain of SSD would be the solution for uniform view of SD as well as UD. Here domain implies all possible values of a property of an object.

Thus a domain based generic model for representing SSD is the identification of our review work. In the next section, expectations from the generic data model are summarized.

3 Expectation from the Generic Domain of SSD

The generic domain based model of SSD must conform to following characteristics.

- *Different structure of the same data*: The same set of data may be reconciled into different order. For example teachers of the department may be ordered by either designation, or seniority.
- *Domain based representation is non-redundant:* In a domain, no value will be repeated. It ensures the existence of unique value only.
- *Domain based representation inherently maintains referential integrity:* No further constraint need to be imposed.
- *No need to store NULL value:* If any value is either not available, or not applicable, or not defined, it will be treated as non existence within the domain.

- *Domain could be defined and populated by the values prior to the actual record instantiated:* In the classical approach as followed using SQL, first a table is defined and then records are inserted. But in the proposed work, even domain could be populated by the values before actual instantiation of the record.
- *Domain is not implied by the Data type:* Values of the domain may be anything that semantically truthful. Values and operations must be pertinent to the domain as per applications.
- *Format*: The format of data should not be bar for comparison. User of the domain can insert or retrieve the data as per his or her semantic format. For example 11th September 2001 or 9/11/20001 means the same in the domain.
- *Unit*: Countable quantity are measured in different units and sometimes processed in another. For example current balance amount of money of customer account in any currency could be asked at any point of time. The internal transformation from one unit to another unit is highly expected.
- *Complement*: A domain should represent the universal set for the property of the object. Any subset of the domain must be a domain. A complement of the set should be supported.
- *Operations*: Update and query operations should be performed efficiently.
- *Domain constraint*: General constraint could be defined on the domain in the conceptual level. For example age of standard ten examinees must be greater than sixteen.
- *Value constraint*: A particular value may be constrained in the domain. In our real life exception is always occurred. The exception data need to be accommodated or expressed through function.
- *Integration with external data:* Today many models have already proposed for advance applications like semantic models. Domain should be mapped with different models.

4 Conclusions

Logically and conceptually, many attempts were made to model SSD. However, a widely accepted generic model for SSD is yet in demand. This extensive survey aims to analyze the state of art works on semi-structured data representation to identify the expected features of such a generic model. This is going to be the basic objective set towards building generic SSD model. The identified characteristics of SSD formally define access structures including artificial intelligence and linguistic support for the next generation database language as well as uniform view of SD, SSD and UD.

References

1. Papakonstantinou, Y., Garcia-Molina, H., Widom, J.: Object Exchange Across Heterogene-ous Information Sources*. In: 11th International Conference on Data Engineering, Taiwan, pp. 251–260 (1995)

2. McHugh, J., Abiteboul, S., Roy, G., Quass, D., Widom, J.: Lore: A Da-tabase Management System for Semistructured Data. SIGMOD Re-cord, 54–66 (1997)
3. Dobbie, G., Xiaoying, W., Ling, W.T., Lee, L.M.: An object-Relationship –Attribute Model for Semi-Structured Data. Technical Report, School of Computing, Singapore (2000)
4. Dobbie, G., Xiaoying, W., Ling, W.T., Lee, L.M.: Designing Semi-Structured Database using ORA-SS Model. In: 2nd International Con-ference on Web Information Systems Engineering, vol. 1, p. 171. IEEE (2001)
5. Hammer, J., McHugh, J., Garcia-Nolima, H.: Semi-Structured Data: The TSIMMIS Experience. In: 1st East-European Workshop on Ad-vances in Databases and Information Systems, Petersburg, Russia, pp. 1–8 (1997)
6. Garcia-Molina, H., Papakonstantinou, Y., Quass, D., Rajaraman, A., Sajiv, Y., Ullman, J., Vassalos, V., Widom, J.: The TSIMMIS Ap-proach to Mediation: Data Models and Languages. J. of Intelligent In-formation Systems 8(2), 117–132 (1997)
7. Novotny, T.: A Content-Oriented Data Model for Semistructured Data. Dateso, 55–66 (2007)
8. Calvanese, D., Giacomo, D.G., Lenzerini, M.: Extending Semi-Structured Data. In: 4th Asia-Pacific Conference on Conceptual Mod-eling, Australian, vol. 67, pp. 11–14 (2007)
9. Giacomo, D.G., Lenzerini, M.: A uniform Framework for Concept Definitions in Description Logics. J. of Artificial Intelligence Research, 87–110 (1997)
10. Magnani, M., Montesi, D.: A Unified Approach to Structured, Semi-Structured and Unstructured Data. TECHREPORT in Education, In-formation Processing and Management 2, 263–275 (2004)
11. Magnani, M., Montesi, D.: Dimensions of Ignorance in a Semi-Structured Data Model. In: 15th International Workshop on Database and Expert Systems Applications. IEEE (2004)
12. Bray, T., Paoli, J., Sperberg-McQueen, C., Maler, E., Yergeau, F.: Ex-tensible Markup Language (XML)1.0, 4th edn., W3CRecommendation, `http://www.altova.com/specs_xml.html`
13. Document Object Model. W3C Recommendation, `http://www.w3.org`
14. Simple API for XML, `http://www.saxproject.org/`
15. Esposito, D.: Manipulate XML Data Easily with the XPath and XSLT APIs in the .NET. In: MSDN Magazine (2003)
16. Dittrich, P.J., Vaz Salles, A.M.: iDM: A Unified and Versatile Data Model for Personal Dataspace Management. In: Pro. of VLDB 2006, pp. 367–378. ACM, Korea (2006)
17. iMeMx project, `http://www.imemex.org`
18. Lahiri, T., Abiteboul, S., Widom, J.: OZONE: Integrating Structured and Semi-Structured Data. In: 7th Int. Con. Database Programming Languages, Scotland (1999)

Observation-Based Fine Grained Access Control for XML Documents

Raju Halder and Agostino Cortesi

DAIS, Università Ca' Foscari Venezia, Italy
{halder,cortesi}@unive.it

Abstract. The eXtensible Markup Language (XML) is recognized as a simple and universal standard for storing and exchanging information on the web. The risk of unauthorized leakage of this information mandates the use of access control at various levels of granularity. In this paper, we extend to the context of XML documents the notion of Observation-based Fine Grained Access Control (OFGAC) which was originally designed for the relational databases. In this setting, data are made accessible at various levels of abstractions depending on their sensitivity level. Therefore, unauthorized users are not able to infer the exact content of an attribute or element containing partial sensitive information, while they are allowed to get a relaxed view of it, according to their access rights, represented by a specific property.

Keywords: Access Control, XML Documents, Abstract Interpretation.

1 Introduction

With more and more information being exchanged, distributed or published through the web, it is important to ensure that sensitive information is being accessed by the authorized web-users only. Disclosure of sensitive information to unauthorized web-users may cause a huge loss to the enterprizes or organizations. Access control policies and their enforcement [2,7,9] emerged as a most effective solution to ensure safety of the sensitive information in an web information system. The granularity of traditional access control mechanism for XML is coarse-grained and can be applied at file or document level only. As a result, any XML file containing data with both public and private protection requirements will have to be split into two files before applying the access control. However, the need of more flexible business requirements and security policies mandates the use of Fine Grained Access Control (FGAC) mechanisms [1,6,10,12,13] that provide safety of the information in XML documents even at lower level such as individual element or attribute level.

In traditional FGAC, the notion of sensitivity of web-information is too restrictive (either public or private) and impractical in some real systems where intensional leakage of the information to some extent is allowed with the assumption that observational power of external observers is bounded. Thus, we need to weaken or downgrading the sensitivity level of web-information, hence, consider

N. Chaki and A. Cortesi (Eds.): CISIM 2011, CCIS 245, pp. 267–276, 2011.

a weaker attacker model. The weaker attacker model characterizes the observational characteristics of attackers and can be able to observe specific properties of the private data.

To cope with this situation, in our previous work [8], we introduced an Observation-based Fine Grained Access Control (OFGAC) mechanism for Relational Database Management System (RDBMS) based on the Abstract Interpretation framework.

In this paper, we extend this approach to the context of XML documents aiming at providing accessibility of sensitive information at various levels of abstractions depending on their sensitivity level. Unauthorized users are not able to infer the exact content of an attribute or element containing partial sensitive information, while they are allowed to get a relaxed view of it represented by specific property, according to their access rights. The traditional fine grained access control can be seen as a special case of the proposed OFGAC framework.

The structure of the paper is as follows: Section 2 provides a motivating example. Section 3 recalls some basic ideas about the policy specification for XML fine grained access control system and the OFGAC framework for RDBMS. In Section 4, we extend the OFGAC framework to the context of XML documents. Finally, in Section 5, we draw our conclusions.

2 A Motivating Example

Various proposals in support of fine-grained XML access control have been introduced in the literature, including View-based [1,5,6], Non-deterministic Finite Automata (NFA)-based [3,12,13], RDBMS-based [10,11,14] etc.

All the proposals above are binary-based, *i.e.* an access control has only two choices: either "allow" or "forbid", resulting into two extreme views to the XML information: either "public" or "private". Sensitive information are visible to the authorized people only, whereas non-sensitive information are visible to all. However, there are many application areas where some data on the web are treated as partially sensitive and a relaxed view of those data is provided to the users at various levels of sensitivity according to their access rights.

Example 1. Consider an XML document that stores customers' information of a bank. Figure 1(a) and 1(b) represent the Document Type Definition (DTD) and its instance respectively. According to the DTD, the document consists of zero or more "customer" elements with three different child elements: "PersInfo", "AccountInfo", "CreditCardInfo" for each customer. The "CreditCardInfo" for a customer is optional, whereas each customer must have at least one bank account represented by "AccountInfo". The element "PersInfo" keeps the record of personal information for the customers.

Suppose, according to the access control policy, that employees in the bank's customer-care section are not permitted to view the exact content of IBAN and credit-card numbers of the customers, while they are allowed to view only the first two digits of IBAN numbers and the last four digits of credit card numbers, keeping other sensitive digits hidden. For instance, in case of the 12 digits credit

```
<?xml version="1.0"? >
<! DOCTYPE BankCusomers [ >
<! ELEMENT BankCusomers(Customer*) >
<! ELEMENT Customer(PersInfo, AccountInfo+, CreditCardInfo?) >
<! ELEMENT PersInfo(Cid, Name, Address, PhoneNo) >
<! ELEMENT Cid (# PCDATA) >
<! ELEMENT Name (# PCDATA) >
<! ELEMENT Address (street, city, country, pin) >
<! ELEMENT street (# PCDATA) >
<! ELEMENT city (# PCDATA) >
<! ELEMENT country (# PCDATA) >
<! ELEMENT pin (# PCDATA) >
<! ELEMENT PhoneNo (# PCDATA) >
<! ELEMENT AccountInfo (IBAN, type, amount) >
<! ELEMENT IBAN (# PCDATA) >
<! ELEMENT type (# PCDATA) >
<! ELEMENT amount (# PCDATA) >
<! ELEMENT CreditCardInfo (CardNo, ExpiryDate, SecretNo) >
<! ELEMENT CardNo (# PCDATA) >
<! ELEMENT ExpiryDate (# PCDATA) >
<! ELEMENT SecretNo (# PCDATA) >
<! ATTLIST Cid IBAN CDATA #REQUIRED ]>
```

(a) DTD

```
<?xml version="1.0"? >
<BankCusomers>
<Customer>
<PersInfo>
<Cid> 140062 </Cid>
<Name> John Smith </Name>
<Address>
<street> Via Pasini 62 </street>
<city> Venezia </city>
<country> Italy </country>
<pin> 30175 </pin>
</Address>
<PhoneNo> +39 3897745774 </PhoneNo>
</PersInfo>
<AccountInfo>
<IBAN> IT10G 02006 02003 000011115996 </IBAN>
<type> Savings </type>
<amount> 50000 </amount>
</AccountInfo >
<CreditCardInfo>
<CardNo> 4023 4581 8419 7835 </CardNo>
<ExpiryDate> 12/15 </ExpiryDate>
<SecretNo> 165 </SecretNo>
</CreditCardInfo>
</Customer>
</BankCusomers>
```

(b) XML document

Fig. 1. A Document Type Definition (DTD) and its instance

card number "4023 4581 8419 7835" and the IBAN number "IT10G 02006 02003 000011115996", a customer-care personnel is allowed to see them as "**** **** **** 7835" and "IT*** ***** ***** **********" respectively, just to facilitate the searching of credit card number and to redirect the account related issues to the corresponding country (*viz*, "IT" stands for "Italy"). In addition, suppose the policy specifies that the expiry dates and secret numbers of credit cards and the deposited amounts in the accounts are fully-sensitive and completely hidden to them.

The traditional FGAC mechanisms are unable to implement this scenario as the IBAN numbers or credit card numbers are neither private nor public as a whole. To implement traditional FGAC, the only possibility is to split the partial sensitive element into two sub-elements: one with private privilege and other with public. For example, the credit-card numbers can be split into two

sub-elements: one with first 12 digits which is made private and the other with last 4 digits which is made public. However, practically this is not feasible in all cases, as the sensitivity level and the access-privilege of the same element might be different for different users, and the integrity of data is compromised. For instance, when an integer data (say, 10) is partially viewed as an interval (say, [5, 25]), we can not split it.

The Observation-based Fine Grained Access Control (OFGAC) mechanism in [8] provides a solution of such scenario in case of RDBMS, and is based on the Abstract Interpretation framework [4]. We will extend this approach to the context of XML documents, giving rise to partial accessibility of the information on the web.

3 Observation-Based Access Control Policies

In this section, we recall some basic ideas from [4,6,8].

Basis of Fine Grained Access Control Policy Specification for XML. Most of the existing proposals on fine grained access control for XML are based on the basic policy specification introduced by Damiani et al. [6] that specifies the access authorization by a 5-tuple of the form $\langle Subject, Object, Action, Sign, Type\rangle$. The "*Subject*" represents the identifiers or locations of the access requests to be granted or rejected. It is denoted by a 3-tuple $\langle UG, IP, SN\rangle$ where UG, IP and SN are the set of user-groups/user-identifiers, the set of completely-specified/patterns-of IP addresses and the set of completely-specified/patterns-of symbolic names respectively. For instance, ⟨ Physicians, 159.101.*.*, *.hospital.com ⟩ represents a subject belonging to the group physicians, issuing queries from the IP address matching with the pattern 159.101.*.* in the domain matching with symbolic name pattern *.hospital.com. The "*Object*" represents the Uniform Resource Identifier (URI) of the elements or attributes in the documents. The URI is specified by the conditional or unconditional path expressions. The "*Action*" is either "read" or "write" or both being authorized or forbidden. The "$Sign \in \{+, -\}$" is the sign of authorization. Sign "+" indicates "allow access", whereas sign "-" indicates "forbid access". The "*Type*" of the access represents the level of access (DTD level or instance level), whether access is applicable only to the local element or applicable recursively to all sub-elements, hard or soft etc. The priority of the type of accesses from highest to lowest are: LDH (Local Hard Authorization), RDH (Recursive Hard Authorization), L (Local Authorization), R (Recursive Authorization), LD (Local Authorization specified at DTD level), RD (Recursive Authorization specified at DTD level), LS (Local Soft Authorization), RS (Recursive Soft Authorization). Since this specification provides users only two choices in accessing the information: either "allow" or "forbid", we call it *Binary-based FGAC Policy* for XML.

Galois Connection and Abstract Representation of Databases. In general, data contained in any database are *concrete* as they belong to concrete domains of

integers, strings, etc, whereas *abstract* representation of these data are obtained by replacing concrete values by the elements from abstract domains representing specific properties of interests. For instance, addresses of the patients in a "Patient" database can be abstracted by the provinces they belong. Here, province is the abstract representation of all the exact locations that are covered by that province. We may distinguish partial abstract database in contrast to fully abstract one, as in the former case only a subset of the data in the database is abstracted. The values of the attribute x are abstracted by following the Galois Connection $(\wp(D_x^{con}), \alpha_x, \gamma_x, D_x^{abs})$, where $\wp(D_x^{con})$ and D_x^{abs} represent the powerset of concrete domain of x and the abstract domain of x respectively, whereas α_x and γ_x represent the corresponding abstraction and concretization functions (denoted $\alpha_x : \wp(D_x^{con}) \rightarrow D_x^{abs}$ and $\gamma_x : D_x^{abs} \rightarrow \wp(D_x^{con})$) respectively. In particular, partial abstract databases are special case of fully abstract databases where for some attributes x the abstraction and concretization functions are identity functions id, and thus, follow the Galois Connection $(\wp(D_x^{con}), id, id, \wp(D_x^{con}))$.

The Observation-based Fine Grained Access Control policy and its enforcement to RDBMS. In OFGAC framework [8], users are permitted to view the sensitive information at various levels of abstractions according to their authorization level. Highly sensitive information are forbidden completely, while partial-sensitive and non-sensitive information are disclosed in an intermediate form (represented by specific properties according to the access control policies) and in its exact form respectively.

Definition 1 (Observation-based Disclosure Policy). *Given a domain of observable properties D, and an abstraction function $\alpha_D : \wp(val) \rightarrow D$, an observation-based disclosure policy op assigned to the observer O is a tagging that assigns each value v in the database state σ a tag $\alpha_D(v) \in D$, meaning that O is allowed to access the value $\alpha_D(v)$ instead of its actual value v.*

Given an observation-based disclosure policy "*op*", the OFGAC framework for RDBMS consists of the following steps:

- Transform the concrete database into an (partial) abstract database by providing an abstract representation of the sensitive data in the database according to "*op*".
- Convert users' queries into the corresponding abstract versions and execute them on the resulting (partial) abstract database.

Observe that the abstraction of foreign/primary key are achieved by using special variable (type-2) in order to maintain integrity constraint. Also, the aggregate functions and set operations are treated differently so as to preserve the soundness.

4 OFGAC for XML

We are now in position to introduce the notion of access control policy specification for XML under OFGAC framework. Then, we apply the OFGAC approach in two directions: view-based and RDBMS-based.

Observation-based Access Control Policy Specification for XML. It is specified by a 5-tuple of the form $\langle Subject, Object, Action, Abstraction, Type \rangle$. The components "*Subject*", "*Object*", "*Action*" and "*Type*" are defined exactly in the same way as in case of FGAC policy specification. The component "*Abstraction*" is defined by the Galois Connection $(\wp(D_x^{con}), \alpha_x, \gamma_x, D_x^{abs})$, where $\wp(D_x^{con})$ and D_x^{abs} represent the powerset of concrete domain of x and the abstract domain of x respectively, and α_x and γ_x represent the corresponding abstraction and concretization functions.

Since the "*Object*" represents either XML element or attribute, the following two cases may arise when "*Abstraction*" is applied on them:

- The "*Object*" represents an intermediate element and "*Type*" is "Recursive" (denoted by "R"). In this case, the abstraction defined in the rule for an element is propagated downwards and applied to all its sub-elements and attributes recursively.
- The "*Object*" represents an attribute and "*Type*" is "Local" (denoted by "L"). In this case, only the attribute value is abstracted by following the Galois Connection specified in the rule.

Table 1. Observation-based Access Control Policy Specification for XML code

Rule	*Subject*	*Object*	*Action*	*Abstraction*	*Type*
R1	customer-care, 159.56.*.*, *.Unicredit.it	/BankCustomers/ Customer/ PersInfo	read	$(\wp(D_x^{con}), id, id, \wp(D_x^{con}))$	R
R2	customer-care, 159.56.*.*, *.Unicredit.it	/BankCustomers/ Customer/ AccountInfo/ IBAN	read	$(\wp(D_{iban}^{con}), \alpha_{iban}, \gamma_{iban}, D_{iban}^{abs})$	L
R3	customer-care, 159.56.*.*, *.Unicredit.it	/BankCustomers/ Customer/ AccountInfo/ type	read	$(\wp(D_{type}^{con}), id, id, \wp(D_{type}^{con}))$	L
R4	customer-care, 159.56.*.*, *.Unicredit.it	/BankCustomers/ Customer/ AccountInfo/ amount	read	$(\wp(D_{amount}^{con}), \alpha_\top, \gamma_\top, \{\top\})$	L
R5	customer-care, 159.56.*.*, *.Unicredit.it	/BankCustomers/ Customer/ CreditCardInfo/ CardNo	read	$(\wp(D_{CardNo}^{con}), \alpha_{CardNo}, \gamma_{CardNo}, D_{CardNo}^{abs})$	L
R6	customer-care, 159.56.*.*, *.Unicredit.it	/BankCustomers/ Customer/ CreditCardInfo/ ExpiryDate	read	$(\wp(D_{ExDate}^{con}), \alpha_\top, \gamma_\top, \{\top\})$	L
R7	customer-care, 159.56.*.*, *.Unicredit.it	/BankCustomers/ Customer/ CreditCardInfo/ SecretNo	read	$(\wp(D_{SecNo}^{con}), \alpha_\top, \gamma_\top, \{\top\})$	L

Example 2. Consider the XML code in Figure 1. The observation-based access control policy under OFGAC framework can be specified as shown in Table 1, where the abstraction functions are defined as follows:

$$\alpha_{CardNo}(\{d_i \ : \ i \in [1 \ldots 16]\}) = **** \ **** \ **** \ d_{13}d_{14}d_{15}d_{16}$$

$$\alpha_\top(X) = \top$$

where X is a set of concrete values and $\top$ is the top most element of the corresponding abstract lattice. The functions α_{iban}, γ_{iban}, γ_{CardNo}, $\gamma_{\top}$ are also defined in this way depending on the corresponding domains. Observe that the identity function *id* is used to provide the public accessibility of non-sensitive information, whereas the functions $\alpha_{\top}$ and $\gamma_{\top}$ are used to provide private accessibility of highly sensitive information by abstracting them with top most element $\top$ of the corresponding abstract lattice.

Given a binary-based access control policy p and an observation-based access control policy *op* in XML format, the FGAC and OFGAC can be implemented in two ways:

- By applying p or *op* directly to the XML documents (view-based) or by rewriting users' XML queries by pruning the unauthorized part in it (NFA-based).
- By taking the support of RDBMS, where the XML documents and the XML policies (p or *op*) are first mapped into the underlying relational databases and the policy SQL respectively, and then the users' XML queries are mapped into equivalent SQL queries and evaluated on those relational databases by satisfying the policy SQL.

Figure 2 depicts a pictorial representation of these approaches. Observe that the application of FGAC *w.r.t.* p results into a binary-based access control system that yields two extreme views to the information: either "allow" or "forbid", whereas the application of OFGAC *w.r.t.* *op*, on the other hand, results into a tunable access control system where partial view of the information at various levels of abstractions is provided.

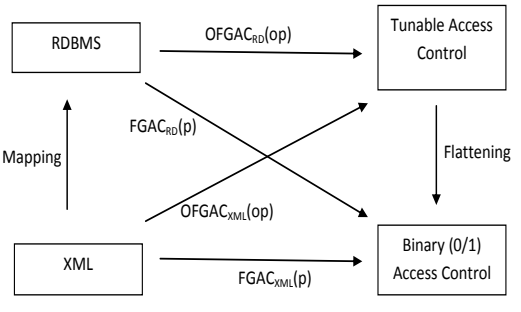


Fig. 2. Pictorial Representation of FGAC Vs. OFGAC

View-based OFGAC for XML. Consider the XML code in Figure 1 and the associated observation-based access control policy specification depicted in Table 1. We know that in view-based approaches for each subject interacting with the

system, separate views are generated with respect to the access rules associated with the subject [6]. Therefore, in our example, the XML view corresponding to the users belonging to "customer-care" section of the bank is depicted in Figure 3.

```
<?xml version="1.0"? >
<BankCusomers>
<Customer>
<PersInfo>
<Cid> 140062 </Cid>
<Name> John Smith </Name>
<Address>
<street> Via Pasini 62 </street>
<city> Venezia </city>
<country> Italy </country>
<pin> 30175 </pin>
</Address>
<PhoneNo> +39 3897745774 </PhoneNo>
</PersInfo>
<AccountInfo>
<IBAN> IT*** ***** ***** ********** </IBAN>
<type> Savings </type>
<amount> ⊤ </amount>
</AccountInfo >
<CreditCardInfo>
<CardNo> **** **** **** 7835 </CardNo>
<ExpiryDate> ⊤ </ExpiryDate>
<SecretNo> ⊤ </SecretNo>
</CreditCardInfo>
</Customer>
</BankCusomers>
```

Fig. 3. View generated for the employees in bank's customer-care section

Consider now the following XML query Q_{xml} issued by a personnel in the customer-care section:

$$Q_{xml} = /BankCusomers/Customer/AccountInfo[@type = "Savings"]$$

The execution of Q_{xml} on the view of Figure 3 returns the following results:

```
<AccountInfo>
<IBAN> IT*** ***** ***** ********** </IBAN>
<type> Savings </type>
<amount> ⊤ </amount>
</AccountInfo>
```

RDBMS-based OFGAC for XML. Consider the XML document in Figure 1 and the observation-based policy specification in Table 1. By following [10], we first map the XML document into relational database representation, partially shown in Table 2. Observe that we do not translate the XML policies into the equivalent SQL statements, rather we put the rules into the relational database itself by associating them with the corresponding elements or attributes. The empty rule in a row specifies that the corresponding element (and its sub-elements and child-attributes) or attribute has public authorization. If any access-conflict occurs for any sub-element, it is resolved simply by adopting *abstraction-take-precedence* policy according to which authorization corresponding to more abstract view overrides the authorization corresponding to less abstract view. The users' XML queries are then mapped into SQL representation and are evaluated on this relational database under OFGAC framework as reported in [8].

Table 2. The equivalent relational database representation of the XML code

(a) "*BankCustomers*"

id	*pid*	*rule*
BC1	null	-

(b) "*Customer*"

id	*pid*	*rule*
C1	BC1	-

(c) "*PersInfo*"

id	*pid*	*rule*
PI1	C1	R1

(d) "*AccountInfo*"

id	*pid*	*rule*
AI1	C1	-

(e) "*CreditCardInfo*"

id	*pid*	*rule*
CI1	C1	-

(f) "*IBAN*"

id	*pid*	*val*	*rule*
IB1	AI1	IT10G 02006 02003 000011115996	R2

(g) "*type*"

id	*pid*	*val*	*rule*
TP1	AI1	Savings	R3

(h) "*amount*"

id	*pid*	*val*	*rule*
AM1	AI1	5000	R4

(i) "*CardNo*"

id	*pid*	*val*	*rule*
CN1	CI1	4023 4581 8419 7835	R5

(j) "*ExpiryDate*"

id	*pid*	*val*	*rule*
EX1	CI1	12/15	R6

Suppose the following XML query Q_{xml} is issued by an employee from customer-care section of the bank:

$$Q_{xml} = /BankCusomers/Customer/AccountInfo[@type = \text{“}Savings\text{”}]/IBAN$$

Since the OFGAC Policies and XML documents are now in the form of relational database, the system translates Q_{xml} into an equivalent SQL query Q_{rdb} as follows:

Q_{rdb} =**SELECT** *Ch_No.val* **FROM** *IBAN Ch_No, type Ch_Tp, AccountInfo P_AccInfo, Customer P_Cust, BankCustomers P_BCust* **WHERE** (*Ch_No.pid = P_AccInfo.id* **AND** *Ch_Tp.pid = P_AccInfo.id* **AND** *Ch_Tp.val = “Savings”*) **AND** *P_AccInfo.pid = P_Cust.id* **AND** *P_Cust.pid = P_BCust.id*

The execution of Q_{rdb} on the database of Table 2, by following the OFGAC technique in [8], yields the following result:

val
IT*** ***** ***** **********

Observe that RDBMS-based approaches suffer from time-inefficiency, whereas view-based approaches, on the other hand, suffer from space-inefficiency. The robustness of the proposed OFGAC system depends on the ability of the external observers to extract sensitive information based on the observable properties of the query results. The possibility of collusion attacks for XML documents under OFGAC framework is same as that of relational databases as described in [8].

5 Conclusions

In this paper, we discussed the extension of the notion of observation-based fine grained access control to the case of XML documents. The traditional FGAC

can be seen as a special case of the proposed OFGAC framework, where the sensitive information are abstracted by the top element $\top$ of their corresponding abstract lattices.

Acknowledgement. Work partially supported by RAS L.R. 7/2007 Project TESLA.

References

1. Bertino, E., Ferrari, E.: Secure and selective dissemination of xml documents. ACM Trans. on Information and System Security 5(3), 290–331 (2002)
2. Bertino, E., Jajodia, S., Samarati, P.: A flexible authorization mechanism for relational data management systems. ACM Trans. on Information Systems 17(2), 101–140 (1999)
3. Bouganim, L., Ngoc, F.D., Pucheral, P.: Client-based access control management for xml documents. In: Proc. of the 13th Int. Conf. on Very Large Data Bases (VLDB 2004), pp. 84–95. VLDB Endowment, Toronto (2004)
4. Cousot, P., Cousot, R.: Abstract interpretation: a unified lattice model for static analysis of programs by construction or approximation of fixpoints. In: Conf. Record of the 6th Annual ACM POPL, pp. 238–252. ACM Press, Los Angeles (1977)
5. Damiani, E., de Capitani di Vimercati, S., Paraboschi, S., Samarati, P.: Design and implementation of an access control processor for xml documents. Journal of Computer and Telecommunications Netowrking 33(1-6), 59–75 (2000)
6. Damiani, E., De Capitani di Vimercati, S., Paraboschi, S., Samarati, P.: A fine-grained access control system for xml documents. ACM Trans. on Information and System Security 5(2), 169–202 (2002)
7. Griffiths, P.P., Wade, B.W.: An authorization mechanism for a relational database system. ACM Trans. on Database Systems 1(3), 242–255 (1976)
8. Halder, R., Cortesi, A.: Observation-based fine grained access control for relational databases. In: Proc. of the 5th Int. Conf. on Software and Data Technologies (ICSOFT 2010), pp. 254–265. INSTICC Press, Athens (2010)
9. Jajodia, S., Samarati, P., Subrahmanian, V.S., Bertino, E.: A unified framework for enforcing multiple access control policies. SIGMOD Record 26(2), 474–485 (1997)
10. Koromilas, L., Chinis, G., Fundulaki, I., Ioannidis, S.: Controlling Access to XML Documents Over XML Native and Relational Databases. In: Jonker, W., Petković, M. (eds.) SDM 2009. LNCS, vol. 5776, pp. 122–141. Springer, Heidelberg (2009)
11. Lee, D., Lee, W.-C., Liu, P.: Supporting XML Security Models Using Relational Databases: A Vision. In: Bellahsène, Z., Chaudhri, A.B., Rahm, E., Rys, M., Unland, R. (eds.) XSym 2003. LNCS, vol. 2824, pp. 267–281. Springer, Heidelberg (2003)
12. Luo, B., Lee, D., Lee, W.-C., Liu, P.: Qfilter: fine-grained run-time xml access control via nfa-based query rewriting. In: Proc. of the 13th ACM Int. Conf. on Information and Knowledge Management (CIKM 2004), pp. 543–552. ACM Press, Washington D.C (2004)
13. Murata, M., Tozawa, A., Kudo, M., Hada, S.: Xml access control using static analysis. ACM Trans. on Information and System Security 9(3), 292–324 (2006)
14. Tan, K.-L., Lee, M.-L., Wang, Y.: Access control of xml documents in relational database systems. In: Proc. of the Int. Conf. on Internet Computing (IC 2001), pp. 185–191. CSREA Press, Las Vegas (2001)

Controlled Access over Documents for Concepts Having Multiple Parents in a Digital Library Ontology

Subhasis Dasgupta[1] and Aditya Bagchi[2]

[1] Kaavo Inc. DA - 29 Sector 1 Salt Lake City, Kolkata 700064, India
sd@kaavo.com
[2] Indian Statistical Institute, 203 B T Road, Kolkata 700108, India
aditya@isical.ac.in

Abstract. This paper proposes a solution to a problem present in Digital Library ontology for accessing concepts having multiple parents. Instead of considering the underlying structure as a tree, authors consider a DAG structure for concept nodes having multiple parents. A hashing mechanism has been developed to avoid change in document annotations against change in ontological structure. The paper highlights the problem and describes the methodologies avoiding the algorithmic details for paucity of space.

1 Introduction

A Digital Library (DL) usually supports documents, related to different subject areas, which may be contributed by many different repositories, seamlessly integrated to offer a composite view to a user. A user while accessing the library is not concerned about the different sources wherefrom the documents are fetched.

In a Digital Library, related documents are usually referred by a common index. This index can be a key word common among the referred documents, can be the common author name or can be a subject name. A subject can again be a topic of another subject, thus forming a hierarchy of subjects. For example, documents on "Balanced Tree" would be subset of the documents on "Tree" which in turn is the subset of "Data Structure" and which again is the subset of "Computer Science". Moreover, in a digital library, same subject name may be reached from different subject areas. For example, documents under "Database" may be contributed by Geographic Information System (GIS) originated from Geography or by Bio-Informatics originated from Biology or by Computer Science & Engineering. So, the subject hierarchy may give rise to a Directed Acyclic Graph. A recent study on the formal model of digital library (DL) has revealed that a DL can be represented as an ontological structure [1], where documents may be classified and stored against appropriate concepts present in the ontology.

A DL environment also has a dynamic user population mostly accessing from remote locations. Other than individual identities, these users are also characterized by other properties that in turn control access to the library. For example, to get access to certain group of documents, one may have to be a member of a particular user-group or must be over certain age or must have a minimum level of academic qualification. Controlled access to digital library objects is a challenging area of research. Any user or group of users, usually designated as subject must have

N. Chaki and A. Cortesi (Eds.): CISIM 2011, CCIS 245, pp. 277–285, 2011.

appropriate authorization to exercise any type of access (Read, Write etc.). A typical authorization model for a DL must also support varying granularity of authorization ranging from sets of library objects to specific portions of objects. A good amount of work has already been done on the secured access of individual objects, particularly text documents in XML or RDF environment [2,3,4,5,6]. Depending on the authorization, access may be granted either to an entire document or a portion of it. Study has been made for both positive and negative authorizations i.e. providing permission to access or explicitly inhibiting access. However, in the ontological structure of a digital library, a user must have access to a concept before accessing the objects (documents) covered by it. Earlier studies on access control models for digital library have considered library object hierarchy and proposed authorization policies for them [7,8]. These hierarchies usually follow tree structures. Even studies on concept level access control for Semantic Web [9,10] have considered tree structures only. However, when represented as ontology, a digital library environment may have a concept derived from more than one concept above it. As mentioned earlier, a concept named *Database* may be reached from Computer Science & Engineering (*CS*), Geographic Information System (*GIS*) or Biology/Bio-informatics (*Bio*). These three concept areas may have distinct or even overlapping user communities. As a result, any document under *Database* may be of interest to more than one of the above three user communities. Research work done so far ensures that a user must possess appropriate authorization to get access to a concept. However, if access to a concept is granted, all documents under it are available to the concerned user. On the other hand, for individual document, controlled access can be provided to different portions of a document. For example, in case of a technical paper, a user may get access to the *Title* and *Abstract* of a paper but the body of the actual paper may not be available to him/her.

Authors of this paper have initiated an Ontology Based Access Control (OBAC) system, particularly for a digital library, in order to support flexible but controlled access to different documents under a concept having multiple parents in. Since a concept may have multiple parents representing multiple subject areas, depending on clearance to a subject area, a user may get access to a document related to that subject area only but not the other documents contributed by other subject areas even when all the documents belong to the same child concept. Motivation under Section 2 explains the situation with example and shows how this OBAC model provides more flexibility to the access control system, so that the controlled access can be imposed not only at the concept level but it may also percolate down to individual documents covered by the concerned concept. Section 3 provides the principles and policies. Section 4 discusses about the implementation. Section 5 draws the conclusion indicating also the future direction of research.

2 Motivation

This paper proposes a flexible access control system for retrieving documents using a digital library ontology supporting an underlying DAG structure to handle multiple parent concept problem. So here a concept may have more than one parent concept in the hierarchy. As a result, the documents under a concept can be categorized against the concepts above it. A user can access a document only if he/she has appropriate authorization to access the category to which the document is placed.

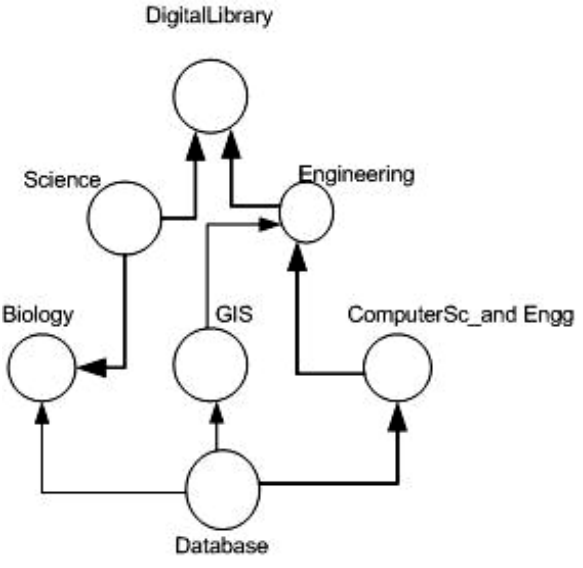


Fig. 1. An ontological structure with the concept *Database* having three parent concepts *Biology, GIS* and *Computer Sc & Engg*

Fig.1 shows an environment where documents covered under the concept *Database* may be contributed by or of interest to any users of the parent concepts. So a document under a child concept can be a member of one or more than one of the parent concepts. Consequently, documents under a child concept having *n* parents, can be classified into (2^n-1) categories. So, the *Database* concept in Fig.1 can be classified into (2^3-1) or 7 categories. Fig.2 shows the Venn diagram corresponding to the concept *Database* having three parent concepts Computer Science (*CS*), Geographic Information System (*GIS*) and Bio-Informatics (*Bio*) as explained earlier. So, a document under the concept *Database* may be of interest to the users of *CS/GIS/Bio* or any combinations of them.

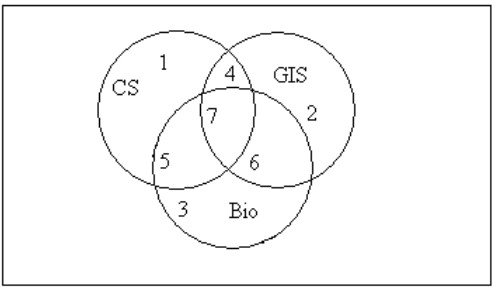


Fig. 2. Possible document categories under the common concept "*DATABASE*"

Situation depicted in Fig.1 and Fig.2 is very common in case of a digital library. However, the present implementations avoid such document classification possibility and keep the child concept under the parent concept that contributes maximum number of documents. So, according to the above example, the concept *Database* would possibly be kept in the *CS* path with all documents under it. Any user of *GIS* or *Bio* community will directly access the *Database* concept and would be able to get all the documents under it. Relevance of a document will be determined by the underlying document retrieval system available with a particular implementation. Present research effort, however, proposes a system where a user will access only the documents relevant to his access path of the ontology and thereby reduces the search time. Moreover, it provides a flexible access control mechanism increasing the security of the overall digital library system. The present paper describes the overall philosophy behind the Ontology Based Access Control (OBAC) system and a possible implementation strategy. Algorithmic details and access control policy algebra have been avoided for paucity of space and would be communicated later in a different forum.

3 Principles and Policies

3.1 Entity of OBAC Model

The OBAC model is represented by a set of *Subject* (S), *Object* (O), *Access Rights* (A) and *Sign* (V).

Depending upon the context and policy, *Subject* can be users, group of users or roles.

Object can be a document under a concept or a concept itself or may even be the entire ontology.

The present effort has considered only read and browse operations and access rights are defined accordingly. A subject getting a read access to a concept also gets the browse access. However, only browse access may also be given. A read access allows a user to get the documents covered by a concept. A browse access, on the other hand, allows a user to browse through a concept node to reach its descendents without accessing the documents covered by it.

Here the *Sign* will be positive for access permission or negative for explicit denial.

Authorization: Authorization is defined by a four tuple (s,o,a,v) where $s \in S$, $o \in O$, $a \in A$ and $v \in V$. If *True,* an authorization represents that the subject s can access the object o with the access right a if v is +ve or cannot access it at all if v is –ve.

Policy: A policy can be a single authorization or a combination of such authorizations. Detail discussion of policy algebra is not within the scope of this paper. Rather the possible implementation strategy would be covered here.

Access Mechanism: A user trying to access the DL would submit his/her credentials to the "Credential Verifier" which in turn would verify whether the concerned user can access the DL at all or not. For example, a user submitting his/her credential as a student of Computer Science (CS) would, by default, get access to documents under the concept CS and also all other documents covered by the concepts placed in the descendant sub-graph of CS in the ontology. The "Credential Verifier" has a "Rule

base" to decide the default category of a user. As part of the credentials, a user may also submit other digital documents that may provide different access rights to a user other than just the membership to a default class or may even inhibit access to any sub-class of documents or to any descendant of an already permitted concept. Detail logical structure and algorithms involved in a "Credential Verifier" is again not within the scope of this paper. After passing through the "Credential Verifier", a user would reach to the "Authorization Server" to receive the appropriate authorizations.

An "Authorization Server" has two purposes. First for a new user, after receiving the output from the "Credential Verifier", it generates the appropriate set of authorizations for different objects applicable for the concerned user. These authorizations are usually of the type,

$$<s,o,a,v> = \text{True/False}$$

Apparently, in an ontological environment with finite number of concepts present, only a representation like $<s,o,a>$ = True/False should suffice signifying that a user (also called subject) '*s*', gets an access to object '*o*' with access right '*a*' if the authorization is positive i.e. True or does not get access at all if authorization is negative i.e. False.

However, the proposed OBAC model accepts the usual inheritance rule of an ontology where access to a concept with certain set of access rights signify that all concepts under it are also accessible with same set of access rights. Since the proposed model so far allows only read and browse access, the policy server may allow a user only browse access to a concept even when the same user has read access (browse included) to its parent concept. So, after allowing (+ve authorization) read access to a parent concept, the Authorization Server may have to issue an explicit -ve authorization to the child concept, so that the concerned user can only browse through such child concept but cannot read the documents covered by it. So,

$<x,c_1,read,+> \wedge <x,c_2,read,->$ where $c_2 < c_1$ (c_2 is a child concept of c_1)
is a valid policy.

The second purpose of the "Authorization Server" is verification of access right for an existing user. When an existing user submits a query, appropriate query processor generates the access plan. This access plan passes through the "Authorization Server", where it verifies whether an access to an object by the concerned user would be permissible or not. For this purpose, the "Authorization Server" not only stores the authorization specifications for each valid user, but also maintains a set of derivation rules that derives other valid authorizations from a set of defined access rights. Once again, this paper is not covering the detail implementation of an "Authorization Server".

Flexible Access Control Policies: Policy specification for flexible access control has been discussed in detail in [14]. In this paper the authors have proposed flexible access control policies for concepts having multiple parents. As described earlier, a concept having 'n' parent concepts can have documents categorized into 2^n-1 classes. In order to explain the flexible access control policy, let's consider that a user has to pay to get read access to a document. So, in order to get read access to documents under the concept *Database* (a composite concept node in the ontology), a user may pay for *CS* or *GIS* or *Bio* or any combination of them. Accordingly, system may define different access control policies for different user groups as shown in Fig.3 referring to the document classes defined in Fig.1.

Document Class	Payment made for Parent nodes	Documents Available (Access Policy-1)	Documents Available (Access Policy-2)
1	CS only	CS only (Class-1)	CS only (Class-1) CS+GIS(Class-4) CS+Bio(Class-5)
:	:	:	:
6	Bio & GIS	GIS only (Class-2) Bio only (Class-3) GIS+Bio (Class-6)	Documents of all classes other than Class-1(CS only)
7	Payment made for all the parent nodes	Access permitted for all documents	Access permitted for all documents

Fig. 3. Multiple Access Policies for Document Classes under Composite Concept Node

Fig.3 shows two possible access control policies for two different groups of users. Policy-1 is definitely stricter than Policy-2. However, user classification has not been considered in this paper. More flexible access control policies can also be defined. Indications of such policies have been made in the conclusion.

This classification of documents demands that the documents under a concept, particularly having multiple parents should be properly annotated so that their classes can be identified against query made under certain declared access control policy and access clearance of the user (appropriate payments made in the present context). So in the example discussed, a document should be annotated by its class number among 1 to 7. As such it is not a problem if the ontology is static. However, in real life situation Digital Library ontology may change. New nodes (concepts) may be added. So for a concept node having multiple parents, two types of updation may take place.

Addition of a Parent Node

Fig.4 shows a situation where a new concept node is added to the ontology as the fourth parent node of the concept *Database.* As a result, the document classification would have to be changed. Earlier, the concept *Database* had 3 parents *CS, GIS* and *BIO*. So, the possible document classes under the concept *Database* were seven (2^3-1=7). Accordingly, the documents were annotated for classification. Now if a new concept X is defined as a parent of *Database* as shown in Fig.4, then the documents under *Database* may be categorized into 2^4-1=15 possible classes. However in the Digital Library, there may not be documents present in all those classes. As shown in Fig.4, addition of new parent X has given rise to two more classes only. On the other hand, documents in class 9 are those which are of interest to both the concepts GIS

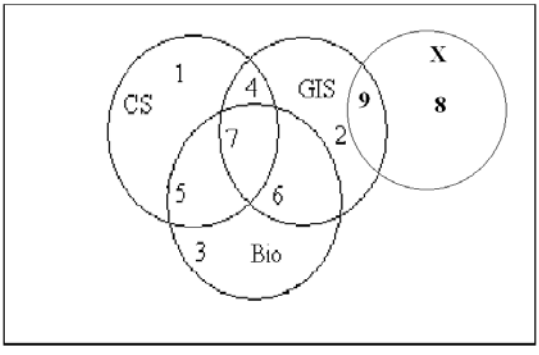


Fig. 4. Change in the ontological structure by addition of a new parent concept

and X. These documents were earlier placed in class 2 and hence to be re-categorized and re-annotated. However, it is not possible to change the annotation of the large number of affected documents. Moreover this restructuring process will also demand change in the authorizations of many users, since such authorizations not only provides access permissions to different concepts, but also to the document classes and thereby to the documents.

Addition of a Child Node

Let's consider that a new child node Pattern Recognition *(PR)* is added for the same set of parent nodes considered earlier i.e. *CS, GIS* and *Bio.* Now, some of the documents stored under the concept *Database* may need to be accessed from *PR* as well (Fig.5). In that case, same document may need to have more than one annotation, one for *Database* and another for *PR.* In order to solve this problem, a hashing technique has been developed to map documents to document classes so that the access mechanism remains invariant to the change in the ontology structure.

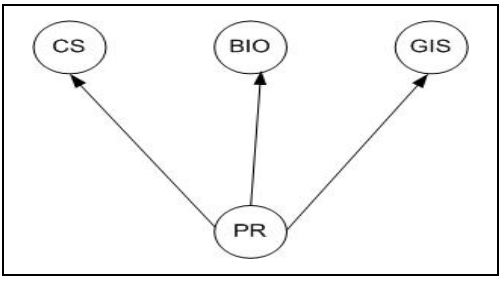


Fig. 5. Change in the ontological structure by addition of a child node

4 Implementation Strategy

First, instead of annotating the documents by the document class under a composite concept node, a unique document-id is assigned to each document stored in the library. In the earlier situation depicted in Fig.1, a hashing algorithm reads the

relevant key words of a document and places its id in one of the 7 buckets representing 7 classes. Under the changed circumstances, since a new concept X has been defined as the fourth parent of *Database*, documents (key word set of each document) can be re-hashed to place their ids in the new set of 9 buckets as shown in Fig.4. As a matter of fact, in the situation shown in Fig.4, only the documents in class 2 need to be re-hashed to decide whether a document-id is to be retained in bucket for class 2 or to be placed in bucket of class 9. Similarly in case of addition of new child node, same document–id may be placed in the appropriate class under *Database* and also in the appropriate class under *PR* without disturbing the actual documents. Detail of the hashing algorithm and management of the hashing process could not be provided here for space constraint. However, standard key word based document retrieval methods can be used for classifying documents and placing them into appropriate document classes.

5 Conclusion

This paper has described a flexible access control mechanism for a Digital Library Ontology. For different group of users different access control strategies can be adopted. For a concept having multiple parents, documents under it have been classified according to the parent concepts. This approach provides fast retrieval to relevant documents. Document annotation problem under different document classes have been solved. The proposed strategy also takes care of structural changes in the ontology.

References

1. Gonçalves, M.A., Watson, L.T., Fox, E.A.: Towards a digital library theory: a formal digital library ontology. Int. J. Digital Libraries 8(2), 91–114 (2008)
2. Bertino, E., Ferrari, E.: Secure and Selective Dissemination of XML Documents. ACM Trans. On Information and System Security 5(3), 290–331 (2002)
3. Gabillon, A.: A Formal Access Control Model for XML Databases. In: Workshop on Secured Data Management, VLDB (2005)
4. Carminati, B., Ferrari, E., Bertino, E.: Securing XML Data in Third Party Distribution Systems. In: Proc. ACM CIKM, pp. 99–106 (2005)
5. Farkas, C., Gowadia, V., Jain, A., Roy, D.: From XML to RDF: Syntax, Semantics, Security, and Integrity. In: Security Management, Integrity and Internal Control in Information Systems, pp. 41–55. Springer, Heidelberg (2006)
6. Damiani, E., De Capitani di Vimercati, S., Paraboschi, S., Samarati, P.: Securing XML Documents. In: Zaniolo, C., Grust, T., Scholl, M.H., Lockemann, P.C. (eds.) EDBT 2000. LNCS, vol. 1777, pp. 121–135. Springer, Heidelberg (2000)
7. Adam, N.R., Atluri, V., Bertino, E., Ferrari, E.: A content-based authorization model for digital libraries. Proc. IEEE TKDE 14(2), 296–315 (2002)
8. Ray, I., Chakraborty, S.: A Framework for Flexible Access Control in Digital Library Systems. In: Damiani, E., Liu, P. (eds.) Data and Applications Security 2006. LNCS, vol. 4127, pp. 252–266. Springer, Heidelberg (2006)

9. Damiani, E., De Capitani di Vimercati, S., Fugazza, C., Samarati, P.: Modality Conflicts in Semantics Aware Access Control. In: Proc. ACM ICWE 2006, pp. 249–256 (July 2006)
10. Qin, L., Atluri, V.: Concept-level Access Control for the Semantic Web. In: Proc. ACM Workshop on XML Security, pp. 94–103 (2003)
11. Caseau, Y.: Efficient Handling of Multiple Inheritance Hierarchies. In: Proc. OOPSLA, pp. 271–287 (1993)
12. van Bommel, M.F., Beck, T.J.: Incremental encoding of multiple inheritance hierarchies supporting lattice operations. Electronic Transactions on Artificial Intelligence 5(001), 35–49 (2000)
13. van Bommel, M.F., Wang, P.: Encoding Multiple Inheritance Hierarchies for Lattice Operations. Data and Knowledge Engineering 5(2), 175–194 (2004)
14. Jajodia, S., Samarati, P., Sapino, M.L., Subrahmanian, V.S.: Flexible support for multiple access control policies. ACM Trans. On Database Systems 26(2), 214–260 (2001)

Secret Image Sharing with Embedded Session Key

Prabir Kumar Naskar, Hari Narayan Khan, Ujjal Roy, Ayan Chaudhuri, and Atal Chaudhuri

Department of Computer Science & Engineering, Jadavpur University, Kolkata 700032, West Bengal, India
{cse.prabir,manik1984,royujjal, ayanchaudhuri27,atalc23}@gmail.com

Abstract. In today's scenario many secret image files are needed to be transmitted over internet for various important purposes. So protection of these files is an important issue. Efficient cryptographic methods are there to protect data but every thing depends on the protection of encryption key. Various secret sharing schemes are available in the literature but computational complexity is very high in most of the cases that leads to single point failure. To overcome this drawback shared cryptography becomes more popular. Here we are suggesting a novel secret sharing scheme which employs simple graphical masking method, performed by simple ANDing for share generation and reconstruction can be done by performing simple ORing the qualified set of legitimate shares. Here sharing and reconstruction are so simple, then our scheme may be applied in many hand held system with very low end processors. Not only that, the generated shares are compressed and each share contains partial secret information. Again decompression is possible only when qualified set of legitimate shares are available, that leads to added protection to the secret and reduced bandwidth requirement for transmission.

Keywords: Threshold Cryptography, Image sharing, Compression, Perfect Secret Sharing (PSS).

1 Introduction

The effective and secure protections of sensitive information are primary concern in commercial, research and military systems. Today secret image is transferred over Internet for various commercial purposes. Therefore, it is also important to ensure data is not being tampered.

Imagine that there is a deadly weapon whose production is an automated process controlled by a computer. The product plan has been loaded to the computer and the computer controls the production by executing the instructions given in the plan. The plan is kept encrypted. Before each production, the plan is decrypted and after the production, decrypted plan is deleted. This guarantees that the production can only be initiated by authorized people. This weapon is so dangerous that no one is allowed to start its production alone. Therefore, the decryption key must be distributed to the

N. Chaki and A. Cortesi (Eds.): CISIM 2011, CCIS 245, pp. 286–294, 2011.

officials such that the production can be started only when an authorized group decides to do so. This can be made possible if there is software loaded to the computer that can split the original encryption key into pieces and reconstruct the key using some of these pieces. In this scenario, we assume that the computer does not store the original encryption key permanently and it works perfectly securely and reliably. It is also important for an information process to ensure data is not being tampered. Encryption methods are one of the popular approaches to ensure the integrity and secrecy of protected information. However, one of the critical vulnerabilities of encryption techniques is single-point-failure. That is the secret information cannot be recovered if the decryption key is lost or the encrypted content is corrupted during the transmission. To overcome this drawback, secret sharing becomes more popular.

This is basically a variant of threshold cryptography, which deals with sharing sensitive secret among a group of n users so that only when a sufficient number k ($k \leq n$) of them come together, the secret can be reconstructed. Well known secret sharing schemes (SSS) in the literature include Shamir [1] based on polynomial interpolation, Blakley[2] based on hyper plane geometry and Asmuth-Bloom[3] based on Chinese Remainder theorem.

Shamir's [1] scheme based on a polynomial of degree (k-1) to any set of k points that lie on the polynomial. The method is to create a polynomial of degree (k-1) as follows-

$$f(x) = d_0 + d_1x^1 + d_2x^2 + - - - + d_{k-1}x^{k-1} \text{ (mod p)} \quad (1)$$

where d_0 is the secret and p is a prime number and the remaining coefficients picked at random. Next find n points on the curve and give one to each of the participants. When at least k out of the n participants revel their points, there is sufficient information to fit an $(k-1)^{th}$ degree polynomial and then the secret value d_0 can be easily obtained by using Lagrange Interpolation.

Blakley [2] used geometry to solve the secret sharing problem. The secret message is a point in a k-dimensional space and n shares are affine hyperplanes that intersect in this point. The set solution $x=(x_1, x_2, x_3,\ldots, x_k)$ to an equation

$$a_1x_1 + a_2x_2 + \text{---} + a_kx_k = b \quad (2)$$

forms an affine hyperplane. The secret, the intersection point, is obtained by finding the intersection of any k of these planes.

Asmuth-Bloom's [3] scheme in which reduction modulo operation is used for share generation and the secret is recovered by essentially solving the system of congruence using Chinese Remainder Theorem (CRT).

Above all secret sharing schemes are regarded as a Perfect Secret Sharing (PSS) scheme because coalition of (k-1) shares doesn't expose any information about the secret. A shortcoming of above secret sharing schemes is the need to reveal the secret shares during the reconstruction phase. The system would be more secure if the subject function can be computed without revealing the secret shares or reconstructing the secret back. This is known as function sharing problem where the function's computation is distributed according to underlying SSS such that distributed parts of computation are carried out by individual user and then the partial results can be combined to yield the final result without disclosing the individual secrets. Various function sharing protocols are been proposed [4], [5], [6], [7], [8], [9], [10] mostly based

on Shamir's secret sharing as the underlying scheme. Some work [11] is also available on Blakley's secret sharing scheme and Asmuth-Bloom scheme [12] as well.

In all of above secret sharing schemes, each share hold the complete secret information in encrypted or ciphered form. We have suggested a different concept, where simple graphical masking (ANDing) technique is used for shared generation and all the shares contain partial secret information and reconstruction is done by simply ORing the predefined minimal set of shares.

The success of the scheme depends upon the mask generation, a step wise algorithm is suggested for such mask design for any (k, n) scheme where n numbers of masks are designed to generate n different shares and any k shares on ORing reconstruct the original secret. Here we have further proposed an unique compression technique on the shares as a solution towards decreasing the bandwidth requirement for transmitting multiple shares to some extent possible.

2 Secret Sharing Algorithm

The proposed work is based upon a novel secret sharing scheme which employs simple graphical masking method using simple ANDing for share generation and reconstruction can be done by simple ORing the predefined minimal number of shares.

2.1 Concept

For better understanding let us consider any secret as a binary bit file (i.e. bit is the smallest unit to work upon, in actual implementation one can consider a byte or group of bytes or group of pixels as the working unit). The secret could be an image, an audio or text etc. We shall decompose the bit file of any size onto n shares in such a way that the original bit file can be reconstructed only ORing any k number of shares where $k \leq n \geq 2$ but in practice we should consider $2 \leq k < n \geq 3$.

Our basic idea is based on the fact that, every share should have some bits missing and those missing bits will be replenished by exactly (k-1) other shares but not less than that. So every individual bit will be missed from exactly (k-1) shares and must be present in all remaining (n-k+1) shares, thus the bit under consideration is available in any set of k shares but not guaranteed in less than k shares. Now for a group of bits, for a particular bit position, (k-1) number of shares should have the bit missed and (n-k+1) number of shares should have the bit present and similarly for different positions there should be different combinations of (k-1) shares having the bits missed and (n-k+1) number of shares having the bits present. Clearly for every bit position there should be ${}^{n}C_{k-1}$ such combinations and in our scheme thus forms the mask of size ${}^{n}C_{k-1}$, which will be repeatedly ANDed over the secret in any regular order. Different masks will produce different shares from the secret. Thus, 0 on the mask will eliminate the bit from the secret and 1 in the mask will retain the bit forming one share. Different masks having different 1 and 0 distributions will thus generate different shares.

Next just ORing any k number of shares we get the secret back but individual share having random numbers of 1's & 0's reflect no idea about the secret.

As an example a possible set of masks for 5 shares with threshold of 3 shares is shown below:

Share-1	:	1	1	1	0	1	1	0	1	0	0
Share-2	:	1	1	1	0	0	0	1	0	1	1
Share-3	:	1	0	0	1	1	1	0	0	1	1
Share-4	:	0	1	0	1	1	0	1	1	0	1
Share-5	:	0	0	1	1	0	1	1	1	1	0

One can easily check that ORing any three or more shares we get all 1's but with less than three shares some positions still have 0's i.e. remain missing.

2.2 Algorithm

Here we are presenting the algorithm for designing the masks for n shares with threshold k.

Step-1: List all row vectors of size n having the combination of (k-1) numbers of 0's and (n-k+1) numbers of 1's and arrange them in the form of a matrix. Obvious dimension of the matrix will be ${}^{n}C_{k-1} \times n$.

Step-2: Transpose the matrix generated in Step-1. Obvious dimension of the transposed matrix will be $n \times {}^{n}C_{k-1}$. Each row of this matrix will be the individual mask for n different shares. The size of each mask is ${}^{n}C_{k-1}$ bits, i.e. the size of the mask varies with the value of n and k.

Pseudo Code for mask generation:

Input: n, k
Output: masks and length of each mask

```
int mask_generator(n, k, mask[n][])
{
    bin[][n] // integer array
    len = 0; //initialization
    max_val = 2^n - 1; //Maximum value of n bit number
    for i=max_val-1  to 0
        //Store binary equivalent of i to bin
        decimal_to_binary(i, bin[len]);
        //If (k-1) no of zeros exist then increment len
        if (zero_check(bin[len], k))
            len++;
        end if
    end for
    rearrange(bin);    //Shuffle the row of bin array
    //Take transpose matrix of bin and store in mask
    transpose(mask, bin);
    return len;
}
```

Let us consider the previous example where n=5 and k=3.

Step-1: List of row vectors of size 5 bits with 2 numbers of 0's and 3 numbers of 1's.

1	1	1	0	0
1	1	0	1	0
1	1	0	0	1
0	0	1	1	1
1	0	1	1	0
1	0	1	0	1
0	1	0	1	1
1	0	0	1	1
0	1	1	0	1
0	1	1	1	0

Dimension of the matrix is ${}^5C_2 \times 5$ i.e. 10×5

Step-2: Take the transpose of the above matrix and we get the desired masks for five shares as listed above in the form of matrix of dimension $5 \times {}^5C_2$ i.e. 5×10. There are five masks each of size 10 bits.

3 Image Sharing Protocol

Here we are presenting stepwise protocol for our image secret sharing scheme. In our scheme, we share both secret data and key. Therefore, every share has two parts, secret share and header share.

3.1 Sharing Phase

Step-1: First construct Header Structure (h) of five fields and put share number (S) in 1st field, total number (n) of shares in 2nd filed, threshold number (k) in 3rd filed, key ($\mathcal{K}$) in 4th filed, and the width of secret image in bytes (B) in 5th filed.

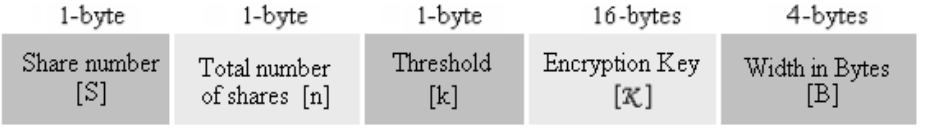


Fig. 1. Header Structure

Step-2: Generates n masks for n individual shares using the proposed mask generation algorithm for n and k.

Step-3: Generate 16-byte digest from the session key ($\mathcal{K}$) defined for encrypting the secret image.

[*Share Generation*]

Step-4: Now select a mask and apply logical AND (byte in the secret image corresponding to bit one of the mask will be retained and that corresponding to bit zero of the mask will be set to zero) repeatedly with the secret image and the zero byte in the generated share corresponding to zero bit of the mask be discarded, this generates one compressed secret share.

Step-5: Then the 1st retained byte (P_1) will be ciphered by the 1st digest byte (Q_1) by the following operation:

$$R_1 = (P_1 \times Q_1) \bmod 251 \tag{3}$$

And 2nd retained byte will be ciphered by the 2nd digest byte Q_2.

[*Header Share*]

Step-6: From each ciphered share we take k number of bytes by which we forms (n × k) matrix (A).

$$\begin{vmatrix} a_{[0,0]} & a_{[0,1]} & - - - & a_{[0,k-2]} & a_{[0,k-1]} \\ a_{[1,0]} & a_{[1,1]} & - - - & a_{[1,k-2]} & a_{[1,k-1]} \\ - - - & & & & - - - \\ a_{[n-1,0]} & a_{[n-1,1]} & - - - & a_{[n-1,k-2]} & a_{[n-1,k-1]} \end{vmatrix}$$

Step-7: Now the header [Figure-1] excluding the leftmost field is also shared by applying following operation-

$$V_i = \sum (a_{[i,\, j]} \times h_{[j]}), \tag{4}$$

where i=0, 1, … n-1 and j=0, 1, …, k-1

Step-8: Next each header share is appended with the share number (S) in the first field and concatenated with the corresponding secret share, which forms one complete share for transmission.

3.2 Reconstruction Phase

[*Header Reconstruction*]

Step-1: First collect k-numbers of share and extract confused header information. Also generates (k × k) matrix (A).

Step-2: Now applying any conventional linear equation solving technique to regenerate the original Header information.

[*Secret Reconstruction*]

Step-3: Once the original Header is reconstructed, we extract the Key ($\mathcal{K}$), and using $\mathcal{K}$ we generate same 16-byte digest string.

Step-4: Now using n and k, extracted from reconstructed Header structure, generate n masks (the masks used in sharing phase) using our mask generation algorithm.

Step-5: According to the share number of the share holder appropriate mask is used to expand the secret share part by inserting zero bytes corresponding to zero bit in the corresponding mask.

Step-6: Ciphered bytes (R_1) corresponding to 1-bit position in the mask, have generated by the equation-(3). So here, we apply following operation to get original byte (P_1).

$$P_1 = (R_1 \times M_1^{-1}) \bmod 251 \tag{5}$$

Where Q_1 is the first digest byte and M_1^{-1} is the multiplicative inverse of Q_1.

Step-7: Next k numbers of expanded secret shares are ORed to reconstruct the original secret. Here we extract the width (B) of the original secret image from Header structure, by which we can easily reconstruct the lossless original secret Image.

4 Analysis of Compression

All masking pattern has equal number of zeros with different distribution only. In every share, we collapse all zero bytes corresponding to zero bit in the corresponding mask. It may be noted that as k is closer to n, more is compression i.e. maximum for k = n.

Next for lossless expanding, knowing n and k we can redesign all n masks using our original mask generation algorithm. According to the share number of the share holder appropriate mask is used to expand the secret share by inserting zero bytes corresponding to zero bit in the corresponding mask. In our example of (3, 5) the mask size is of 10 bits and every mask has 4 zeroes, thus every secret can be compressed by approximately 40%, obviously the compression varies with (k, n). (In case of an example of (5, 6) the mask size is 15 bits and every mask has 10 zeroes, thus compression will be 66.6%).

Table 1. Shows compression rate for different (k, n)

Total number of Shares (n) = 5			
Threshold (k)	Length of Masking Pattern	Number of zero in masking pattern	Approximate Compression Rate (%)
2	5	1	20
3	10	4	40
4	10	6	60
5	5	4	80
Total number of Shares (n) = 6			
2	6	1	16
3	15	5	33
4	20	10	50
5	15	10	66
6	6	5	83

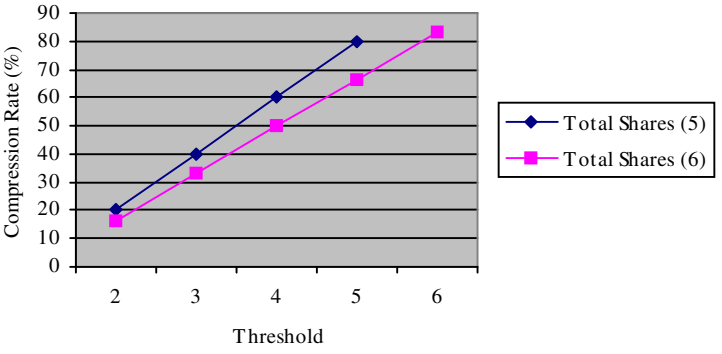
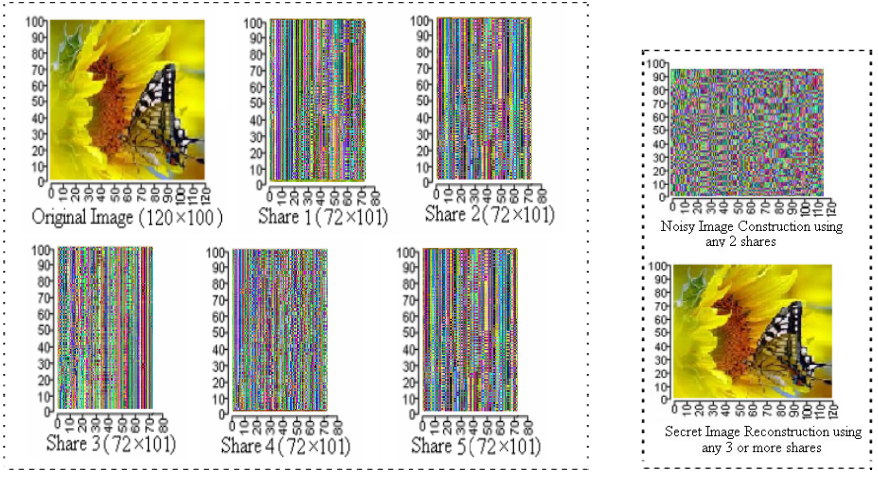


Fig. 2. Threshold vs. compression rate



a. Sharing Phase b. Reconstruction Phase

Fig. 3. (a) Original secret image and Five shared images, (b) Reconstructed Noisy image and original Image

5 Strength of the Protocol

Here we use an image file as a secret. But our proposed scheme is equally applicable for any binary file such as audio (.wav), Text etc. In our scheme if and only if numbers of collating shares are equal to k or more, then only the original secret image is reconstructed; otherwise reconstructed image will be completely noisy. Because fewer shares can not reconstruct the original header, thus we can not have either right key ($\mathcal{K}$) or the information to construct the correct masking pattern. So, our proposed scheme can claim to be a Perfect Secret Sharing (PSS) Scheme.

Here all generated shares are compressed and hold partial secret information in encrypted form. That provides strong protection of the secret file and reduced the

bandwidth of transmission medium. Not only that, only when legitimate group of shares are come together, then we only reconstruct the original secret information.

6 Conclusion

We present a novel secret sharing approach which is PSS and generates compressed noisy shares. Like other group of researchers the shares can be sent in some cover medium. The cover medium may be Image, Audio, and Video. Therefore noisy shares become innocent shares that protect from attackers eyes. However, all of these methods need cover file (i.e. the meaningful shares) size bigger than the secret but in our case cover image of same size as the secret is good enough. Our future effort will try to reduce the size of the cover image further i.e. cover image size may be lesser than the secret.

Acknowledgments. We are thankful to the department of Computer Science & Engineering of Jadavpur University, Kolkata, for giving us the platform for planning and developing this work in departmental laboratories.

References

1. Shamir, A.: How to share a secret? Comm. ACM 22(11), 612–613 (1979)
2. Blakley, G.: Safeguarding cryptographic keys. In: Proc. of AFIPS National Computer Conference (1979)
3. Asmuth, C., Bloom, J.: A modular approach to key safeguarding. IEEE Transaction on Information Theory 29(2), 208–210 (1983)
4. Desmedt, Y.: Some Recent Research Aspects of Threshold Cryptography. In: Okamoto, E. (ed.) ISW 1997. LNCS, vol. 1396, pp. 158–173. Springer, Heidelberg (1998)
5. Desmedt, Y., Frankel, Y.: Threshold Cryptosystems. In: Brassard, G. (ed.) CRYPTO 1989. LNCS, vol. 435, pp. 307–315. Springer, Heidelberg (1990)
6. Desmedt, Y., Frankel, Y.: Shared Generation of Authenticators and Signatures. In: Feigenbaum, J. (ed.) CRYPTO 1991. LNCS, vol. 576, pp. 457–469. Springer, Heidelberg (1992)
7. Desmedt, Y., Frankel, Y.: Homomorphic zero knowledge threshold schemes over any finite abelian group. SIAM Journal on Discrete Mathematics 7(4), 667–675 (1994)
8. Huang, H.F., Chang, C.C.: A novel efficient (t, n) threshold proxy signature scheme. Information Sciences 176(10), 1338–1349 (2006)
9. De Santis, A., Desmedt, Y., Frankel, Y., Yung, M.: How to share a function securely? In: Proc. of STOC 1994, pp. 522–533 (1994)
10. Shoup, V.: Practical Threshold Signatures. In: Preneel, B. (ed.) EUROCRYPT 2000. LNCS, vol. 1807, pp. 207–220. Springer, Heidelberg (2000)
11. Bozkurt, Kaya, Selcuk, Guloglu: Threshold Cryptography Based on Blakely Secret Sharing. Information Sciences
12. Kaya, K., Selcuk, A.A.: Threshold Cryptograhy based on Asmuth-Bloom Secret Sharing. Information Sciences 177(19), 4148–4160 (2007)

Command and Block Profiles for Legitimate Users of a Computer Network

Anna M. Bartkowiak

[1] Institute of Computer Science, University of Wroclaw
Joliot Curie 15, 50-383 Wroclaw, Poland
aba@ii.uni.wroc.pl,
[2] Wroclaw High School of Applied Informatics, Wroclaw, Poland

Abstract. Intruders and masqueraders are a plague in computer networks. To recognize an intruder, one firstly needs to know what is the normal behavior of a legitimate user. To find it out, we propose to build pairs of profiles called 'command and block profiles'. Schonlau data (SEA) are used for illustration of the concept and its usability in work with real data. The elaborated data contain observations for 50 users; for each of them a sequence of 15,000 system calls was recorded. Data for 21 users are pure; data for the remaining 29 users are contaminated with activities of alien (illegitimate) users. We consider only the uncontaminated data (for the 21 users). 5 out of 21 investigated users seem to change their profiles during work time. Some trials have shown that the proposed simple method may also recognize a big part of alien implanted blocks.

Keywords: computer security, legitimate user, intruders, alien blocks, masquerade, Schonlau data, unix commands, anomaly, outliers.

1 Introduction, Statement of the Problem

Nowadays, when working in a computer network, we are exposed to many attacks of alien intruders and masqueraders, who want to get some information to which we have access and they have not. How to detect such illegitimate users? There has been quite a lot of research on this topic. One idea is to inspect the sequence of system calls issued by a user working in a computer network. It is anticipated that an alien (that is: illegitimate) user behaves somehow differently as the legitimate one. A test rig for the problem was provided by [1]. The authors gathered also some data (called now Schonlau or SEA data). The data set contains system calls issued by 50 legitimate users; for each user a sequence of 15,000 system calls was recorded. These legitimate data were interspersed randomly by alien blocks (coming from other 20 users) simulating insider hackers. The task is to find the implants. The problem has been elaborated by many researchers using sophisticated methods, see [2]–[11]; further 38 references on masquerade detection may be found in [11]. But, to the authors knowledge, no publication has concentrated only on the legitimate users appearing in the SEA data, considering the question: To what degree data for one legitimate user working over

N. Chaki and A. Cortesi (Eds.): CISIM 2011, CCIS 245, pp. 295–304, 2011.

longer time are similar in its parts? In the following we elaborate that question. In next section (2) the Schonlau's experiment and its data are described. Section 3 contains the new proposal to use for comparison of users data (and parts of them) the introduced *command and block profiles*. These are constructed for individual users. We show also two examples of contaminated data, where the implanted blocks are immediately visible. It is found, that for most of the legitimate users the sub-parts of their data are consistent. Section 4 looks at all the 21 legitimate users in a holistic way and projects them onto a self-organizing map (SOM). Section 5 contains some discussion and closing remarks.

2 Schonlau's Experiment and the Derived Data

Schonlau et. al. [2] have made the following experiment: They recorded data of 70 users working in a computer network of an institution. For each user they recorded 15,000 system calls (unix commands). The recorded sequence for each user was subdivided into blocks of 100 commands each, yielding 150 blocks; these blocks were further subdivided into three parts: I. blocks 1–50, II. blocks 51–100, and III. blocks 101–150. The data of 50 users were retained as the proper data for further analysis. These data were subjected to a simulated intrusion: block I remained always intact (pure, uncontaminated), but blocks from part II and III were randomly exchanged with blocks of users 51-70 left apart.

The data set containing the 50 possibly 'invaded' users, with the information, which blocks are *alien* (i.e. contaminated), is publicly available [1], it will be called in the following the *SEA* (from Schonlau et al.) data set. The problem is: Given the data of the 50 users, recognize the alien blocks. Table 1 shows the statistics of possessing alien blocks by each of the 50 users from the set SEA. The 21 users with 0 alien blocks will be the subject of our further analysis.

Table 1. Part (a): Number of pure and contaminated blocks ($|contbl|$) for 50 users included into Schonlau data (SEA). Part (b): Relabelled 'pure' (i.e. non contaminated) users and number of commands $|cmd|$ issued by each pure user

(a) all users in SEA numeration # 1, ..., 50
and number of contaminated blocks $|contbl|$

SEA #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
$\|contbl\|$	0	3	11	2	0	0	13	0	24	13	0	6	0	0	6	10	0	6	0	0	0	0	1	21	9
SEA #	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
$\|contbl\|$	13	0	3	1	3	0	0	0	12	1	6	2	9	0	0	3	20	16	6	5	4	0	2	0	0

(b) non contaminated users relabelled to 1, ..., 21
and number of different commands $|cmd|$ employed by each of them

no.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
SEA	1	5	6	8	11	13	14	17	19	20	21	22	27	31	32	33	39	40	47	49	50
$\|cmd\|$	164	100	91	104	138	192	100	139	110	157	74	128	130	80	97	155	133	136	97	163	133

3 Analysis of Non Contaminated Data, by Individual Users

3.1 Reorganizing the Sequence of Each User into a Data Matrix

Taking the sequence of 15,000 commands of each user, we found in first place, how many different commands were employed by that user. This was done using the Matlab function `unique`. The number of unique commands u employed by each user is shown in part (b) of Table 1. One may notice that the counts vary between 70 and 198. Generally, one may expect here about 100 – 150 commands.

Next an incidence matrix $\mathbf{X}$ of size $150 \times u$ was calculated, with rows $(=150)$ denoting blocks and columns $(=u)$ denoting commands employed by the given user. The elements x_{ij} of $\mathbf{X}$ denote, how many times the jth command $(j = 1, \ldots, u)$ has appeared in block no. i. For further analysis, the 150 rows of $\mathbf{X}$ were subdivided into 3 parts: Part I: rows 1–50; Part II: rows 51–100; Part III: rows 101–150. These parts have served as the base for constructing command-and-block profiles described in next subsection.

3.2 Defining Command and Block Profiles

We start from the fact that each user is represented by a data matrix $\mathbf{X}_{150\times u}$. The matrix is split into three parts, each containing 50 blocks of commands. First of all, we calculate from each part a row vector $\mathbf{s}_{1\times u} = (s_1, \ldots, s_u)$ denoting the frequency of specific commands issued in that part. For three parts of $\mathbf{X}$ we obtain 3 such vectors say $\mathbf{v}^I$, $\mathbf{v}^{II}$, and $\mathbf{v}^{III}$. The three vectors are displayed in respective scatter plots depicting v_j against j, $(j = 1, \ldots, u)$. Two such plots, for users 11 and 33 (SEA numeration), are shown in the left column of Fig. 1. We call such plots ***command profiles***. They depict the usage of commands (in the ordering provided by the Matlab function `unique`). The consecutive appearing in time is destroyed, only the frequency of appearing is recorded. One may inspect such profiles by 'eye' and notice some specificities about usage of the commands, in particular whether there are some which are employed exceedingly frequently, or perhaps, some of them are not used at all.

Looking at the exemplary displays in Fig. 1 one may notice, that user no. 11 behaves consistently, and all three his command profiles look alike. What concerns user 33, one might say that – to some degree – all his command exhibit also some similarity, but not so very close. Thus, we should use additional means to confirm the similarity of behavior of that user. A formal tool for comparing two command profiles is described in subsection 3.3. As other tool, we propose here to construct another graph, which we call ***block profile***.

The proposed block profile is computed separately for each of the parts I, II, III. It considers individually the appearance of commands within each block. Each block is now viewed as a multivariate data vector characterized by u attributes, which are frequencies of the u unique commands found by Matlab:

$$\mathbf{x}_i = (x_{i1}, \ldots, x_{iu}), \quad i = 1, \ldots, 50. \tag{1}$$

We are interested in distribution of the data vectors $\{\mathbf{x}_i\}$, in particular in possible outliers. There are many methods of finding outliers [12]. A very good method is computing Mahalanobis distances. Yet this method is not applicable, because we have more variables then samples. However, ordinary Euclidean distances are applicable. Thus for each block (i), its squared Euclidean distance from the origin of the multivariate coordinate system is computed:

$$d^2_{Eu}(i) = d^2_{Eu}(\mathbf{x}_i - \mathbf{0}) = \sum_{j=1}^{u} x^2_{ij}, \quad i = 1, \ldots, 50. \tag{2}$$

The distances are computed separately for parts I, II, III of the given user's data. The values of $d^2_{Eu}(i)$ put against i, the no. of the block, constitute the block profile of the user. Each user has 3 such profiles, which may be superimposed. Exemplary block profiles are shown in right column of Fig. 1.

Points representing blocks coming from different parts of the data are represented by different markers and colors. The red line appearing in the block

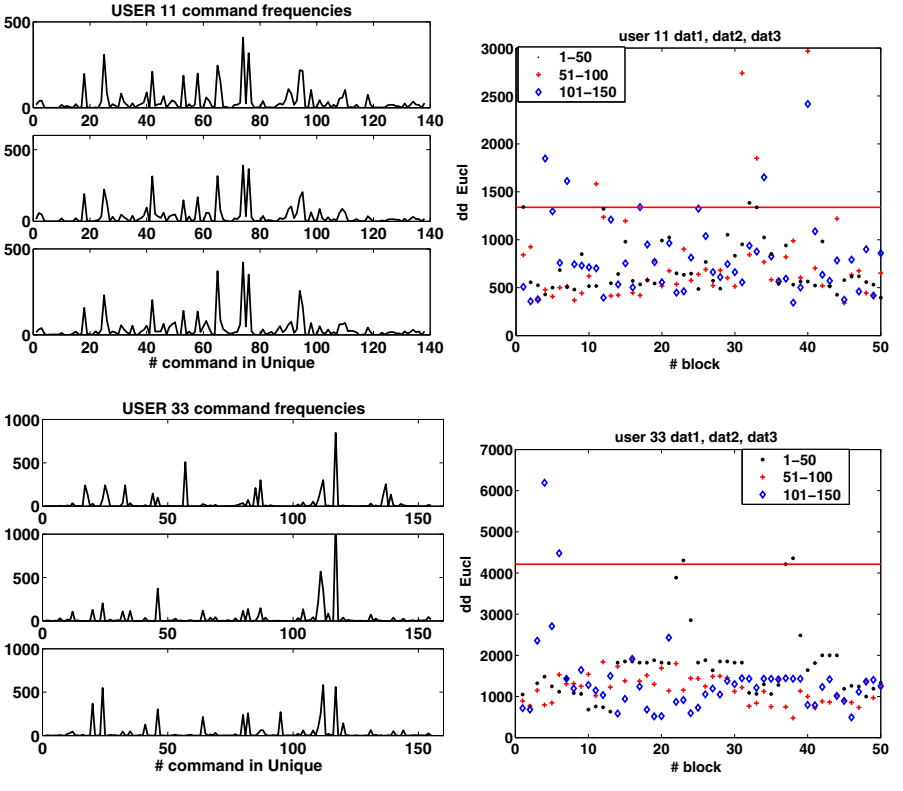


Fig. 1. Non-contaminated data. Command profiles (at the left) and block profiles (at the right) constructed for legitimate, non-invaded users U11 and U33 (SEA numeration)

profiles is an alpha=0.05 delimiter of the upper distances $d^2(i),\ i = 1, \ldots, 50$ evaluated from part I of the data. Points above this line may be deemed as outliers. A similar line might be constructed at the bottom of the exhibit. In subsection 3.4 we tell more, how to draw such delimiting lines.

Generally, the command profile helps in stating a change of the behavior with respect of commands usage, and the block profile – in identifying outliers, what is also seen in Fig. 2 constructed for contaminated data. The two profiles complement each other.

The displays shown in both Fig. 1 and Fig. 2 look quite convincing: the command profiles have their individual character for each of the involved users and the outliers appearing in Fig. 2 are really outstanding. None the less, we

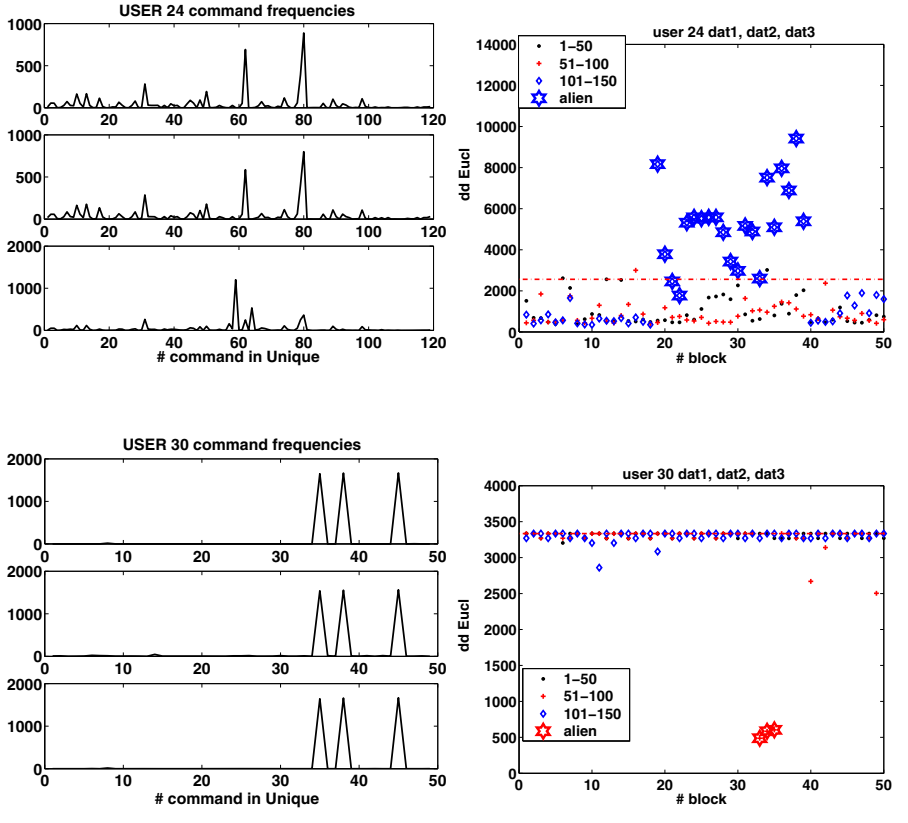


Fig. 2. Contaminated data. Command profiles (left) and block profiles (right) constructed for users U24 and U30 (SEA numeration) with 21 and 3 alien blocks appropriately. The alien blocks are marked as big stars. User U24 got 21 alien blocks in Part III of his data (marked in blue); user U30 with a systematic strange behavior got 3 alien blocks in Part II of his data (marked in red).

would be more glad to work with some statistical indices, permitting to gauge the normality or abnormality of the displays.

3.3 Similarities between Command Profiles Belonging to One User

We will consider now frequency profiles of non contaminated users recorded as data vectors:

$$\mathbf{v}_i^t = (v_{i1}^t, \ldots, v_{iu}^t), \quad i = 1, \ldots, 21, \tag{3}$$

where u denotes the number of unique commands issued by the given user, and $t = 1, 2, 3$ indicates, whether this vector was calculated from the first, second or third part of the data. **Notice that now the investigated users appear in the relabelled numeration** $1, \ldots, 21$.

The behavior of each user (i), $(i = 1, \ldots, 21)$ is characterized by his three command profiles

$$\{\mathbf{v}_i^I,\ \mathbf{v}_i^{II},\ \mathbf{v}_i^{III}\},$$

each profile evaluated from 5000 commands issued sequentially by that user. We ask now: Are the 3 profiles of the given user similar each other, or - in other words - consistent? To answer this question, we have computed for each user three correlation coefficients between his profiles: $(\mathbf{v}_i^I,\ \mathbf{v}_i^{II})$, $(\mathbf{v}_i^I,\ \mathbf{v}_i^{III})$, $(\mathbf{v}_i^{II},\ \mathbf{v}_i^{III})$ and obtained three coefficients r_{12}, r_{13}, r_{23}, also their average $\bar{r}$.

All they appeared positive, quite large, most of them above 0.75. The three largest averages of $\bar{r}$ are: 0.991, 0.992, 0.988 (for users No.s 2, 3, 4). Five users behave inconsistently: the averages of their triplets r_{12}, r_{13}, r_{23} is lower than 0.7, as shown below:

No.	r_{12}	r_{13}	r_{23}	$\bar{r}$
6	0.943	0.460	0.613	0.672
10	0.531	0.258	0.511	0.433
13	0.663	0.413	0.825	0.634
16	0.697	0.488	0.758	0.648
19	0.986	0.322	0.326	0.545

Summarizing these results: As might be expected, most of the uncontaminated users behave consistently during the entire investigated period of their activity. This sounds optimistic. But there are some of them (e.g. No. 6, 10, 13, 16, 19 listed above) which behave differently, what is seen from their triplets r_{12}, r_{13}, r_{23}. Here one might rise doubts whether the respective command profiles were really computed for data coming from the same user.

3.4 Drawing Lines Delimiting Outliers in Block Profiles

We wish to draw a line separating 5% of the largest distances $d^2(i)$, $(i = 1, \ldots, 50)$ evaluated using formula (2). The delimiting line will be established for each user on the basis of blocks 1-50 of his data (constituting part I of his data). The theoretical distribution of the $d^2(i)$-s is unknown, therefore we will relay on

empirical distribution. We will seek for the 0.05-th upper empirical quantile of the distribution of the $d^2(i)$-s. There are together 50 blocks constituting part I of the data of each user. The simplest way to find the sought quantile is:
• order all the distances in ascending order $\{ d^2_{(j)}\}$, $(j = 1, \ldots, 50)$
• compute the sought 0.05 quantile $q_{0.05}$ as

$$q_{0.05} = (d^2_{(47)} + d^2_{(48)})/2. \tag{4}$$

The quantile $q_{0.05}$ may be also obtained by bootstrap, which is a more sophisticated method, however also more time consuming. In this paper have used the simple formula (4), which provided reasonable results.

For the given data, assuming alpha=0.05, the expected number of distances surpassing the derived quantile $q_{0.05}$ is $50\times0.05 = 2.5$. For a new sample of N blocks the number of distances surpassing the derived value $q_{0.05}$ is approximately binomial $b(x, p, N)$ with $x = 0, 1, 2, \ldots, N$. Substituting $p = 0.05,\ N = 50$, we obtain the following probabilities of appearing x=0, 1, ... items (squared distances) beyond the established limit:

x	0	1	2	3	4	5	6	7	8	9
p(x)	0.0769	0.2025	0.2611	0.2199	0.1360	0.0658	0.0260	0.0086	0.0024	0.0006

The most probable number of items exceeding the established boundary $q_{0.05}$ is: 2, 3, 4. This concerns one part of the data, containing 50 blocks of commands altogether.

So far we have analyzed the data sequences for the 21 users which were claimed to be pure, that is non contaminated by alien blocks. Each user was considered in his own space of commands which was of dimension 74–192 (see Table 1, part b). We proceed now to have a global look at the 21 users by creating a holistic view of all the 21 users by Kohonen's self-organizing map.

4 A Holistic View of the 21 Non Contaminated Users

Now we will construct a common incidence matrix for all the 21 non contaminated users. To do it, we have to find the set of unique commands issued by all these 21 users. The cardinality of this set equals u=598. Using this information, we compute for each user his three command profiles and put them together into a command profiles matrix $\mathbf{V_{63\times598}}$. For ease of interpretation, the 3 profiles for each user are labelled firstly by the # of the user (i=1, ..., 21) and next by one of the letters *a, b, c* indicating from which part of the data the given profile was constructed (*a* symbolizes Part I containing blocks 1-50; *b* – Part II with blocks 51-100; *c* – Part III with blocks 101-150) The data matrix $\mathbf{V}$ was standardized statistically ('by var'), to have mean = 0 and variance = 1, and served as input to the Matlab Som Toolbox [14], which produced – upon our request – two maps shown in Fig. 3.

The map consists of 7×6=42 hexagons corresponding to 42 topographic regions (Voronoi regions) of the elaborated 598-dimensional command space. The

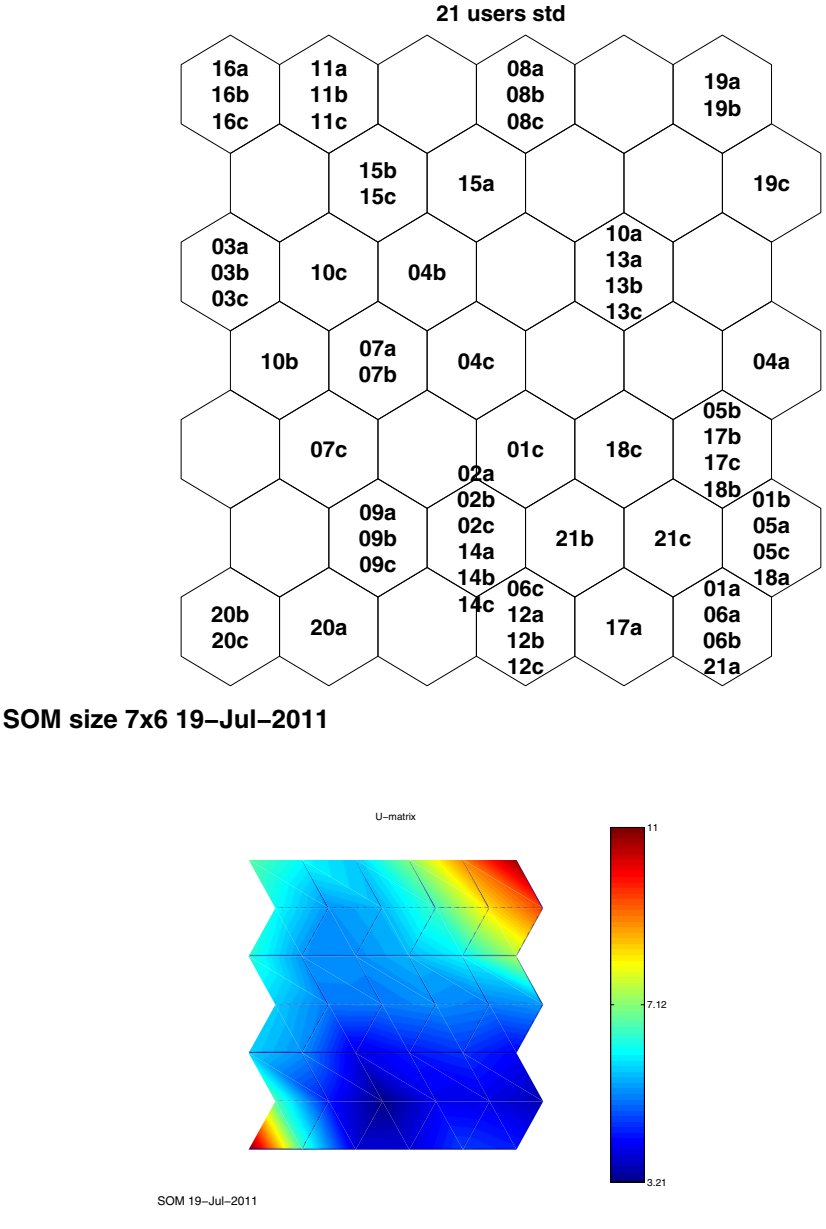


Fig. 3. Kohonen's map (size: 7×6) for standardized data - from 63 command profiles obtained for 21 non contaminated users. The upper map shows the similarity of command usage by users 1, ..., 21 in three periods (a, b, c) of their working time. Notice the non consistency in behavior of user #10. The *umati* map at bottom (see the on-line version in color) shows the topological distances of the nodes in the considered 598-dimensional space of commands. It appears that only 2 users (#19 and #20) are sticking out and are far away of the main bulk of the data.

map was constructed in such a way that the topology of the regions in the data space and on the map is preserved as much, as possible (see [13] how to do it). For our data the final topographical error is equal to 0.063 (see [13,14] for definitions), so we got in the map quite a good representation of location of the users in the data space.

The upper map in Fig. 3 shows shows a clustering of users (No.s 1 – 21) and parts a, b, c of their data in the 598-dimensional space of different (unique) commands employed by the 21 users. One may notice that the profiles a, b, c belonging to the same user are most frequently located in the same hexagon (eventually in neighboring hexagons), which means that data for the given users are consistent. One may observe also that the profiles of user #10 (with the smallest average correlation coefficient $\bar{r} = 0.43$) belong to 3 different nodes, one of them separated by two cells from the others. Each hexagon is represented by so called code-book vector having a dual representation: as the center of the hexagon on the map (map 'node') and as a prototype vector in the 598-dimensional data space (located in the corresponding Voronoi region).

The bottom map in Fig. 3 shows the true distances between the centers of the neighboring Voronoi regions – by interpolated colors between the distances among the map nodes (the umati technique [14]). The magnitude of the distances may be read up from a color bar (see the on-line version in color). One may notice that dark blue color corresponds to very near centers of Voronoi regions, and dark red indicates big distances. One may notice two distant points-users: no 19 and no 20 (outliers). They are located in opposite corners of the map. This means that they use specific sets of commands. The three parts of their data are relatively consistent: they are located in neighboring nodes.

5 Discussion and Closing Remarks

We have proposed a simple and quick method to investigate the behavior of a legitimate user working in a network. Generally, the proposed methods work well; there is also an indication that they may detect a major part of the implants hidden in the data of the remaining 29 users (this needs independent investigation). It was also stated that several (perhaps four or five) of the 'pure' users change their behavior with elapsing time.

So far we have analyzed the data sequences for the 21 users which were claimed to be pure, that is non contaminated by alien blocks. Firstly, each user was considered in his own space of commands, which was of dimension 74–192 (see Table 1, part b). We stated that 5 of these users exhibit a greater inconsistency between the three parts of their data, which is visible both in their command and block profiles. Next we have gathered all the command profiles into one big matrix $\mathbf{V}$ of size 63×598, where 598 is the number of unique commands employed by the 21 users. Each user was represented by his three profiles denoted as a, b, c. A holistic display of all the users and parts of their data was obtained in a Kohonen's self-organizing map. The views in the map confirmed our previous observation: most of the users behave consistently during time of their work in the network, which may be captured using the proposed command and block profiles.

A much more appealing problem is to apply the proposed methodology to data contaminated by alien blocks. This is the topic for an independent elaboration, which was not completed so far.

We end with two quotations from the survey paper [11]: "Insider threat is a nascent research field ... Building effective systems for detecting insider attacks remains an open challenge."

References

1. Schonlau, M.: Masquerading used data, web page, http://www.schonlau.net
2. Schonlau, M., et al.: Computer intrusion: detecting masquerades. Statistical Science 16, 1–17 (2001)
3. Bartkowiak, A.M.: Anomaly, novelty, one-class classification: a comprehensive introduction. International Journal of Computer Systems and Industrial Management Applications 3, 061–071 (2011), http://www.mirlabs.net/ijcisim/index.html
4. Kim, H.-S., Cha, S.-S.: Empirical evaluation of SVM-based masquerade detection using UNIX command. Computers & Security 24, 160–168 (2005)
5. Guan, X., Wang, W., Zhang, X.: Fast intrusion detection based on non-negative matrix factorization model. J. of Network and Computer Applications 32, 31–44 (2009)
6. Wang, W., Guan, X., Zhang, X.: Processing of massive audit data streams for real-time anomaly intrusion detection. Computer Communications 31, 58–72 (2008)
7. DiGesu, V., LoBosco, G., Friedman, J.H.: Intruders pattern identification, pp. 1–4. IEEE (2008) 978-1-4244-2175-6/08 ©2008
8. Sodiya, A.S., Folorunso, O., Onashoga, S.A., Ogunderu, O.P.: An improved semi-global alignement algorithm for masquerade detection. Int. J. for Network Security 13, 31–40 (2011)
9. Bertacchini, M., Fierens, P.I.: Preliminary results on masquerader detection using compression based similarity metrics. Electronic Journal of SADIO 7(1), 31–42 (2007), http://www.dc.uba.ar/sadio/ejs
10. Posadas, R., Mex-Perera, C., Monroy, R., Nolazco-Flores, J.: Hybrid Method for Detecting Masqueraders Using Session Folding and Hidden Markov Models. In: Gelbukh, A., Reyes-Garcia, C.A. (eds.) MICAI 2006. LNCS (LNAI), vol. 4293, pp. 622–631. Springer, Heidelberg (2006)
11. Salem, M.B., Hershkop, S., Stolfo, S.J.: A survey of insider attack detection research. In: Insider Attack and Cyber Security: Beyond the Hacker, pp. 69–90. Springer, Heidelberg (2008)
12. Bartkowiak, A.: Outliers in biometrical data: what's old, what's new. Int. J. of Biometrics 2(1), 2–18 (2010)
13. Kohonen, T.: Self-organising maps. Springer, Heidelberg (1995)
14. Vesanto, J., Himberg, J., Alhoniemi, E., Parhankangas, J.: SOM Toolbox for Matlab 5. Som Toolbox team, Helsinki University of Technology, Finland, Libella Oy, Espoo, 1–54 (2000), http://www.cis.hut.fi/projects/somtoolbox/

Songs Authentication through Embedding a Self Generated Secure Hidden Signal (SAHS)

Uttam Kr. Mondal[1] and Jyotsna Kumar Mandal[2]

[1] Dept. of CSE & IT,
College of Engg. & Management, Kolaghat,
Midnapur (W.B), India
[2] Dept. of CSE,
University of Kalyani,
Nadia (W.B), India
uttam_ku_82@yahoo.co.in, jkm.cse@gmail.com

Abstract. In this paper, an algorithm has been proposed to provide security to digital songs through amplitude modulation along with generating a secure hidden signal with an authenticating code without affecting its audible quality. Generating the hidden authenticating signal with the help of amplitude modulation for selected phases of song signal is the first phase of proposed technique followed by fabrication of authenticating code using higher frequencies above audible range. The embedded hidden secure signal as well as authenticating code will use to detect and identify the original song from similar available songs. A comparative study has been made with similar existing techniques and experimental results are also supported with mathematical formula based on Microsoft WAVE (".wav") stereo sound file.

Keywords: Average absolute difference (AD), maximum difference (MD), mean square error (MSE), normalized average absolute difference (NAD), normalized mean square error (NMSE), song authentication, songs authentication through embedding a self generated secure hidden signal (SAHS).

1 Introduction

Today's creative organizations are facing competitive market for spreading business and holding market. Creating a quality product involved a lot of investment as well as money. People are finding easier way to put less effort or investing money and producing product for existence in this contemporary market. Some of them are applying technology for making piracy of original versions and producing lower price products. This intension is a frequent phenomenon for digital audio/video industry with improvement of digital editing technology [4]. Even, it is quite harder to listeners to find the original from pirated versions. Therefore, it is a big challenge for business person, computer professional or other concern people to ensure the security criteria of originality of songs [1, 2] and protect from releasing the duplicate versions.

In this paper, a framework for protecting originality of a particular song with the help of unique secret code obtained through amplitude modulation and generating a

N. Chaki and A. Cortesi (Eds.): CISIM 2011, CCIS 245, pp. 305–313, 2011.

secure hidden signal with other authenticating code without changing its audible quality has been presented. Separating amplitudes and phases of song signal and generating the hidden authenticating signal with the help of amplitude modulation for selected phases of song signal is the first phase of proposed technique followed by fabrication of authenticating code using higher frequencies above audible range. Embedded hidden secure signal as well as authenticating code can easily distinguish the original from similar available songs. It is experimentally observed that added or subtracted extra values will not affect the song quality but provide a level of security to protect the piracy of song signal.

Organization of the paper is as follows. Embedding secret magnitude values and selecting a specific magnitude and covering the total audible range of the song are presented in section 2.1. The extraction embedding cover signal is performed in section 2.2. Experimental results are given in section 3. Conclusions are drawn in section 4. References are given at end.

2 The Technique

The scheme fabricates the secure hidden signal with help of amplitude modulation followed by generating authenticating code. Algorithms namely SAHS - AM and SAHS - SRS are proposed as double security measure, the details of which are given in section 2.1 and section 2.2 respectively.

2.1 Amp-Modulating of Song Signal (SAHS - AM)

Modulating the amplitude of song signal in adaptive manner to produce another secure signal is fabricated with the help of selected phases of song signal. The procedure of generating secret signal is depicted in the following algorithm.

Algorithm:
Input: Original song signal.
Output: Generating a secure signal for song authentication.
Method: Separating amplitude and phase of original song and modulating the amplitude values are depicted in the figure 1 and details as follows.

Step 1: Separating amplitude and phase from original song signal produces two signals representing amplitude and phase characteristics of it respectively.

Step 2: Modulation of amplitude signal by using adaptive quantization technique with fixed quantization step (let, fixed quantization level is Δ which is the average magnitude value in the range of 1-200 Hz, this range generally is not recognized by audio system), i.e., every position and its consecutive position the separation would be $\pm\Delta$. Therefore, the difference should be $\pm\Delta$ for each consecutive positions.

Step 3: Find the ratio of phase and amplitude signal, generate the ratio signal R(n) which will represent the ratio of the phase and amplitude values. Hence, the difference of two consecutive amplitude values is $\pm\Delta$, therefore, the difference between the maximum and the minimum would be less.

Step 4: Apply reverse process of step 1 to get the modified song signal back.

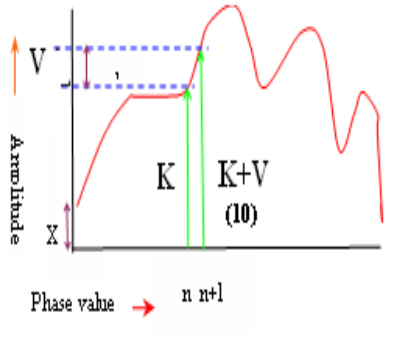


Fig. 1. Sampling of ratio signal

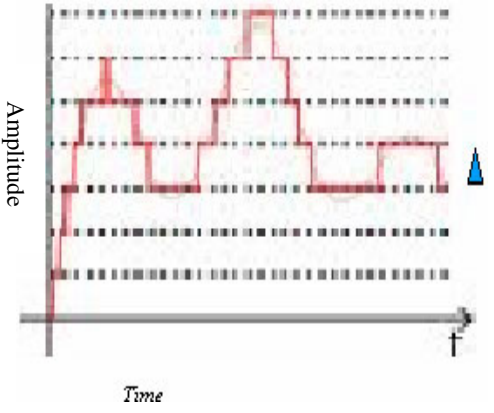


Fig. 2. Modulation of amplitude of the signal

2.2 Sampling of Ratio Signal (SAHS - SRS)

The sampling of ratio signal of the proposed technique is given below.

Algorithm:
Input: Sample values of ratio signal
Output: Modified song signal with hidden authenticating secret code in higher frequencies.
Method: Procedure of sampling of ratio signal is depicted in the figure 2 and details as follows.

Step 1: Let, X is the amplitude value of 1st position of the generated ratio signal; put the X in the higher frequency position at 20,002Hz (beyond human audible capability).

Step 2: Though the amplitude variation for consecutive position of ratio signal is constant, therefore, it would be incremented or decremented by a fixed value V (say). Now, represent incremented V and decremented V by 1 and 0 respectively. Therefore, if the amplitude values of consecutive positions are K+V, K+V,V,V, K+V,V, K+V,V,V,V, K+V,V,V, K+V,V, K-V, V,V, K+V V,V, K+V, V,.. (where K is amplitude value of previous positions)[figure 2] then, it will be represented by 10, 10, 00, 00, 10, 00, 10, 00, 00, 00, 10, 00, 00, 10, 00, 01, .. respectively.

Step 3: Take the represented values as a group of 2 and put together side by side to represent a real number as follows 0.1010, 0, 0.1, … (from the value representation of step 2 [in sequence]).

Step 4: Add the represented real numbers in the next available position of higher frequencies (above 20,000 Hz [in sequence]) which is next to X value.

2.3 Authentication

The authentication algorithm of the proposed technique is given below.

Algorithm:
Input: Modified song signal with embedded secret code.
Output: Modified song signal with authenticating secret code in higher frequencies.
Method: Adding secret code in specified way over song signal is provided the unique authentication of song signal.

Step 1: Find average value in the range of 1 to 200 Hz and put in frequency position of 20,001 Hz which is the value of quantization level of amplitude signal.
Step 2: The average value (step 1), the value of X and consecutive V value (using SAHS – SRS) will be used for identifying unique properties of original song without hampering its audible quality.

Therefore, if there is any changes occur during processing, it will create a difference with the authenticating codes that present in the higher frequencies region of the song signal and changing a position will create difference with the hidden code in that region. In the case of stereo type song the whole process will be repeated for 2^{nd} channel of song signal.

2.4 Extraction

The decoding is performed using similar mathematical calculations. The algorithm of the same is given below.

Algorithm:
Input: Modified song signal with embedded authenticating code in higher frequency range.
Output: Original song signal.
Method: The details of extraction of original song signal are given below.

Step 1: Apply FFT over x to get magnitude values in frequency domain of song signal, says Y(n), n is the total range of frequencies of song signal.
Step 2: Remove all the secret codes from higher frequencies region (above 20,000Hz).
Step 3: Apply inverse FFT to get back the sampled values of original song signal.

3 Experimental Results

Encoding and decoding technique have been applied over 10 seconds recorded songs, the song is represented by complete procedure along with results in each intermediate step has been outlined in subsections 3.1.1 to 3.1.3. The results are discussed in two sections out of which 3.1 deals with result associated with SAHS and that of 3.2 gives a comparison with existing techniques.

3.1 Results

For experimental observation, strip of 10 seconds classical song (‘100 Miles From Memphis’, sang by Sheryl Crow) has been taken. The sampled value of the song is given in table 1 as a two dimensional matrix. Figure 1 shows amplitude-time graph of the original signal. The output generated in the process is shown in figure 3 (number of sampled values is 441000). Figure 4 shows the generated amplitude signal [Figure 4(a)], phase signal [Figure 4(b)] and ratio signal [Figure 4(c)] respectively. Figure 6 shows the difference of frequency ratio of original and modified song with authenticating code. From figure 6 it is seen that the deviation is very less and there will not affect the quality of the song at all.

3.1.1 Original Recorded Song Signal (10 seconds)

The values for sampled data array x(n,2) from the original song is given in table 1. Whereas the graphical representation of the original song, considering all rows (441000) of x(n,2) is given in the figure 3.

Table 1. Sampled data array x(n,2)

Sl no	x(k,1)	x(k,2)
…	…	…
	0	0.0001
	0.0000	0.0000
	-0.0009	-0.0009
	-0.0006	-0.0007
	-0.0012	-0.0012
	-0.0014	-0.0014
	-0.0016	-0.0017
	-0.0023	-0.0022
	-0.0027	-0.0027
	-0.0022	-0.0021
…	…	…

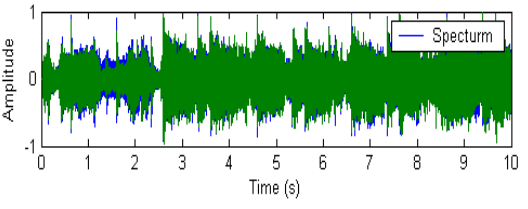
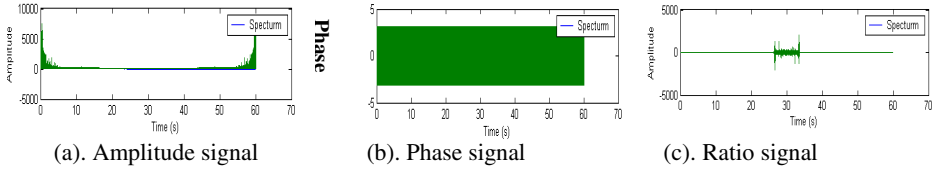


Fig. 3. Original song (‘100 Miles From Memphis’, sang by Sheryl Crow)



(a). Amplitude signal (b). Phase signal (c). Ratio signal

Fig. 4. Component signals

3.1.2 Modified Song After Modulating Amplitude and Adding Authenticating Code (10 seconds)

The graphical representation of the modified song signal is shown in the figure 5.

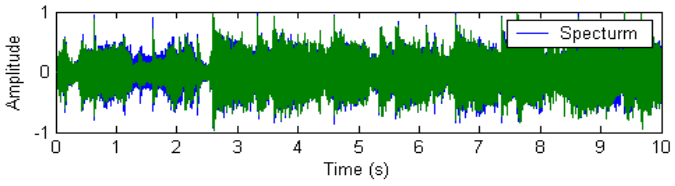


Fig. 5. Modified song after modulating amplitude and adding authenticating code

3.1.3 The Difference of Magnitude Values Between Original Signal and Modified Signal

The graphical representation of the magnitude differences of original and modified songs is shown in the figure 6.

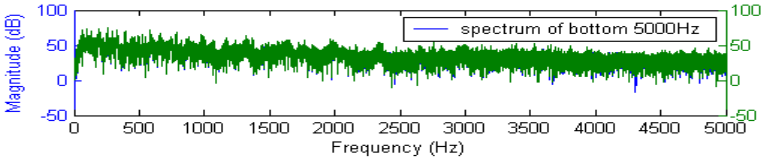


Fig. 6. The difference of magnitude values between signal figure 3 and figure 5

3.2 Comparison with Existing Systems

Various algorithms [5] are available for embedding information with audio signals. They usually do not care about the quality of audio but we are enforcing our authentication technique without changing the quality of song. A comparative study of properties of our proposed method with Data Hiding via Phase Manipulation of Audio Signals (DHPMA)[3] before and after embedding secret message/modifying parts of signal (16-bit stereo audio signals sampled at 44.1 kHz.) is given in table 2, table3 and table4. Average absolute difference (AD) is used as the dissimilarity measure between original and modified song to justify the modified song. Whereas a lower value of AD signifies lesser error in the modified song. Normalized average absolute difference (NAD) is quantization error is to measure normalized distance to a range between 0 and 1. Mean square error (MSE) is the cumulative squared error between the embedded song and the original song. A lower value of MSE signifies lesser error in the embedded song. The SNR is used to measure how much a signal has been tainted by noise. It represents embedding errors between original song and modified song and calculated as the ratio of signal power (original song) to the noise power corrupting the signal. A ratio higher than 1:1 indicates more signal than noise. The PSNR is often used to assess the quality measurement between the original and a modified song. The higher the PSNR represents the better the quality of the modified song. Thus from our experimental results of benchmarking parameters (NAD, MSE, NMSE, SNR and PSNR) in proposed method obtain better performances without affecting the audio quality of song.

Table 3 gives the experimental results in terms of SNR (Signal to Noise Ratio) and PSNR (Peak signal to Noise Ratio). Table 4 represents comparative values of Normalized Cross-Correlation (NC) and Correlation Quality (QC) of proposed algorithm with DHPMA. Table 5 shows PSNR, SNR, BER (Bit Error Rate) and MOS (Mean opinion score) values for the proposed algorithm. Here all the BER values are 0. The figure 7 summarizes the results of this experimental test. It is seen from the results that the performance of the algorithm is stable for different types of audio signals.

Table 2. Metric for different distortions

Sl No	Statistical parameters for differential distortion	Value using SAHS	Value using DHPMA
1	MD	0.4476	3.6621e-004
2	AD	0.0016	2.0886e-005
3	NAD	0.0133	0.0063
4	MSE	5.9602e-006	1.4671e-009
5	NMSE	4.2810e+003	8.4137e-005

Table 3. SNR and PSNR

Sl No	Statistical parameters for differential distortion	Value using SAHS	Value using DHPMA
1	Signal to Noise Ratio (SNR)	36.3154	40.7501
2	Peak Signal to Noise Ratio (PSNR)	52.0750	45.4226

Table 4. Representing NC and QC

Sl No	Statistical parameters for correlation distortion	Value using SAHS	Value using DHPMA
1	Normalised Cross-Correlation(NC)	1	1
2	Correlation Quality (QC)	-0.1144	-0.5038

Table 5. Showing SNR, PSNR BER, MOS

Audio (1s)	SNR	PSNR	BER	MOS
Song1	36.3154	52.0750	0	5
Song2	24.7558	34.3455	0	5
Song3	38.0870	58.6785	0	5
Song4	27.4947	41.7886	0	5
Song5	19.5848	31.3647	0	5

Quality rating (Mean opinion score) of the proposed technique has been computed by using equation (1).

$$Quality = \frac{5}{1 + N * SNR} \tag{1}$$

where N is a normalization constant and SNR is the measured signal to noise ratio. The ITU-R Rec. 500 quality rating is perfectly suited for this task, as it gives a quality rating on a scale of 1 to 5 [6]. Table 6 shows the rating scale, along with the quality level being represented.

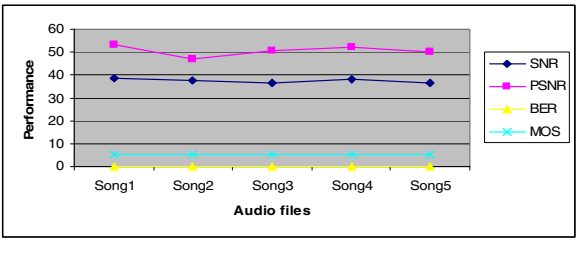


Fig. 7. Performance for different audio signals

Table 6. Quality rating scale

Rating	Impairment	Quality
5	Imperceptible	Excellent
4	Perceptible, not annoying	Good
3	Slightly annoying	Fair
2	Annoying	Poor
1	Very annoying	Bad

4 Conclusion and Future Work

In this paper, an algorithm for generating the hidden authenticating signal with the help of amplitude modulation for selected phases of song signal and embedding secret code in the higher frequencies region of song signal has been proposed which will not affect the song quality but it will ensure to detect the distortion of song signal characteristics. Additionally, the proposed algorithm is also very easy to implement.

This technique is developed based on the observation of characteristics of different songs but the mathematical model for representing the variation of those characteristics after modification may be formulated in future. It also can be extended to embed an image into an audio signal instead of numeric value. The perfect estimation of percentage of threshold numbers of sample data of song that can be allow to change for a normal conditions will be done in future with all possibilities of errors.

References

1. Mondal, U.K., Mandal, J.K.: A Practical Approach of Embedding Secret Key to Authenticate Tagore Songs (ESKATS). In: Wireless Information Networks & Business Information System Proceedings (WINBIS 2010), vol. 6(1), pp. 67–74. organized by Rural Nepal Technical Academy (Pvt.) Ltd., Nepal (2010)
2. Mondal, U.K., Mandal, J.K.: A Novel Technique to Protect Piracy of Quality Songs through Amplitude Manipulation (PPAM). In: International Symposium on Electronic System Design (ISED 2010), pp. 246–250 (2010) ISBN 978-0-7695-4294-2
3. Xiaoxiao, D., Bocko, M.F., Ignjatovic, Z.: Data Hiding Via Phase Manipulation of Audio Signals. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2004), vol. 5, pp. 377–380 (2004) ISBN 0-7803-8484-9
4. Erten, G., Salam, F.: Voice Output Extraction by Signal Separation. In: ISCAS 1998, vol. 3, pp. 5–8 (1998) ISBN 07803-4455-3
5. Katzenbeisser, S., Petitcolas, F.A.P.: Information Hiding Techniques for Steganography and Digital Watermarking. Artech House, Norwood (2000) ISBN 978-1-58053-035-4
6. Arnold, M.: Audio watermarking: Features, applications and algorithms. In: IEEE International Conference on Multimedia and Expo., New York, NY, vol. 2, pp. 1013–1016 (2000)

Secure Money Transaction in NFC Enabled Mobile Wallet Using Session Based Alternative Cryptographic Techniques

Riti Chowdhury and Debashis De

Department of Computer Science and Engineering, West Bengal University of Technology, BF-142, Sector – I, Salt Lake City, Kolkata-700064, India
dr.debashis.de@gmail.com

Abstract. Mobile wallet is very useful for day to day money transaction. It is an essential requirement to build utmost level of security. NFC technology introduces a new gateway to enable secure mobile wallet money transaction providing direct point to point device communication. This paper proposes a method to support alternative cryptographic algorithms for each new transaction and each of these transactions are based on a session with different cryptographic key generated for each session. Information used during mobile transaction resides in the cryptographic SIM card. The point to point mobile money transaction in NFC enabled mobile wallet is carried out over a secure channel using WAP.

Keywords: Mobile Wallet, NFC Technology, Cryptographic SIM card, WAP.

1 Introduction

Mobile phones are getting smarter providing a wide range of services and applications over the wireless network which are bringing the globe in front of the mobile phone owner. If someone wants to access the exact fare or make a payment with discount, the calculations become very complex and time consuming. Again the worst case happens when someone loses physical wallet and needs to remember which cards the physical wallet contained and manually cancel each one, and again has to apply for the new cards. This problem can be solved by replacing the physical wallet with a digital wallet integrated into an existing mobile device like a cell phone. Unfortunately the intruders are becoming more powerful by modifying, intercepting, and fabricating user's private and/or financial data applying several techniques and the mobile phones leads to out of service. The effect is awful in case of m-commerce but we can solve the problem using the strategy proposed in this paper. Now a days NFC technology is very popular one and can be used with a session based alternative cryptographic algorithm strategy to provide security in mobile phone based transaction with other mobile phone or with the shops that contains PoS. The Cryptographic SIM [3] [4] card plays an important role to protect confidential information and keys needed for this purpose. This strategy makes the mobile phone money transaction secured, protecting transaction information against man-in-middle attack, interception, modification and fabrication [8].

N. Chaki and A. Cortesi (Eds.): CISIM 2011, CCIS 245, pp. 314–323, 2011.

2 Related Work

The recent work has been made in different area of NFC enabled digital wallet. NFC devices work in a peer to peer communication fashion which results more secure money transaction facility using NFCIP [13]. The one time password [5], [6] introduces in way of safe money transaction. It makes a strong effect against phishing attack and also unauthorized access of user data. The cryptographic SIM card [3], [4] is also introduced to store user's confidential data and the secret keys related to encryption technology. Secure channel through which transaction leads can be achieved by using different protocol. Here we have used Wireless Application Protocol (WAP) providing secure channel .The cryptographic algorithm for secure money transaction has been also introduced. Different method has been developed for NFC enabled secure money transaction has been developed [9] [10] [12].

3 NFC Background

3.1 NFC Technology

NFC is a set of short-range wireless technologies, requiring a distance of 20 cm or less but works well within the range of 10 cm or less [2]. NFC operates at 13.56 MHz and at data rate ranging from 106 Kbit/s to 848 Kbit/s [1]. NFC involves an initiator and a target; the initiator generates a Radio Field that has the capability to power a passive target.

3.2 Advantages of NFC

NFC is a wireless communication technology that supports device to device interaction within a few centimeters without data hacking or any distortion. This technology can be used for the money transaction over the wireless mobile phones or with any PoS. NFCIP is a mode of NFC that supports the peer to peer communication between initiator and target in two modes [13]. In active mode the initiator and target both are responsible for generating RF field, but in passive mode only the initiator generates RF field. NFCIP works while sending some data to which target responds. The aforesaid fact is depicted in Fig1.

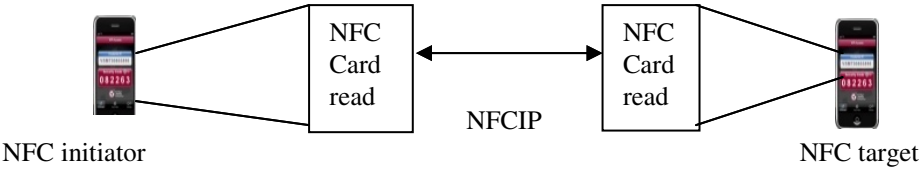


Fig. 1. NFC Communication for direct device to device transaction

4 Secure NFC Communication

The NFC enabled mobile phones are very useful for direct device to device money transaction. This concept can be implemented in case of mobile wallet. This paper proposes an innovative method of using alternative cryptographic algorithms. The cryptographic key for the selected algorithm is changed in each new transaction based on a session. The banks that support mobile wallet money transaction facility are needed to make an agreement on all possible algorithms that will be used for direct mobile to mobile money transaction. The mobile wallet should facilitate the algorithms among the standard list negotiated by the banks. The cryptographic manager contains the algorithms among the decided list only and is embedded with the operating systems used by mobile phones.

4.1 Alternative Cryptographic Algorithm Concept

The bank server contains a list of cryptographic algorithms standardized previously; say CAL. Mobile Os contains some or all of these algorithms CA1, CA2, and CA3…. CAn from the CAL All these algorithms reside in the cryptographic manager, which are alternatively selected for each money transaction Ti. Suppose T (i) is a transaction that selects a specific CAi where i=1|2|……|n and CAi belongs to any of the algorithm in CAL used by the mobile wallet. When the mobile wallet user goes for the next transaction T (i+1), CAi is altered to the newer one i.e. CAj. CAj is any algorithm in CAL except CAi. Transaction authority (TA) has a fixed list of encryption algorithms that are previously negotiated in its secure storage. For each transaction requested by a Mobile wallet these algorithms alter. Trusting authority (TA) uses a random Cryptographic algorithm generator function Ranf () which randomly selects cryptographic algorithm for each session Ssn. The Ranf () function selects CA such that CAi! =CAj where Ti and Tj are two consecutive transactions. When a user requests for the transaction TA runs Ranf () and according to the output of Ranf () the cryptographic algorithm is selected. The information about the output is sent to the digital wallet using a secure channel to protect data from interception or information hacking on the fly. A person who is trying to observe the data over the network for a long time will be unable to trace the algorithm as it is getting changed for each new money transaction. For each new money transaction Ti the Cryptographic key Ki key is changed to the newer one according to the algorithm CAi selected for this Ti request. As a result it is quite difficult to suspect the data used in the transaction moving over a secure channel in a wireless environment. All the keys are maintained in a secure storage within the bank server for an individual transaction session. DES, AES, ECC IDEA and RC5 are the cryptographic algorithms that can be used for this purpose. The secure communication channel is made using Wireless Transport Layer Security (WTLS) and Wireless Transaction Protocol (WTP). Cryptographic algorithm negotiation using random cryptographic function Ranf () is depicted in the Fig.2.

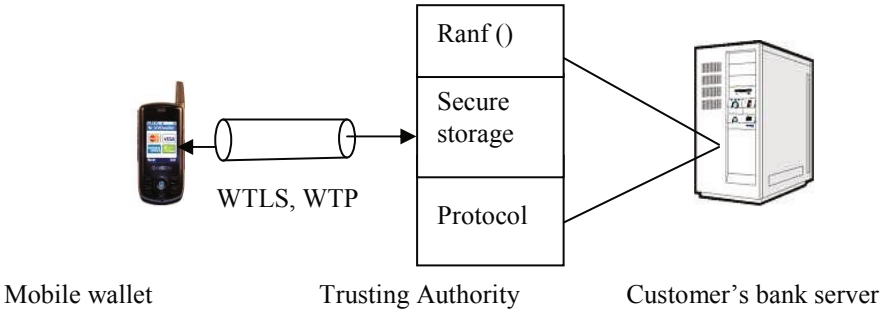


Fig. 2. Cryptographic algorithm negotiation using random crypto function Ranf

5 Operational Overview

5.1 Session Management

When a user makes a transaction request to its issuer bank, first a session Ssn is created for that specific transaction Ti by the bank server. To fulfill this purpose a global clock is required to maintain synchronization in each different operational element (OE) i.e. Mobile wallet, Customer's bank, Merchant's bank, and Point of Sell. Maintaining global clock by different distributed operational element is difficult. This paper suggests an idea for synchronizing each operation that is running over geographically distributed areas across the network. Ssn is the time needed to perform the whole transaction and Ssnlft is the remaining time available for the transaction after any operation is performed by an operational element. Here only the customer's bank server takes the responsibility to calculate Ssn and Ssnlft for each operation. After each request, response and acknowledgement the customer's bank server receives, it calculates the Ssnlft and sends Ssnlft value to the other elements that are going to perform the next operation so on.

Notations.

Ssn: Maximum Allowable Session for Each Transaction i.e. the Maximum allowable time duration needed to complete a money transaction.
OETi: Operational Execution Time for i th operation.

Session Management Algorithm.

Ssn is determined by the customer's bank server. This is a predetermined fixed valued attribute. This can be used to restrict an intruder trying to collect information of user's credentials. Maximum allowable OETi value is kept in the secure storage element of the customer's bank server. For each new communication occurs between the OEs, the OE that is going to perform the next operation sends current OETi value

needed to the customer's bank server .Customer's bank server then checks and confirms the operation to be executed only if the required OETi value of the OE for current operation is less than or equals to that of its maximum value. Session will get time out when Ssnlft = 0. If any of the OE gets the information that Ssnlft = 0, OE discards the information and informs the Customer's bank server that the transaction Ssn is destroyed. Customer's bank server then broadcasts the session destroy message to all the OE currently involved in the transaction operation.

Session Management algorithm is performed as follows:

Ssnlft = (Ssn- OET1); where i=1 for the initialization.
For i=2 to n {
Ssnlft = (Ssnlft-OETi); while OETi is the time needed by the latest operation that has been currently executed. i=2, 3, 4, ……,n;
}
n is the maximum allowable number of operations for a transaction.
if (Ssnlft = = 0) then transaction will be stopped;
Else transaction continues;

5.2 Transaction Setup Algorithm

The user keeps closer his/her MW to the PoS with which he/she wants to perform the transaction Ti. The transaction setup algorithm initializes the transaction where PoS requests for the payment to MW and the cryptographic algorithm to be selected is decided by the TA, after receiving PReq from MW. The mobile owner has to touch and confirm the transaction in his/her MW [10] as a user request PReq for transaction and PReq goes to the trusting authority (TA). TA links with the digital wallet making a secure channel by WTSL, WTP through which further communication will be executed. TA makes a session Ssn for transaction Ti. If all the transaction operations are not completed during this time, a new request is needed to be made by MW and again a new session is created. All the data running over the wireless network is encrypted by negotiating an encryption algorithm CA. There are different encryption algorithm techniques available which are altered for each new transaction. With the transaction request MW sends to TA a list of cryptographic algorithms CAL that are supported by MW for this transaction. TA selects CA among these algorithms by executing Ranf (CAL) function and informs the selected cryptographic algorithm to MW.

Notations.

{ MW, PoS, TA, CB, MB} : Customer's mobile wallet, Point of Sell device like PoS in the shop or another NFC mobile wallet, Trusting Authority, Customer's Bank Server, Merchant's Bank server respectively.
CAL: List of Cryptographic Algorithms
PMReq: MW Payment Message Request
PoSReq: POS Payment Message Request
TRequest: Time Instant of PMReq Generation

h(x): Hash Function on the message x
PReq: Payment Request. PReq = {PMreq,h (CAL,TRequest)}
AT: Allow Transaction
DT: Disallow Transaction
MAmnt: Payment Amount
Confpm: Confirm Payment message
Ranf (CAL): Random Function for selecting the cryptographic algorithm to be used.
CAkey: Cryptographic Algorithm Key
EN(y): Encryption of the message y with the cryptographic algorithm selected

The Transaction Setup Algorithm is as follows:

1. Initialization: PoS → MW: PoSReq, MAmnt

   MW → PoS: Confpm

2. MW → TA: PReq ; PReq = h {PMReq, (CAL,TRequest)}
3. TA → MW: If For this Ssn

   TA: Ranf (CAL) {

   Op= {CA}}

   b. for each Op: CAkey = new key;

4. TA → MW: EN (AT, Ssnlft, Op, CAkey);

Direct NFC enabled device to device payment is shown in Fig.3.

Fig. 3. Direct NFC enabled device to device payment [11]

5.3 Authentication Request Algorithm

This algorithm matches customer's banking information that has been sent with the customer's secret database stored in the bank server. If TA accepts the transaction request it sends MW an acknowledgement asking for the MW authentication request AuthanReq .TA executes a random generator function Rancp () to generate one time password Cp to MW [5]. The user is asked for two factor authentications where user has some secret information i.e. x and what he is i.e. y. This information is sent to the TA using the negotiated encryption technique. TA compares the identification information given by the customer with the database of the customer stored in CB server and confirms if data are successfully matched to MW or otherwise session Ssn is

destroyed and TA send this information to MW. The other details of information such as expiry date or valid user credit balance is below also verified. If it is not proper then session Ssn is destroyed and the transaction is discarded.

Notations.

Rancp (): Random number generator for one time password.
Cp: Output of Rancp ()
CuIN(): Customer's Unique Identification number request.
CuFP():Customer's Unique Identity request ; such as fingerprint.
CuIN(x): Customer's Unique Identification number x given by Customer's Bank.
CuFP(y): Customer's Unique Identity; in this case y is considered as finger print.
AuthenReq: Customer Authentication Request.
Req={Rancp(),CuIN(),CuFP()}
AuthenRes: Customer Authentication Response.
AuthenRes = {Cp,CuIN(x),CuFP(y)}
CPIN: Customer's Personal Identification Number
PD:Payment Details. PD= {MAmnt, Maccnt, MB}
Avalamnt: Available Amount in Customer's Account
FP: Finger Print stored in database of the customer's bank;
DoE: Date of Expiry of customer's account.
Maccnt: Merchant's Account Number
MINVaL: Minimum Amount needed for an account to exists

The Authentication Request Algorithm is as follows:

1) TA → MW: EN (Authanreq)
AuthanReq= {Rancp(),CuIN(),CuFP(),Ssnlft}
MW → TA: EN (AuthenRes)
AuthanRes= {Cp,CPIN(x), CuFP(y),OETi}
2) TA → CB: Check (Cp,CPIN,FP){
if TA → (Cp,PIN,FP) = CB → (Cp,CPIN(x), CuFP(y))
Oau=matched;
TA → MW: message of successful customer's authentication with Ssnlft;
MW → TA: EN (PD): PD=(MAmnt,Maccnt,MB,OETi)
TA → CB: ssnlft;for valid ssnlft
If (avalamnt>(MAmnt+ MINVaL)||DoE!=current date);
dAmnt= Deduct MAmnt from Avalamnt;
CB → MW: EN (dAmnt,ssnlft)
Else session timeout;
}
3) Else
TA → CB: Check (Cp,CPIN,FP){
Oau = unmatched;
TA → MW: unsuccessful customer's authentication;
EXIT from transaction;
}

5.4 Money Transfer Algorithm

For a successful authentication the transaction is made and CB transfer Mamnt to MB and also deduct the value of Mamnt from CB. The money transfer algorithm transfers the payment from customer's bank account to merchant's bank account. CB sends payment amount to MB's Maccnt and MB sends successful transaction completion message to PoS and PoS displays successful payment message to its screen.

The money transaction algorithm is as follows:

1) CB → MB: EN (Maccnt, dAmnt, ssnlft)
   Message (Credit Maccnt by dAmnt)
2) MB → PoS: EN(transaction update massage with MAmnt,ssnlft)
3) PoS → MW: Payment complete message

Operational overview for NFC enabled mobile transaction in a mobile wallet is shown in Fig.4.

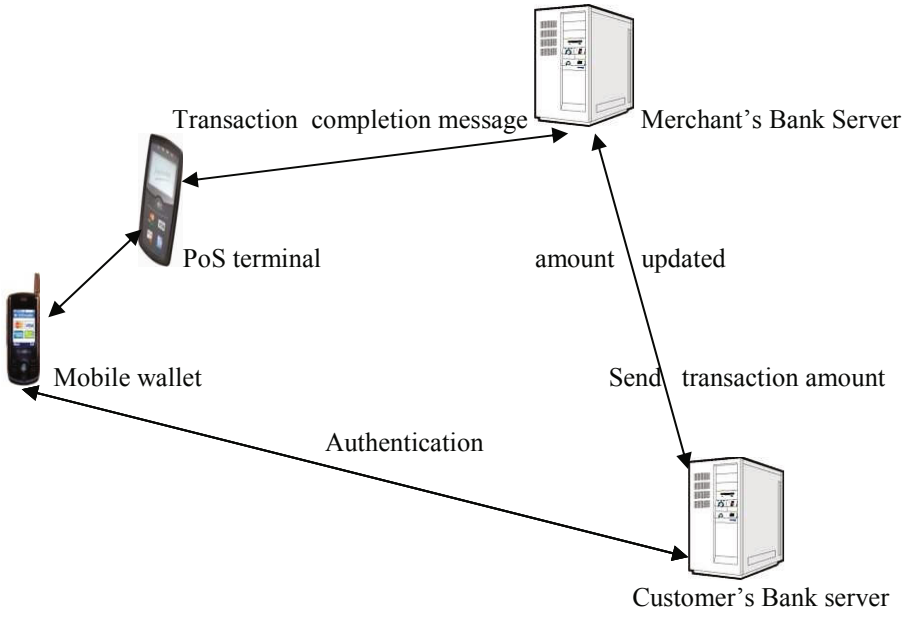


Fig. 4. Operational overview for NFC enabled mobile transaction

6 Protocol Used for the Transaction

The transaction is to be made using a secure channel. WAP is used to make a secure channel though which the information runs over the network. WTLS and WTP protocol layers make the aforesaid secure channel between OEs. WTLS operates above the transport layer and provides security in transport service interface for

secure communication. This protocol supports for optimized handshaking and dynamic key refreshing. WTLS and WTP support the cryptographic techniques for wireless mobile money transaction.

7 Cryptographic SIM Card

A new concept about the SIM card of the mobile phone is that it contains a secure storage. It can be used to store user's confidential data and/or the private data such as encryption keys, digital certificates that are needed for mobile phone transaction. The current capacity of SIM is at least 32 KB [3]. So a number of cryptographic information can be stored within the SIM card. As the SIM card has unique ISDN assigned to each card holder he/she can easily block or make the card invalid if the phone or SIM card is lost. Further transaction can also be blocked if the SIM card is lost. The SIM card of next generation will contain more memory and will be able to keep the cryptographic algorithms in SIM card [4]. So our proposed strategy works well when it comes to the mobile phone money transaction.

8 Conclusion

The alternative cryptographic algorithms based on a session concept enhance the security for money transaction in NFC mobile wallet. It enables the system with new cryptographic technologies in a flexible manner. Symmetric key cryptographic algorithms can be used for easier management and time-space minimization instead of asymmetric cryptographic algorithm. The cryptographic SIM card holds all the transaction details during a transaction session. This prevents fraud transaction even if the mobile wallet has been stolen. WTLS makes secure channel to carry out a secure transaction.

Acknowledgement. Authors are grateful to DST for the project under fast track young scientist scheme on Dynamic Optimization of Green Mobile Networks: Algorithm, Architecture and Applications having reference number SERC/ET-0213/2011 under which this paper has been completed.

Reference

1. Benyó, B., Sódor, B., Fördős, G., Kovács, L., Vilmos, A.: A generalize approach for NFC application development. In: 2nd International Workshop on Near Field Communication, pp. 45–50. IEEE Press, Monaco (2010)
2. Chen, W., Hancke, G.P., Mayes, K.E., Lien, Y., Chiu, J.-H.: NFC Mobile Transactions and Authentication Based on GSM Network. In: 2nd International Workshop on Near Field Communication, NFC, pp. 83–89. IEEE Press (2010)
3. Kálmán, G., Noll, J.: SIM as Secure Key Storage in Communication Networks. In: Proceedings of the 3rd International Conference on Wireless and Mobile Communications, pp. 55–61. IEEE Press, Guadeloupe (2007)

4. Jara, A., Zamora, M.A., Skarmeta, A.F.G.: Secure use of NFC in medical envi-ronments. In: 5th European Workshop on RFID Systems and Technologies, pp. 1–8. IEEE Press, Bremen (2009)
5. Leung, C.-M.: Depress phishing by CAPTCHA with OTP. In: International Conference on Anti-Counterfeiting, Security, and Identification in Communication, pp. 187–192. IEEE Press, Hong Kong (2009)
6. Tiwari, P.B., Joshi, S.R.: Single sign-on with one time password. In: First Asian Himalayas International Conference in Internet, pp. 1–4. IEEE Press, Kathmandu (2009)
7. Ondrus, J., Pigneur, Y.: An Assessment of NFC for Future Mobile Payment Systems. In: International Conference on the Management of Mobile Business, pp. 43–53. IEEE Press, Toronto (2007)
8. Mulliner, C.: Vulnerability Analysis and Attacks on NFC-Enabled Mobile Phones. In: Proc. International Conference on Reliability and Security, pp. 695–700. IEEE Press, Fukuoka (2009)
9. Pasquet, M., Reynaud, J., Rosenberger, C.: Secure payment with nfc mobile phone in the smarttouch project. In: Proceedings of the 2008 International Symposium on Collaborative Technologies and Systems, pp. 95–98. IEEE Press, Irvine (2008)
10. Raghuwanshi, Pateria, R.K., Singh, R.P.: A new protocol model for verification of payment order information integrity in online Epayment system. In: World Congress on Nature & Biologically Inspired Computing, pp. 1665–1668. IEEE Press, Coimbatore (2009)
11. The Transformational Power of NFC, http://www.mobimatter.com/category/topic/nfc-rfid/
12. Fun, T.S., Beng, L.Y., Likoh, J., Roslan, R.: A light-weight and Private Mobile Payment Protocol by Using Mobile Network Operator. In: Proceedings of International Conference on Computer and Communication Engineering, pp. 162–166. IEEE Press, Kuala Lumpur (2008)
13. Grunberger, S., Langer, J.: Analysis and test results of tunneling IP over NFCIP-1. In: First International Workshop on Near Field Communication, pp. 93–97. IEEE Press, Hagenberg (2009)

Author Index