

Tran Khanh Dang Roland Wagner
Erich Neuhold Makoto Takizawa
Josef Küng Nam Thoai (Eds.)

LNCS 8860

Future Data and Security Engineering

First International Conference, FDSE 2014
Ho Chi Minh City, Vietnam, November 19–21, 2014
Proceedings

FDSE 2014



Springer

Commenced Publication in 1973

Founding and Former Series Editors:

Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

David Hutchison

Lancaster University, UK

Takeo Kanade

Carnegie Mellon University, Pittsburgh, PA, USA

Josef Kittler

University of Surrey, Guildford, UK

Jon M. Kleinberg

Cornell University, Ithaca, NY, USA

Friedemann Mattern

ETH Zurich, Switzerland

John C. Mitchell

Stanford University, CA, USA

Moni Naor

Weizmann Institute of Science, Rehovot, Israel

C. Pandu Rangan

Indian Institute of Technology, Madras, India

Bernhard Steffen

TU Dortmund University, Germany

Demetri Terzopoulos

University of California, Los Angeles, CA, USA

Doug Tygar

University of California, Berkeley, CA, USA

Gerhard Weikum

Max Planck Institute for Informatics, Saarbruecken, Germany

Tran Khanh Dang Roland Wagner
Erich Neuhold Makoto Takizawa
Josef Küng Nam Thoai (Eds.)

Future Data and Security Engineering

First International Conference, FDSE 2014
Ho Chi Minh City, Vietnam, November 19-21, 2014
Proceedings

Volume Editors

Tran Khanh Dang

Nam Thoai

Ho Chi Minh City University of Technology

268 Ly Thuong Kiet Street, District 10, Ho Chi Minh City, Vietnam

E-mail: {khanh, nam}@cse.hcmut.edu.vn

Roland Wagner

Josef Küng

Johannes Kepler University Linz

Altenberger Straße 69, 4040 Linz, Austria

E-mail: {rwagner, jkueng}@faw.jku.at

Erich Neuhold

University of Vienna

Währinger Straße 29, 1190 Wien, Austria

E-mail: erich.neuhold@univie.ac.at

Makoto Takizawa

Hosei University

3-7-2, Kajino-machi, Koganei-shi, Tokyo 184-8584, Japan

E-mail: makoto.takizawa@computer.org

ISSN 0302-9743

e-ISSN 1611-3349

ISBN 978-3-319-12777-4

e-ISBN 978-3-319-12778-1

DOI 10.1007/978-3-319-12778-1

Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014952231

LNCS Sublibrary: SL 3 – Information Systems and Application, incl. Internet/Web and HCI

© Springer International Publishing Switzerland 2014

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

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

In this volume we present the accepted contributions to the First International Conference on Future Data and Security Engineering (FDSE 2014). The conference took place during November 19–21, 2014, in Ho Chi Minh City, Vietnam, at the HCMC University of Technology, the most famous and prestigious university in the south of Vietnam.

The annual FDSE conference is a premier forum designed for researchers, scientists, and practitioners interested in state-of-the-art and state-of-the-practice activities in data, information, knowledge, and security engineering to explore cutting-edge ideas, present and exchange their research results and advanced data-intensive applications, as well as to discuss emerging issues on data, information, knowledge, and security engineering.

The call for papers resulted in the submission of 66 papers. A rigorous peer-review process was applied to all of them. This resulted in 23 full accepted papers (34.8%), which were presented at the conference. Every paper was reviewed by at least three members of the international Program Committee who were carefully chosen based on their knowledge and competence. This careful process resulted in the high quality of the contributions published in this volume. The accepted papers were grouped into the following sessions:

- Big data analytics and applications
- Security and privacy engineering
- Crowdsourcing and social network data analytics
- Biometrics and data protection in smart devices
- Cloud data management and applications
- Advances in query processing and optimization

In addition to the papers selected by the Program Committee, two internationally recognized scholars delivered keynote speeches: “Data Mining on Forensic Data: Challenges and Opportunities,” presented by Professor M-Tahar Kechadi from University College Dublin, Ireland, and “Further Application on RFID with Privacy-Preserving,” presented by Professor Atsuko Miyaji from Japan Advanced Institute of Science and Technology, Japan.

In the first keynote speech, Professor M-Tahar Kechadi talked about data mining on forensic data. The abstract of the speech is briefly summarized as follows: “The success of the current ICT technologies has reached a level in which we generate huge amount of data outpacing our ability to process it and learn from it. Moreover, the future of these technologies depends heavily on the way security threats, trust, and privacy within this “cloudy” environment are dealt with. Current computing environments are exposing their customers to huge risks for their business and reputation, as attacks and cybercrimes are becoming of huge concern. Therefore, these customers need to have confidence in their providers

in following accredited security and privacy practices. For instance, over the last decade, the number of crimes that involve computers and the Internet has grown at a rapid pace and the problem of scale will only escalate the cybercrime phenomenon. Without doubt, the next generation of computing will depend highly on the way large-scale cybercrimes are dealt with, both in developing appropriate and efficient approaches to this problem and also on the amount of effort we put into tackling this phenomenon. In this presentation, we discuss how the investigators can store evidential data and conduct forensics analysis. We focus on the data mining aspect for automatically pre-processing, analyzing the data, and extracting evidence. This will involve redesigning the data mining process to deal with forensic data that are noisy, dispersive, incomplete, and inconstant, which are very challenging.”

In the second keynote speech, Professor Atsuko Miyaji discussed important issues relevant to privacy-preserving in RFID applications. The main contents of the speech are summarized as follows: “Wireless technologies are being developed rapidly to construct smart communications with digital data. Networked devices automatically communicate among themselves in order to carry out efficient information transaction. A radio-frequency identification (RFID) tag with unique identification numbers uses radio waves to transmit data at a distance. RFID is going to support new technologies such as the Internet of Things (IoT) and Machine-to-Machine (M2M). The RFID system actuates various security and privacy issues concerning its owners and holders without any knowledge of users. Preventing unauthorized access to the owner data (confidentiality), tag tracing (link-ability), identification of the owners (anonymity) are some necessary privacy-protection issues of RFID systems. A supply chain management (SCM) controls and manages all the materials and information in the logistics process from acquisition of raw materials to product delivery to the end user. It is crucial to construct protocols that enable the end user to verify the security and privacy not only of the tags but also the path that the tag passes through, which means path authentication. Path authentication is a further application of RFID. In this talk, we summarize previous works to realize authentication schemes, and then redefine a privacy notion of RFID. We present a new direction that enables the end user to verify the security and privacy not only of tags but also paths that tags go through. More clearly, if a path-authenticated tag reaches the end of its supply chain, then the path ensures that no intermediate reader was omitted (or selected wrongly) by a tag, either deliberately or not. This would yield not only convenience and efficiency of delivery but also quality of product.”

The success of FDSE 2014 was the result of the efforts of many people, to whom we would like to express our gratitude. First, we would like to thank all authors who submitted papers to FDSE 2014, especially the two invited speakers. We would also like to thank the members of the committees and external reviewers for their timely reviewing and lively participation in the subsequent

discussion in order to select such high-quality papers published in this volume. Last but not least, we thank the Faculty of Computer Science and Engineering, HCMC University of Technology for hosting FDSE 2014.

November 2014

Tran Khanh Dang
Roland Wagner
Erich Neuhold
Makoto Takizawa
Josef Küng
Nam Thoai

Organization

General Chair

Roland Wagner

Johannes Kepler University Linz, Austria

Steering Committee

Elisa Bertino

Purdue University, USA

Kazuhiko Hamamoto

Tokai University, Japan

Abdelkader Hameurlain

Paul Sabatier University, Toulouse, France

Dieter Kranzmueller

Ludwig Maximilians University, Germany

Beng Chin Ooi

National University of Singapore, Singapore

A. Min Tjoa

Technical University of Vienna, Austria

Nam Thoai

HCMC University of Technology, Vietnam

Program Committee Chairs

Tran Khanh Dang

HCMC University of Technology, Vietnam

Erich Neuhold

University of Vienna, Austria

Makoto Takizawa

Hosei University, Japan

Publication Chair

Josef Küng

Johannes Kepler University Linz, Austria

Publicity Chairs

Lam Son Le

HCMC University of Technology, Vietnam

Quan Thanh Tho

HCMC University of Technology, Vietnam

Hoang Tam Vo

National University of Singapore, Singapore

Local Organizing Chairs

Tran Tri Dang

HCMC University of Technology, Vietnam

Tran Ngoc Thinh

HCMC University of Technology, Vietnam

Le Thanh Sach

HCMC University of Technology, Vietnam

Program Committee

Pedro Antunes	Victoria University of Wellington, New Zealand
Stephane Bressan	National University of Singapore, Singapore
Alain Bui	University of Versailles Saint-Quentin-en-Yvelines, France
Hyunseung Choo	Sungkyunkwan University, South Korea
Somsak Choomchua	King Mongkut's Institute of Technology Ladkrabang, Thailand
Nguyen Tuan Dang	University of Information Technology, VNUHCM, Vietnam
Thanh-Nghi Do	Can Tho University, Vietnam
Dirk Draheim	University of Innsbruck, Austria
Phuong Hoai Ha	University of Tromsø, Norway
Kazuhiko Hamamoto	Tokai University, Japan
Yo-Sung Ho	Gwangju Institute of Science and Technology, Korea
Tran Van Hoai	HCMC University of Technology, Vietnam
Nguyen Viet Hung	University of Trento, Italy
Nguyen Quoc Viet Hung	École polytechnique fédérale de Lausanne, Switzerland
Tran Nguyen Hoang Huy	University of Vienna, Austria
Trung-Hieu Huynh	Industrial University of Ho Chi Minh City, Vietnam
Ryutaro Ichise	National Institute of Informatics, Japan
Tomohiko Igasaki	Kumamoto University, Japan
Muhammad Ilyas	University of Sargodha, Pakistan
Koichiro Ishibashi	University of Electro-Communications, Japan
Hiroshi Ishii	Tokai University, Japan
Eiji Kamioka	Shibaura Institute of Technology, Japan
M-Tahar Kechadi	University College Dublin, Ireland
Nhien An Le Khac	University College Dublin, Ireland
Ismail Khalil	Johannes Kepler University Linz, Austria
Khamphong Khongsomboon	National University of Laos, Laos
Hiroaki Kikuchi	Meiji University, Japan
Surin Kittitornkun	King Mongkut's Institute of Technology Ladkrabang, Thailand
Andrea Ko	Corvinus University of Budapest, Hungary
Hilda Kosorus	Johannes Kepler University Linz, Austria
Pierre Kuonen	Applied Science University, Fribourg, Switzerland
Lam Son Le	HCMC University of Technology, Vietnam
Tan Kian Lee	National University of Singapore, Singapore
Fabio Massacci	University of Trento, Italy

Ngo Nguyen Nhat Minh	University of Trento, Italy
Atsuko Miyaji	Japan Advanced Institute of Science and Technology, Japan
Takumi Miyoshi	Shibaura Institute of Technology, Japan
Hiroaki Morino	Shibaura Institute of Technology, Japan
Thanh Binh Nguyen	HCMC University of Technology, Vietnam
Benjamin Nguyen	Inria Rocquencourt, Versailles, France
Khoa Nguyen	National ICT Australia, Australia
Phan Trong Nhan	Johannes Kepler University Linz, Austria
Masato Oguchi	Ochanomizu University, Japan
Eric Pardede	La Trobe University, Australia
Cong Duc Pham	University of Pau, France
Nguyen Khang Pham	Can Tho University, Vietnam
Phung Huu Phu	University of Gothenburg, Sweden, and University of Illinois at Chicago, USA
Sathaporn Promwong	King Mongkut's Institute of Technology Ladkrabang, Thailand
Tran Minh Quang	National Institute of Informatics, Japan
Pesesteer Racskó	Corvinus University of Budapest, Hungary
Akbar Saiful	Institute of Technology Bandung, Indonesia
Tran Le Minh Sang	University of Trento, Italy
Christin Seifert	University of Passau, Germany
Erik Sonnleitner	Johannes Kepler University Linz, Austria
Reinhard Stumtner	Software Competence Center Hagenberg, Austria
Zoltán Szabó	Corvinus University of Budapest, Hungary
David Taniar	Monash University, Australia
Chivalai Temiyasathit	King Mongkut's Institute of Technology Ladkrabang, Thailand
Quoc Cuong To	Inria Rocquencourt, Versailles, France
Shigenori Tomiyama	Tokai University, Japan
Ha-Manh Tran	International University, Vietnam
Tuan Anh Truong	University of Trento, Italy
Hong Linh Truong	Vienna University of Technology, Austria
Truong Minh Nhat Quang	Can Tho University of Technology, Vietnam
Osamu Uchida	Tokai University, Japan
Hoang Tam Vo	National University of Singapore, Singapore
Pham Tran Vu	HCMC University of Technology, Vietnam
Qing Wang	Australian National University, Australia
Edgar Weippl	Technical University of Vienna, Austria
Shigeki Yamada	National Institute of Informatics, Japan

External Reviewers

Nguyen Ngoc Thien An	University College Dublin, Ireland
Truong Quynh Chi	HCMC University of Technology, Vietnam
Nguyen Van Doan	Japan Advanced Institute of Science and Technology, Japan
Truong Quang Hai	HCMC University of Technology, Vietnam
Ngo Chan Nam	Data Security Applied Research Lab, HCMUT, Vietnam
Tran Thi Que Nguyet	HCMC University of Technology, Vietnam
Huynh Van Quoc Phuong	Data Security Applied Research Lab, HCMUT, Vietnam
Nguyen Dinh Thanh	Data Security Applied Research Lab, HCMUT, Vietnam
Nguyen Thi Ai Thao	HCMC University of Technology, Vietnam
Le Thi Bao Thu	HCMC University of Technology, Vietnam
Le Thi Kim Tuyen	Sungkyunkwan University, South Korea

Table of Contents

Big Data Analytics and Applications

On Context- and Sequence-Aware Document Enrichment and Retrieval towards Personalized Recommendations	1
<i>Hilda Kosorus, Peter Regner, and Josef Küng</i>	
Forests of Oblique Decision Stumps for Classifying Very Large Number of Tweets	16
<i>Van T. Le, Thu M. Tran-Nguyen, Khang N. Pham, and Nghi T. Do</i>	
Performance Evaluation of a Natural Language Processing Approach Applied in White Collar Crime Investigation	29
<i>Maarten van Banerveld, Nhien-An Le-Khac, and M-Tahar Kechadi</i>	
An Efficient Similarity Search in Large Data Collections with MapReduce	44
<i>Trong Nhan Phan, Josef Küng, and Tran Khanh Dang</i>	

Security and Privacy Engineering

Memory-Based Multi-pattern Signature Scanning for ClamAV Antivirus	58
<i>Nguyen Kim Dien, Tran Trung Hieu, and Tran Ngoc Thinh</i>	
Constructing Private Indexes on Encrypted Data for Outsourced Databases	71
<i>Yi Tang, Ji Zhang, and Xiaolei Zhang</i>	
An Extensible Framework for Web Application Vulnerabilities Visualization and Analysis	86
<i>Tran Tri Dang and Tran Khanh Dang</i>	
A Combination of Negative Selection Algorithm and Artificial Immune Network for Virus Detection	97
<i>Vu Thanh Nguyen, Toan Tan Nguyen, Khang Trong Mai, and Tuan Dinh Le</i>	
De-anonymising Set-Generalised Transactions Based on Semantic Relationships	107
<i>Hoang Ong and Jianhua Shao</i>	
An Implementation of a Unified Security, Trust and Privacy (STP) Framework for Future Integrated RFID System	122
<i>Mohd Faizal Mubarak, Jamalul-Lail Ab Manan, and Saadiyah Yahya</i>	

Crowdsourcing and Social Network Data Analytics

Toward a Nexus Model Supporting the Establishment of Business Process Crowdsourcing	136
<i>Nguyen Hoang Thuan, Pedro Antunes, and David Johnstone</i>	
Link Prediction in Social Networks Based on Local Weighted Paths	151
<i>Danh Bui Thi, Ryutaro Ichise, and Bac Le</i>	
An Architecture Utilizing the Crowd for Building an Anti-virus Knowledge Base	164
<i>Nguyen Hoang Thuan, Pedro Antunes, David Johnstone, and Minh Nhat Quang Truong</i>	

Biometrics and Data Protection in Smart Devices

Two-Way Biometrics-Based Authentication Scheme on Mobile Devices	177
<i>Duong-Tien Phan, Toan-Thinh Truong, Minh-Triet Tran, and Anh-Duc Duong</i>	
Prospective Cryptography in NFC with the Lightweight Block Encryption Algorithm LEA	191
<i>Ha Van Nguyen, Hwajeong Seo, and Howon Kim</i>	
Enhance Fuzzy Vault Security Using Nonrandom Chaff Point Generator	204
<i>Minh Tan Nguyen, Quang Hai Truong, and Tran Khanh Dang</i>	
Smart Card Based User Authentication Scheme with Anonymity	220
<i>Toan-Thinh Truong, Minh-Triet Tran, and Anh-Duc Duong</i>	

Cloud Data Management and Applications

Cloud-Based ERP Solution for Modern Education in Vietnam	234
<i>Thanh D. Nguyen, Thanh T.T. Nguyen, and Sanjay Misra</i>	
Heuristics for Energy-Aware VM Allocation in HPC Clouds	248
<i>Nguyen Quang-Hung, Duy-Khanh Le, Nam Thoai, and Nguyen Thanh Son</i>	

Advances in Query Processing and Optimization

Information-Flow Analysis of Hibernate Query Language	262
<i>Agostino Cortesi and Raju Halder</i>	

Investigation of Regularization Theory for Four-Class Classification in Brain-Computer Interface	275
<i>Le Quoc Thang and Chivalai Temiyasathit</i>	
Enhancing Genetic Algorithm with Cumulative Probabilities to Derive Critical Test Scenarios from Use-Cases	286
<i>An T. Tran, Tho T. Quan, and Thuan D. Le</i>	
Towards a Semantic Linked Data Retrieval Model	300
<i>Van Bich Nguyen and Dang Tuan Nguyen</i>	
Author Index	311

On Context- and Sequence-Aware Document Enrichment and Retrieval towards Personalized Recommendations

Hilda Kosorus, Peter Regner, and Josef Küng

Institute of Application Oriented Knowledge Processing,
Johannes Kepler University, Linz, Austria
{hkosorus, jkueng}@faw.jku.at, pregner@faw.at
www.faw.jku.at

Abstract. The amount of unstructured data has grown exponentially during the past two decades and continues to grow at even faster rates. As a consequence, the efficient management of this kind of data came to play an important role in almost all organizations.

Up to now, approaches from many different research fields, like information search and retrieval, text mining or query expansion and reformulation, have enabled us to extract and learn patterns in order to improve the management, retrieval and recommendation of documents. However, there are still many open questions, limitations and vulnerabilities that need to be addressed.

This paper aims at identifying the current major challenges and research gaps in the field of “document enrichment, retrieval and recommendation”, introduces innovative ideas towards overcoming these limitations and weaknesses, and shows the benefits of adopting these ideas into real enterprise content management systems.

Keywords: document enrichment, document retrieval, document recommendation, ontology learning, document annotation, information retrieval, enterprise content management.

1 Introduction

Advances in areas such as information retrieval, machine learning, data mining and knowledge representation have played an important role in dealing with and making use of an ever growing amount of textual information. Various existing approaches have enabled us up to now to extract and learn patterns in order to improve the management, retrieval and interpretability of information. However, there are still many open questions, limitations and vulnerabilities that need to be addressed. Many basic functionalities of content management systems are still not able to provide satisfactory results.

Based our experience with mid- and large-sized organizations we came to the conclusion that most of the techniques, from the previously mentioned research

areas, are rarely used in enterprise content management (ECM) scenarios. Innovative approaches for classification, retrieval and recommendation have barely passed the evaluation or pilot status.

This paper aims at addressing current challenges in complex enterprise content scenarios and the research gaps within the field of “document enrichment, retrieval and recommendation”, given a very large amount of unstructured and semi-structured domain-specific textual documents (e.g., application forms and documents, health care and medical documents, news articles, etc.). Additionally, we present new ideas towards overcoming these challenges and limitations.

We introduce novel techniques and approaches in the application scenario of efficient ECM within private and public organizations. We base our claims on the review and analysis of the current state-of-the-art in the related research fields, and on the discovered weaknesses and limitations of existing employed methods.

The contribution of this paper is threefold: first, it gives an overview of the most recent related work in the related fields of ontology learning, document enrichment, retrieval and recommendation; secondly, it introduces innovative ideas towards document enrichment, retrieval and recommendation; and, thirdly, it presents the expected benefits and profit of the implementation and employment of these new techniques.

The main innovative idea of this paper is exploring, extracting and using various types of contextual information and sequences to enable a more accurate, efficient and effective document retrieval and recommendation. We organize this idea into three main research goals. Our first goal is to use multiple knowledge bases (domain-specific ontologies, structured Web data, user groups and networks, external structure of documents, document meta-data, etc.) to infer, extract and enrich/annotate documents with two types of contextual information: content-dependent and -independent. Secondly, we plan to use these document annotations to perform relevant, meaningful document retrieval to user queries. And thirdly, as our main objective, we aim at developing a personalized, sequence-based recommendation of documents by using the history of document access sequences and the previously generated annotations and knowledge bases.

The rest of the paper is organized in the following way. In the next section, we give an overview of the most recent and relevant related work in the field of ontology learning, document enrichment, retrieval and recommendation. Then, in Section 3 we present in more detail the previously mentioned research goals. Finally, in Section 4, we show the benefits and consequences of implementing and adopting these ideas into real ECM systems.

2 Related Work and Limitations

During the conceptualization of our idea and the analysis of the state-of-the-art, we identified four main areas that are directly related to our research question (i.e. how to improve ECM systems?) and to the proposed approach: ontology learning, document enrichment and annotation, document retrieval and document recommendation.

In some cases, we intend to use the reviewed techniques and methods, in other cases, we plan to improve and extend them to allow for better results, more efficiency and effectiveness when working with ECM systems. In the following, we will review some of the most recent and related findings in the aforementioned research fields.

2.1 Ontology Learning

Ontologies are formal and explicit specifications in the form of concepts and relations of shared conceptualizations [9]. They are a means for knowledge representation and are used as structural frameworks to organize information. Ontologies constitute a basic building block of the Semantic Web, allowing agents to exchange, share, reuse and reason about concepts and relations using axioms.

An ontology can be thought of as a directed graph consisting of concepts as nodes and relations as edges between nodes. The definition of ontologies requires a formal (i.e. natural language independent) representation using formal languages (e.g. Web Ontology Language or OWL) that allows reasoning and defining constraints.

The extent of relational and axiomatic richness gives rise to a wide spectrum of ontology types: from glossaries, thesauri and taxonomies with simple hierarchical structures (also referred to as informal or lightweight ontologies), to formal taxonomies, frames and description logics (i.e. formal, heavyweight ontologies).

Ontology learning from text is the process of identifying terms, concepts, relations and axioms from textual information and using them to construct and maintain an ontology [18]. There are five types of output in ontology learning: terms, concepts, taxonomic relations, non-taxonomic relations and axioms. In order to obtain each of these outputs, certain tasks need to be performed: text pre-processing, term extraction, concept formation, (taxonomic and non-taxonomic) relation discovery and axiom learning. The techniques employed for these tasks were adopted from established research fields, like information retrieval, machine learning, data mining, natural language processing and knowledge representation and reasoning, and can be classified into statistics-based, linguistics-based, logic-based and hybrid approaches.

In the past two decades there have been several independent surveys conducted on the topic of ontology learning. The most prominent works are the ones recorded in [18] and [20]. Zhou [20] identifies five relevant issues in this area: (a) the importance of representation; (b) the involvement of humans, which remains highly necessary; (c) the need for common benchmarks for evaluating ontologies; (d) more research needed for the discovery of non-taxonomic relations; and (e) more effort required in making existing techniques operational on cross-domain text on a Web-scale. Wong et al. [18] review the current state of the art in the area of ontology learning from several perspectives: techniques, evaluation methods, existing ontology learning systems; review recent advances and current trends, and present future research directions.

Evaluation is an important aspect in ontology learning that allows experts to assess the resulting ontologies and, in some cases, guide and refine the learning

process. Existing evaluation approaches can be grouped into the following categories:

- Task-based evaluation – assess the adequacy of ontologies in the context of other applications;
- Corpus-based evaluation – use domain-specific data sources to determine to what extent are ontologies able to cover the respective domain; and
- Criteria-based evaluation – determine how well the created ontologies adhere to a set of criteria.

Some of the most prominent ontology learning systems are: ASIUM, Text-to-Onto, TextStorm/Clouds, SYNDIKATE, OntoLearn, CRCTOL and OntoGain.

Recent challenges in the field of ontology learning techniques include the improvement of term extraction and concept formation and the relation discovery tasks. Additionally, the learning of ontologies from social data and across different languages has also been a topic of interest in the past decade. To further improve the performance of term extraction and relation discovery, recent research work focused on constructing very large text corpora from the Web, as well as on using (semi-)structured Web data (e.g. Wikipedia.) and social data. However, the results obtained so far require further investigation.

2.2 Query Expansion and Document Enrichment

The ever growing amount of available information gave rise to many different challenges concerning storage, management and, especially, retrieval. Text retrieval is the process by which users seek to find documents relevant to the subject of their query.

When searching for information, users usually describe their information need using several keywords, which may happen to be different from the words used in the actually relevant documents. Since language is inherently ambiguous and since queries tend to be rather short, the returned results might lack documents relevant to the user's request. This problem is known as the vocabulary problem [7] or the word mismatch problem. Therefore, the ineffectiveness of information retrieval systems is mainly caused by the inaccuracy with which a query formed by a few keywords models the actual user information need [5].

In order to improve the results of document retrieval and bridge the gap between the user's search intent and the document's full-text, two main approaches were adopted: automatic query expansion (AQE) [5] and automatic document annotation. Among these, we also mention here other methods: interactive query refinement, relevance feedback, word sense disambiguation and search results clustering. One of the most natural and successful technique was the idea of query expansion. The idea behind query expansion is to expand the query term issued by the user with suitable, related terms in order to match the documents' vocabulary. According to [10], query expansion leads to higher recall, however, it decreases the retrieval precision. This is usually caused by the fact that the context of the query is not known or the terms used in the query

are rather conceptual and, therefore, the expansion of the query might still not match the user's search intent. An alternative solution to the vocabulary problem is automatic document annotation or enrichment.

Document enrichment or annotation refers to the process of automatically expanding documents with annotations using external knowledge bases. These annotations can vary from meta-data (e.g. author, time stamp of creation, location, system information, etc.) to content-dependent information (e.g. topic, keywords, concepts, etc.). The idea of document enrichment emerged from the need of retrieving relevant and meaningful documents to user queries.

Current research work in this area uses strictly the document's content and use some kind of external (semi-)structured data, such as Wikipedia [10] or other knowledge bases to enrich documents and enable a more efficient and effective retrieval.

2.3 Document Retrieval

Document retrieval, as part of the larger research area of information retrieval (IR), is the process of obtaining structured or unstructured textual resources relevant to an information need from a collection of document resources. The techniques used in this direction vary from set-theoretic (e.g. Boolean, fuzzy) to algebraic (e.g. vector space models, latent semantic indexing), probabilistic (e.g. probabilistic relevance model, Latent Dirichlet Allocation) or feature-based models [2,17].

Evaluating the performance of information retrieval methods has been a cornerstone of IR and many different measures have been proposed to assess the results of retrieval systems. Most common evaluation measures (precision, recall, f-measure, fall-out) assume some kind of ground truth notion of relevancy: every document is known to be either relevant or non-relevant to a particular query. In practice, however, there might be different shades of relevancy.

There has been extensive research done in the area of information retrieval. We focus here on the narrower field of document or text retrieval and review a selection of popular techniques: vector space models, query expansion, Latent Dirichlet Allocation (LDA).

Vector space models are algebraic models for representing text documents (or other objects) as vectors of term-weights. Terms, in this context, are typically single words, keywords or short phrases. The dimensionality of the vector, and, therefore, of the model itself is given by the number of terms in the vocabulary. One of the best known schemes for weighting terms in a document is the classic term frequency-inverse document frequency (TF-IDF) model. It reflects how important is a term to a document within a collection.

One of the most natural and successful techniques in information retrieval is to expand the original query with other words that best capture the actual user intent and produces a more useful query, i.e., a query that is more likely to retrieve relevant documents [20]. This technique is known as automatic query expansion (AQE). In the past decades, a vast number of approaches have been developed that use various data sources and employ sophisticated methods for

finding new terms correlated with the original query terms [3,8,13,14]. In [3], a new framework for query reformulation was introduced that uses multiple information sources (a large web corpus, articles and titles from Wikipedia) for weighting the explicit query concepts as well as selecting relevant and diverse set of weighted expansion terms.

Although very successful and broadly adopted by many retrieval systems, recent research work suggests that while AQE has a high recall, it lacks retrieval precision [10].

A somewhat complementary technique to AQE is automatic document annotation and enrichment, which was described in more detail in the previous section. The idea behind this approach is to bridge the gap between the user's query and the available documents by enriching the extracted document-terms using external knowledge bases (e.g. taxonomies, ontologies, structured Web data, etc.). Recent research results show that such methods can outperform state-of-the-art query expansion and vector space model-based approaches [10].

Both techniques of query expansion and document enrichment require some kind of matching method between the query terms and the document terms. These can vary from simple set comparisons to more advanced semantic similarity measures. An overview of state-of-the-art similarity measures is presented in [11], where the authors also introduce a new short-text semantic similarity measure.

2.4 Document Recommendation

Recommender systems are defined as systems that produce individualized recommendations as output in order to guide the user in a personalized way to interesting and useful objects in a large space of possible options [4].

Document recommendation is essential for user-oriented document management systems and search engines. Accurate, useful and relevant recommendation could reduce the users' search time, improve the user's interaction with the underlying system and enable a more efficient work [6]. The most widely adopted approaches to document recommendation rely on document content, explicit or implicit user feedback, and user profiles. Approaches based on collaboration filtering (CF) produce recommendations by computing the similarity or the correlation existing between the items from user activity logs.

Recent research in this area tries to leverage the sequential data within user query logs to identify and learn search intent in order to satisfy the user's information need and improve the quality of recommendations. The main techniques behind such approaches rely on probabilistic models or search query graphs [12,16].

In [6], a context-aware document recommendation method is introduced that uses variable length Markov models to model the sequential user access. The main idea behind this method is that a user's current working context could be inferred from his/her previous activity sequence and recommendation can be more accurate by taking such activity context into account.

Modeling query graphs from user activity logs has been widely adopted for Web search engines to identify search intents and produce reliable, useful recommendations [1,16,19]. A more recent work [16] tries to leverage three types of objects: queries, web pages and Wikipedia concepts collaboratively for learning generic search intents and constructs a heterogeneous graph to represent multiple types of relationships between them.

3 Approach

Despite the very rich and vast amount of related work in these four fundamental research fields, many challenges and open issues have been identified (see Section 2) that require further investigation and new perspectives on the matter. Motivated by this research gap and based on the previous work of the authors [11,12,15], we introduce new concepts and ideas towards an improved document management experience with enterprise content management (ECM) systems.

The aim of this work is to address these challenges in the restricted domain of document enrichment, retrieval and recommendation in order to develop generic, domain- and language-independent solutions that enable an improved and efficient user work experience. Our proposed approach will use various external knowledge bases and structured Web data sources to enrich the collection of documents with two categories of context information: content-dependent (i.e. semantic context) and content-independent (or external information), encompassing a wide spectrum of document context. In order to do this, we need to (semi-)automatically construct domain-specific knowledge bases that match the requirements and specifications of their individual, complex application area (e.g. health care and medical documents, news archives, digital libraries, etc.).

Based on the two types of context information, we plan to develop innovative document retrieval and recommendation approaches that allow users to access fast and reliably only the relevant and useful documents. Moreover, we aim to leverage not only the document annotations, but also the sequential data of user activity log in order to make predictions about the user's future information needs. The history of the users' sequential document access allows us to collaboratively learn patterns and identify common search goals.

In order to (semi-)automatically construct the domain-specific knowledge base, we intend to use state-of-the-art techniques and tools from the field of ontology learning. However, we will focus our research towards improving specific tasks and addressing current challenges, like the discovery of non-taxonomic relations, concept formation and axiom learning; defining and developing evaluation approaches to assess the resulting ontology and refine the learning process. We want to address the knowledge acquisition bottleneck (i.e. acquiring the knowledge necessary for learning ontologies) by using multiple sources of structured Web data (e.g. Wikipedia) and the problem of coping with very large document collections. Our aim is to learn ontologies through an iterative process, moving from initial and lightweight to formal, heavyweight ontologies, while continuously improving and extending with the growing database of documents.

The learned ontologies will serve as a knowledge base for document enrichment. While current approaches focus mainly on content-based annotations, our goal is to enrich documents with both content-dependent and -independent information by using multiple knowledge bases, with the aim to enable an improved and efficient document retrieval experience. Additionally, existing methods towards matching user queries to documents use rather simple, limited similarity measures (Jaccard index [7], Dice’s coefficient) that are not able to identify complex or conceptual search requests and retrieve relevant results.

In order to address the vocabulary problem and improve the retrieval quality, we intend to develop a hybrid approach towards document retrieval that combines both query expansion and document enrichment (content-dependent and -independent annotations). Moreover, to enable a better matching between queries and documents, innovative semantic similarity measures will be adopted based on the learned ontologies. Our goal is to allow users to perform complex, conceptual queries, while still maintaining a high retrieval quality.

While document retrieval answers to temporary information needs, in application domains where users manage, maintain and request documents to perform clear tasks over a longer period of time (e.g. maintaining a patient’s file case for diagnosis, writing reports over previous related articles), identifying user goals and predicting future search intents can play a crucial role in improving the user’s work experience.

To address this important aspect, we direct our main focus towards the development of innovative context-aware and personalized recommendation methods that leverage the available history of user document access and the existing document annotations, both content-dependent and -independent context information.

Based on the above mentioned limitations and challenges, we formulate the following research objectives:

1. advances in ontology learning,
2. content-dependent and -independent document enrichment,
3. context-aware document retrieval, and
4. context- and sequence-aware personalized document recommendation,

in the presence of:

- very large unstructured and semi-structured document collections,
- complex application domains,
- conceptual user queries, and
- sequential user activity data.

In the following, we will elaborate in detail on the above mentioned research objectives and present our first ideas and approaches with respect to each of them.

3.1 Advances in Ontology Learning

Research work in the field of ontology learning has matured a lot over the past two decades and many techniques have been developed for the various tasks

within the learning process. However, there are still certain challenges and limitations that need to be addressed in order to allow automatic, heavyweight, application-independent and cross-language ontologies to be developed. One of our goals is to advance the current research in this field and address particular limitations that will improve the performance of the next technical goals.

A detailed study of the state-of-the-art in ontology learning has revealed several weaknesses and open issues (see Section 2.1). We intend to tackle and improve the following aspects:

- The knowledge acquisition bottleneck by using (semi-)structured collaboratively maintained Web data resources (e.g. Wikipedia, DBPedia, Austrian Social Security Law, organization-specific process instructions, etc.), an already existing initial large set of domain-specific documents and other taxonomies (WordNet, Wiktionary, OpenThesaurus, etc.);
- Data cleanliness – manage noise, richness, diversity and validity of the resources used for ontology learning;
- Cross-language ontology learning – allow ontologies to be language-independent;
- Cope with very large domain-specific document collections – these will not only serve as a basis for the initial construction of the ontology, but also during the entire process, at each iteration;
- Move from lightweight to formal ontologies through an iterative process – extending the initial lightweight ontologies to a full-fledged ones;
- Improve ontology learning tasks:
 - Term extraction,
 - Concept formation,
 - Taxonomic and non-taxonomic relation discovery, and
 - Axiom learning.

Figure 1 shows the concept behind the role of the ontology learning and the document enrichment tasks, and the interaction and dependency between their constituent parts.

3.2 Content-Dependent and -Independent Document Enrichment

Our second objective is to address the vocabulary (i.e. term mismatch) problem and bridge the gap between the user’s information need (given by the search query terms) and the documents’ vocabulary and enable complex, conceptual queries by using two categories of document context information (see Figure 1):

1. *Content-dependent* – refers to any additional information related to the semantic content of the document (semantically related concepts, terms, synonyms, etc.), and
2. *Content-independent* – refers to aspects that are independent of the actual content of the document, but still relevant in certain retrieval situations (e.g. meta-data, document type, system context, author, time, geographic location, etc.).

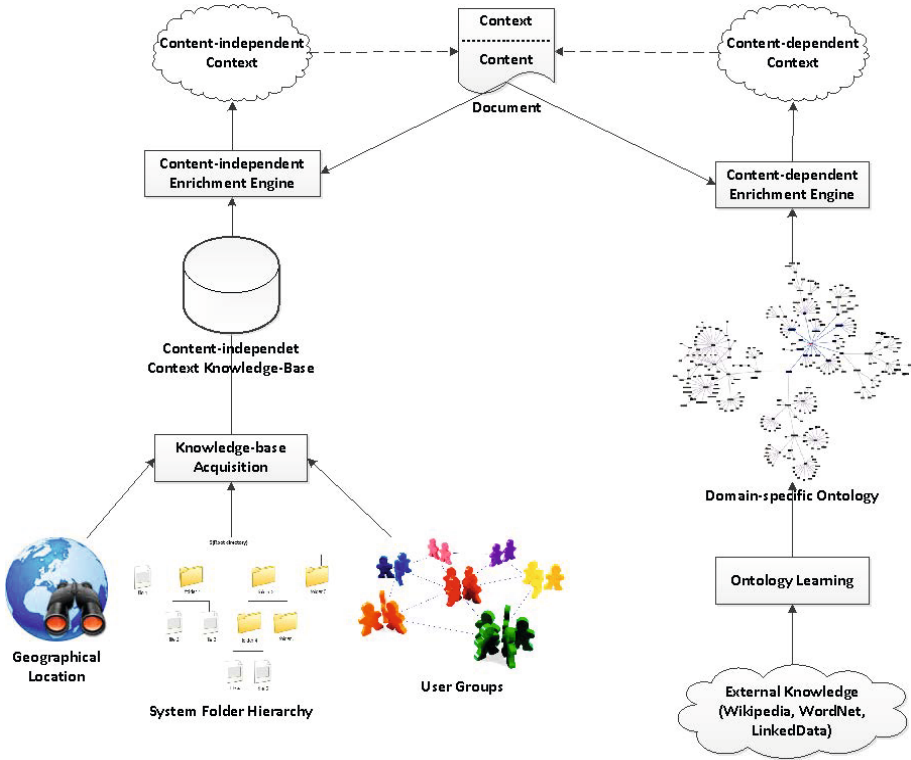


Fig. 1. Concept – Ontology learning and document enrichment

Based on these two types of contextual data, we formulate the following objectives:

- Content-dependent document enrichment – based on the constructed ontology, identify relevant and meaningful terms and concepts to annotate documents with; and
- Content-independent document enrichment: system context, user groups and networks, geographical context, structural context (possibly also based on various knowledge-bases).

The idea behind this approach is that by using a rich document context (both content-dependent and -independent) we will be able to cope with conceptual queries and significantly improve the retrieval and recommendation results. Our aim is to help the users be more efficient in their tasks and answer their information need by guiding them towards useful and relevant documents.

3.3 Context-Aware Document Retrieval

The second step towards solving the well-known vocabulary problem and towards significantly improving the document retrieval quality is the development of an

innovative document retrieval approach. In the future we intend to develop a hybrid approach towards document retrieval that combines both query expansion and document annotations (content-dependent and -independent annotations), which has not been done so far. Moreover, to enable a better matching between queries and documents, innovative semantic similarity measures will be adopted based on the learned ontologies. Our goal is to allow users to perform complex, conceptual queries, while still maintaining a high retrieval quality.

With this new method we want to address the persistent challenge in the field of document retrieval: returning only relevant, useful and informative results to the user's search request in the presence of:

- complex domains,
- very large unstructured and semi-structured document collections, and
- conceptual queries,

by modeling these domains using representative ontologies and enriching documents with content- dependent and -independent information.

Figure 2 shows the concept behind our context-aware document retrieval approach and the interaction behind the relevant system components for the document retrieval task.

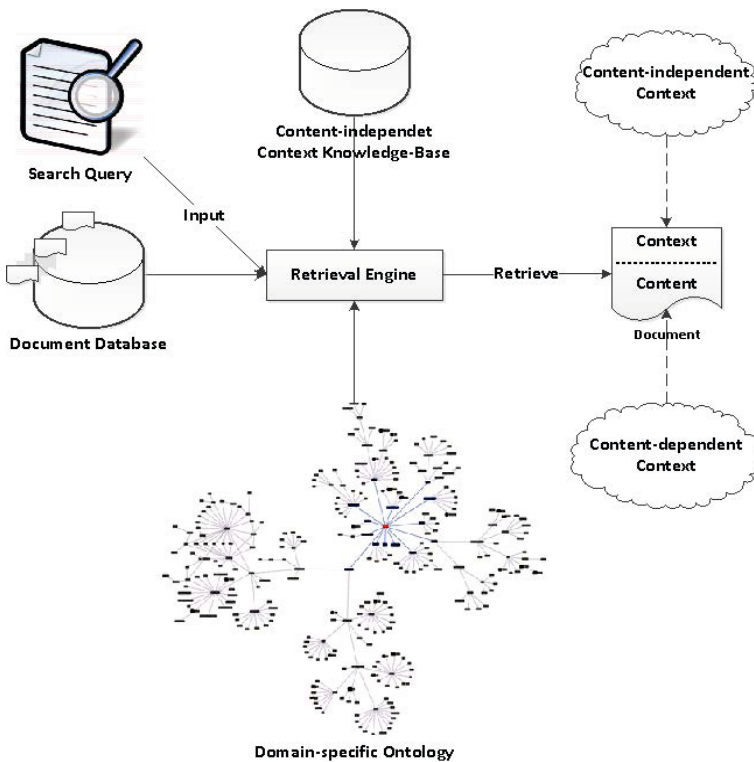


Fig. 2. Concept – Context-aware document retrieval

3.4 Context- and Sequence-Aware Personalized Document Recommendation

The first three objectives serve as a foundation for the achievement of the fourth and main goal.

The above described new document retrieval approach addresses only temporary, momentary user information needs and it can be effective and efficient only in particular situations (e.g. when the accomplishment of the user's current task requires a single successful query).

However, when users manage, maintain and request documents to perform a related chain of tasks over a longer period of time (e.g. maintaining a patient's file case for diagnosis, writing reports over previous related articles, etc.), identifying user goals and predicting future search intents can play a crucial role in improving the user's work experience.

Our fourth and the main technical objective is, therefore, to develop a personalized document recommendation approach that uses the rich document context annotations (see objective B) and the user activity log (i.e. sequence of document access). The aim of this approach is to

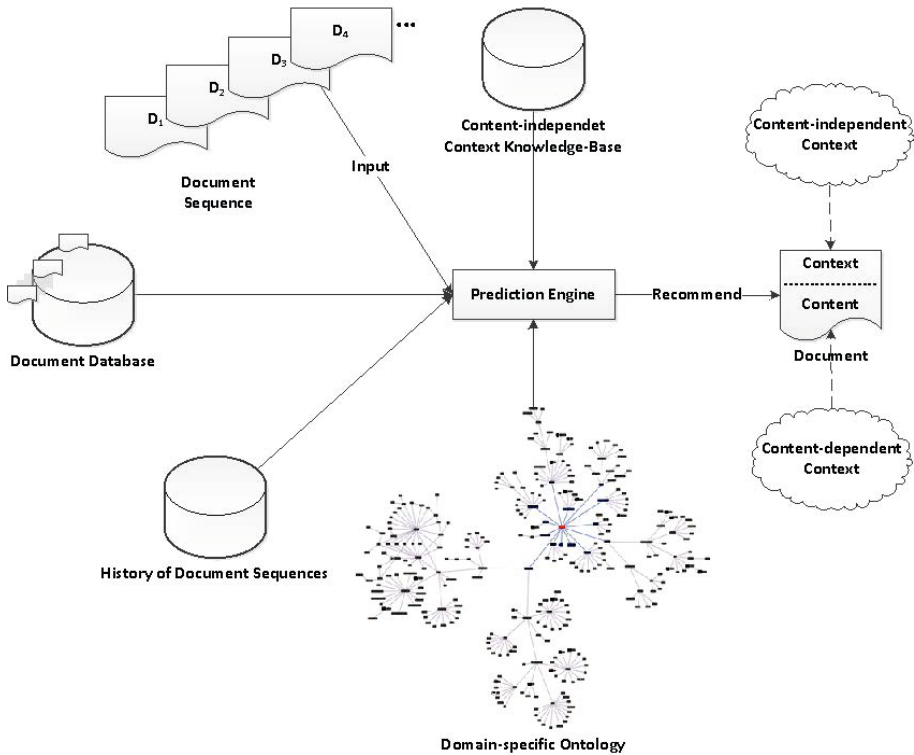


Fig. 3. Concept – Context- and sequence-aware personalized document recommendation

- identify user goals,
- extract navigation patterns and mind-maps, and
- predict the user’s search intent.

Figure 3 shows the concept behind the context- and sequence-aware personalized document recommendation approach and highlights the relevant components of the overall system that enable the execution of this task.

4 Benefits and Consequences

From our experience with enterprise content management (ECM) implementations at different organizations we have learned that, in most cases, the techniques of document classification and enrichment of documents with additional valuable information are barely exploited, in contrast to their potential benefits. Although automatic document annotation is essential for effective retrieval and recommendation, in some cases, categorization of documents and meta-data extraction is performed manually by end users. In most cases, however, the process of manual document enrichment is not feasible, given that real-life ECM projects typically start with a fairly large set of unstructured documents.

Our findings will leverage the potentials of ECM systems and will positively influence the way users process documents. Our approach to enrich documents, not only with content-dependent, but also with content independent information, perfectly fits the needs of modern ECM systems. With features like user logs, document-based collaboration, document-based workflows or project spaces, they are a valuable source for additional context information. An open ECM software architecture should support the integration of additional context information from external sources.

Current ECM systems are used enterprise-wide and more and more across enterprise borders. Internal and external users from different locations and organizations, with different background knowledge, process documents in a collaborative way. As a consequence, single-domain- and static-ontologies fail to deliver the expected results. These scenarios require multiple and complex domains and ontologies that grow on-demand in terms of size and quality. These aspects constitute the focus of our work and represent the foundation for our retrieval and recommendation techniques.

Effective document search relies on a rich set of additional content-dependent and -independent information to be extracted and assigned. This increases the chances of retrieving the appropriate documents when searching, regardless of the user’s background knowledge. Beyond this, we expect a substantial improvement of the system’s recommendation ability when using the document annotations and the history of user activities, especially in the case of users who process documents in a collaborative way. Our work aims at improving the user experience with ECM systems by enhancing documents with valuable information, improving the retrieval and recommendation capability in order to deliver only the information the user needs at a given point in time.

5 Conclusion and Future Work

In this paper we identified four research fields relevant for the advancement and improvement of enterprise content management (ECM) systems and we gave an overview of their current state-of-the-art and challenges. We argued that nowadays existing methods adopted for ECM do not satisfy the user needs when it comes to efficiently and effectively retrieve and manage documents.

We introduced novel ideas towards the advancement of document enrichment, retrieval and recommendation, which we organized in four main research objectives. Our goal is to use multiple sources of context information (i.e. content-dependent and -independent) to enrich documents, and enable relevant and meaningful document retrieval. Further on, we aim at developing a personalized, sequence-based recommendation of documents by using the history of document access sequences and the previously generated annotations and knowledge bases.

The benefits of the expected results are manifold. All in all, we predict an increased efficiency and effectiveness of the users when adopting these new generation of enterprise content management systems. In the future we want to dedicate our attention towards the development and implementation of the presented ideas, then analyze and evaluate the results through various experiments. To start with, we plan to conduct a survey with different organizations that provide ECM systems to their users in order to empirically verify the assumptions made in Section 4 with respect to the challenges and limitations of current ECM systems.

References

1. Baeza-Yates, R.: Graphs from search engine queries. In: van Leeuwen, J., Italiano, G.F., van der Hoek, W., Meinel, C., Sack, H., Plášil, F. (eds.) SOFSEM 2007. LNCS, vol. 4362, pp. 1–8. Springer, Heidelberg (2007)
2. Baeza-Yates, R.A., Ribeiro-Neto, B.: Modern Information Retrieval. Addison-Wesley Longman Publishing Co., Inc., Boston (1999)
3. Bendersky, M., Metzler, D., Croft, W.B.: Effective query formulation with multiple information sources. In: Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, WSDM 2012, pp. 443–452. ACM, New York (2012)
4. Burke, R.: Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction* 12(4), 331–370 (2002)
5. Carpineto, C., Romano, G.: A survey of automatic query expansion in information retrieval. *ACM Comput. Surv.* 44(1), 1:1–1:50 (2012)
6. Chen, J., Chen, T., Guo, H., Yu, T., Wang, W.: Context-aware document recommendation by mining sequential access data. In: Proceedings of the 1st International Workshop on Context Discovery and Data Mining, ContextDD 2012, pp. 6:1–6:7. ACM, New York (2012)
7. Furnas, G.W., Landauer, T.K., Gomez, L.M., Dumais, S.T.: The vocabulary problem in human-system communication. *Commun. ACM* 30(11), 964–971 (1987)
8. Giunchiglia, F., Kharkevich, U., Zaihrayeu, I.: Concept search: Semantics enabled syntactic search. In: Proceedings of the Workshop on Semantic Search (SemSearch 2008) at the 5th European Semantic Web Conference (ESWC 2008), pp. 109–123 (2008)

9. Gruber, T.R.: A translation approach to portable ontology specifications. *Knowl. Acquis.* 5(2), 199–220 (1993)
10. Köhncke, B., Siehndel, P., Balke, W.-T.: Bridging the gap – using external knowledge bases for context-aware document retrieval. In: Urs, S.R., Na, J.-C., Buchanan, G. (eds.) *ICADL 2013*. LNCS, vol. 8279, pp. 11–20. Springer, Heidelberg (2013)
11. Kosorus, H., Bögl, A., Küng, J.: Semantic similarity between queries in QA system using a domain-specific taxonomy. In: Maciaszek, L.A., Cuzzocrea, A., Cordeiro, J. (eds.) *ICEIS (1)*, pp. 241–246. SciTePress (2012)
12. Kosorus, H., Küng, J.: Learning-oriented question recommendation using variable length Markov models and Bloom’s taxonomy. *T. Large-Scale Data- and Knowledge-Centered Systems* (2014)
13. Lüke, T., Schaer, P., Mayr, P.: Improving retrieval results with discipline-specific query expansion. In: Zaphiris, P., Buchanan, G., Rasmussen, E., Loizides, F. (eds.) *TPDL 2012*. LNCS, vol. 7489, pp. 408–413. Springer, Heidelberg (2012)
14. Malaisé, V., Hollink, L., Gazendam, L.: The interaction between automatic annotation and query expansion: a retrieval experiment on a large cultural heritage archive. In: *Proceedings of the Workshop on Semantic Search (SemSearch 2008) at the 5th European Semantic Web Conference (ESWC 2008)*, pp. 44–58 (2008)
15. Nadschläger, S., Kosorus, H., Bögl, A., Küng, J.: Content-based recommendations within a QA system using the hierarchical structure of a domain-specific taxonomy. In: Hameurlain, A., Tjoa, A.M., Wagner, R. (eds.) *23rd International Workshop on Database and Expert Systems Applications, DEXA 2012, Vienna, Austria, September 3-7*, pp. 88–92. IEEE Computer Society (2012)
16. Ren, X., Wang, Y., Yu, X., Yan, J., Chen, Z., Han, J.: Heterogeneous graph-based intent learning with queries, web pages and wikipedia concepts. In: *Proceedings of the 7th ACM International Conference on Web Search and Data Mining, WSDM 2014*, pp. 23–32. ACM, New York (2014)
17. Singhal, A.: Modern information retrieval: a brief overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering* 24, 2001 (2001)
18. Wong, W., Liu, W., Bennamoun, M.: Ontology learning from text: A look back and into the future. *ACM Comput. Surv.*, 44(4), 20:1–20:36 (2012)
19. Zhou, D., Zhu, S., Yu, K., Song, X., Tseng, B.L., Zha, H., Giles, C.L.: Learning multiple graphs for document recommendations. In: *Proceedings of the 17th International Conference on World Wide Web, WWW 2008*, pp. 141–150. ACM, New York (2008)
20. Zhou, L.: Ontology learning: State of the art and open issues. *Inf. Technol. and Management* 8(3), 241–252 (2007)

Forests of Oblique Decision Stumps for Classifying Very Large Number of Tweets

Van T. Le¹, Thu M. Tran-Nguyen², Khang N. Pham², and Nghi T. Do²

¹ Faculty of Computer Science and Engineering, University of Technology, VNUHCM
No 268, Ly Thuong Kiet Street, District 10, Ho Chi Minh City, Vietnam
ltvan@cse.hcmut.edu.vn

² College of Information Technology, Can Tho University
No 1, Ly Tu Trong Street, Ninh Kieu District, Can Tho City, Vietnam
{tnmthu, pnkhang, dtngghi}@cit.ctu.edu.vn

Abstract. Our investigation aims at constructing oblique decision stump forests to classify very large number of twitter messages (tweets). Twitter sentiment analysis is not a trivial task because tweets are short and getting generated at very fast rate. Supervised learning algorithms can thus be useful to automatically detect positive or negative sentiments. The pre-processing step performs the cleaning tasks and the representation of tweets using the bag-of-words model (BoW). And then we propose oblique decision stump forests based on the linear support vector machines (SVM) that is suitable for classifying large amounts of high dimensional datapoints. The experimental results on twittersentiment.appspot.com corpora (with 1,600,000 tweets) show that our oblique decision stump forests are efficient compared to baseline algorithms.

Keywords: Sentiment analysis in Twitter, the bag-of-words model, oblique decision stump forests, supervised learning.

1 Introduction

The World Wide Web has become the core medium for sharing ideas to others as friends or public. It is a one big huge forum for people who have the same interests in the newly introduced topic about personalities, politicians, products or services to discuss and give their opinions. An opinion may be enclosed in a piece of news published by a new agency, a blog post, a feed in a social web site or in somebody's personal web page. Especially, Twitter - a micro-blogging system is one of the most visited social network sites with 645 millions active registered users¹. This site allows users to post and read text-based messages of up to 140 characters, known as "tweets". The information provided from Twitter is too massive as we could get average number of 58 millions tweets per day. Hence it is much interesting if we could track and understand sentiment and opinion of the grand public by analysing their tweets. Sentiment analysis of tweets is one of the most current researches as there has been many applications [1], [2]. It is possible to measure attitudes expressed within a short message. In the customer relationship

¹ <http://www.statisticbrain.com/twitter-statistics>

management, this is very useful if a product manufacturer or seller could get feedbacks from customers about products qualities, compare consumer opinions of its products and those of its competitors to enhance better its products. Before making a purchase decision, a customer should read all reviews, opinions from his friends or other product buyers. As summarized by the review papers of [3], [4], [5], many researches have been carried out for analysing opinions, extracting sentiments from texts. However, opinions are commonly expressed in unstructured natural language data, understanding (automatically by computer program) the semantic meanings of these opinions (positive, negative, supportive, etc.) is a big challenging issue.

Our paper aims at classifying the sentiment in Twitter. The pre-processing step includes the cleaning tasks and the representation of Twitter messages (tweets) by using the bag-of-words (BoW) model without any feature selection strategy. It brings out massive, very-high-dimensional datasets (called noisy data [6] having many input features with each one containing only a small amount of information, i.e. the BoW representation of tweets). And then we investigate two new learning algorithms, called Bagging and Arcx4 of oblique decision stump that is suitable for classifying large number of datapoints in very-high-dimensional input space. Our forest algorithms aim at constructing multiple oblique decision stumps classifiers in the way of Bagging [7] and Arcx4 [8] to form an ensemble of classifiers more accurate than a single one. The main idea is to increase noise robustness of decision stumps [9] (used as "weak learners") with the multivariate node splitting based on the linear support vector machines (SVM [10]). The numerical test results on *twittersentiment.appspot.com* corpora [11] (with 1,600,000 tweets) show that our forests are more accurate than baseline algorithms, including Multinomial Naïve Bayes (MNB [12]), Maximum Entropy (MaxEnt [13]) and SVM [10].

The paper is organized as follows. Section 2 briefly presents the sentiment classification in Twitter and related work. Section 3 introduces our forests of oblique decision stumps for classifying large amounts of tweets in high dimensional input space. The experimental results are presented in section 4. We then conclude in section 5.

2 Classifying the Sentiment in Twitter

2.1 Classification of Tweets

The sentiment classification in Twitter can be considered as a text categorization problem where the task is to classify tweets into the emotions (positive, neutral or negative feelings). As illustrated in [14], [3], and [4], the sentiment classification in Twitter is a difficult task due to very compressed form and irregular structure of tweets. The Twitter users post short messages with the average length of 14 words or 78 characters. Table 1 presents the example of emotion-tweets.

There are two major approaches to classify tweets [3], [5]. The first one is based on the semantic orientation of tweets. [15] proposed the semantic concepts that tend to have a more consistent correlation with positive or negative sentiment. [16], [17], [18] studied the lexical-based techniques for identifying sentiment. The second one [14], [1], [19], [20] addresses the problem as a text classification. Two recent papers [21] and [22] presented an overview of machine learning in short text classification.

Table 1. Example of $\langle Emotion; Tweet \rangle$ binome

<i>Emotion</i>	<i>Message</i>
<i>negative</i>	I hungryyyy, so hungryyyyyy ...I dined with an ice cream
<i>negative</i>	Need a hug
<i>negative</i>	Photoshop, I hate it when you crash
...	...
<i>neutral</i>	hi there
<i>neutral</i>	jQuery UI 1.6 Book Review - http://cfbloggers.org/?c=30631
<i>neutral</i>	@johncmayer is Bobby Flay joining you?
...	...
<i>positive</i>	i looooooooooove uuuu
<i>positive</i>	Helloooooooooo Tweeeeeeeeters!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!! Wazzup??????
<i>positive</i>	ha ha. I meant that I hated the day....not any fellow twits.
<i>positive</i>	WOOOOO! Xbox is back
...	...

The sentiment classification task consists of the data cleaning, the representation of tweets and supervised classification techniques, including Multinomial Naïve Bayes (MNB [12]), Maximum Entropy (MaxEnt [13]) and SVM [10]. Our research in this paper falls into the second direction.

2.2 Data Cleaning

The tweets are collected from many different media, e.g. computers, ipads, cell phones. And then, the tweets include usernames, hyperlinks, misspellings, emoticons, slang, informal words, irregular words and repeated letters (table 1). Therefore, [1] proposed pre-processing steps to perform the tweets cleaning. Twitter usernames, hyperlinks are converted to token *USERNAME*, *URL*, respectively. The repeated letters and the emoticons are removed from tweets. The emoticons and the acronyms are also converted to the ordinary words.

2.3 Data Representation

The bag-of-words (BoW) model [23], [24] is a simplest representation in order to transform the unstructured textual dataset to a structured one (i.e. a data table). The m tweets are then represented by a $m \times (n + 1)$ matrix *BoW* where:

- the first n columns represent a discriminant word (n words)
- the last column is the class label (positive, neutral or negative)
- $BoW[i][j]$ is the frequency of the j^{th} word appearing in the i^{th} tweet

Although the tweet often contains very few words (about 7 words) but large amounts of the tweets has extremely large number of different words, in the order of hundred thousands. It leads to the representation matrix BoW being large amounts of high dimensional datapoints. These issues are really challenging for supervised learning [25].

2.4 Machine Learning in Tweets Classification

Machine learning in automated text classification [26], [27] is a relatively old study. Classifiers are built using one of the machine learning methods and trained on a BoW representation of texts in very-high-dimensional input space. [12] proposed MNB for the text categorization. The study of the linear SVM in text classification are found in [28], [29]. Their comparative study showed that the linear SVM outperforms Rocchio, k nearest neighbors (kNN), Naïve Bayes (NB), Bayesian network (BN) and decision trees in text classification. More recent researches in sentiment analysis of tweets [14], [1], [19], [20] used MNB, SVM and MaxEnt and applied them on unigrams, bigrams and parts of speech as features. Their experiments also demonstrated that SVM on unigrams returned the best result in classifying positive or negative sentiment. The investigation in our paper is to extend our boosting of SVM [30] to build forest algorithms for classifying large number of datapoints in very-high-dimensional input space.

3 Forests of Oblique Decision Stumps

Our forest algorithms aim at using multiple oblique decision stumps classifiers to form an ensemble of classifiers more accurate than a single one. The performance of ensemble-based learning algorithms is decomposed into two key measures, called bias and variance [31]. Bias is the systematic error term (independent of the learning sample) and variance is the error due to the variability of the model with respect to the learning sample randomness. And then, the success of ensemble classifiers is to reduce the variance and/or the bias in learning models. Bagging [7] and Random forest [6] aim at reducing the variance of a learning algorithm without increasing its bias too much. Boosting [32] and Arcx4 [8] try to simultaneously reduce the bias and the variance. These approaches illustrate how to build accurate models with practical relevance for classification.

Ensemble-based learning techniques use decision stumps [9] as "weak learners" (or "base learners"). A decision stump is an one-level decision tree consisting of the root node which is directly connected to the terminal nodes. The decision stump learning algorithm selects a single attribute (univariate) for node splitting as done by decision tree algorithms [33], [34]. Thus, the strength of decision stumps is reduced, particularly when dealing with datasets having dependencies among dimensions (see figure 1). Due to this situation, one can thus build oblique decision trees using the multivariate splitting criteria [35]. Recently ensemble of oblique decision trees [36] using SVM [10], has attracted much research interests.

Our forest algorithms constructs a collection of oblique decision stumps in the same framework of classical Bagging and Arcx4. The main difference is that each oblique decision stump in the forest uses the linear SVM for performing multivariate non-terminal node splitting (using the combination between attributes, instead of choosing a best one

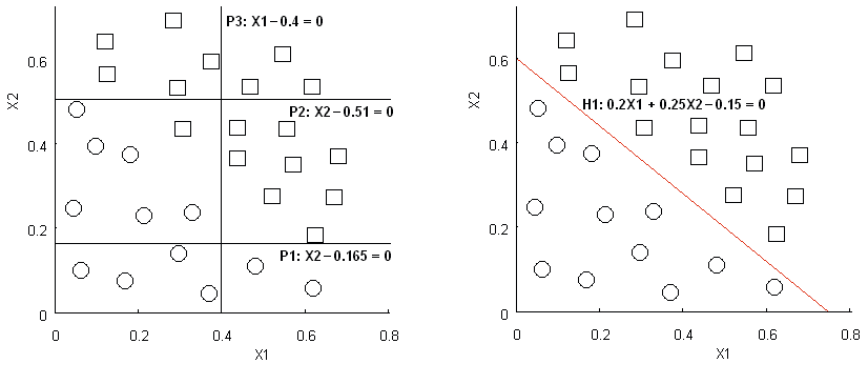


Fig. 1. Uni-variate (left) and bi-variate (right) node splitting

for node splitting). Our proposal is thus an hybridization of decision stump with SVM. SVM models are here used in the growing phase to create the oblique decision stumps.

3.1 Linear SVM for Oblique Decision Stumps

Let us consider a linear binary classification task, as depicted in Figure 2, with m datapoints x_i ($i = 1, \dots, m$) in the n -dimensional input space R^n , having corresponding labels $y_i = \pm 1$. For this problem, the SVM algorithms [10] try to find the best separating plane (denoted by the normal vector $w \in R^n$ and the scalar $b \in R$), i.e. furthest from both class $+1$ and class -1 . It can simply maximize the distance or the margin between the supporting planes for each class ($x \cdot w - b = +1$ for class $+1$, $x \cdot w - b = -1$ for class -1). The margin between these supporting planes is $2/\|w\|$ (where $\|w\|$ is the 2-norm of the vector w). Any point x_i falling on the wrong side of its supporting plane is considered to be an error, denoted by z_i ($z_i \geq 0$). Therefore, SVM has to simultaneously

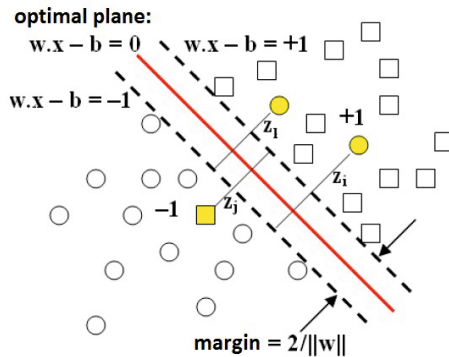


Fig. 2. Linear separation of the datapoints into two classes

maximize the margin and minimize the error. The standard SVMs pursue these goals with the quadratic programming of (1).

$$\begin{aligned} \min \Psi(w, b, z) &= \frac{1}{2} \|w\|^2 + C \sum_{i=1}^m z_i \\ \text{s.t. : } &y_i(w \cdot x_i - b) + z_i \geq 1 \\ &z_i \geq 0 \end{aligned} \quad (1)$$

where the positive constant C is used to tune errors and margin size.

The plane (w, b) is obtained by solving the quadratic programming (1). Then, the classification function of a new datapoint x based on the plane is:

$$\text{predict}(x) = \text{sign}(w \cdot x - b) \quad (2)$$

Unfortunately, the computational cost requirements of the SVM solutions in (1) are at least $O(m^2)$, where m is the number of training datapoints, making classical SVM intractable for large datasets. Therefore, we propose to use the stochastic gradient descent algorithm (SGD) [37], [38] to deal with the SVM problem in the linear complexity $O(m)$.

The SVM problem in quadratic programming (1) is reformulated in an unconstrained problem. We can ignore the bias b without generality loss. The constraints $y_i(w \cdot x_i) + z_i \geq 1$ in (1) are rewritten as follows:

$$z_i \geq 1 - y_i(w \cdot x_i) \quad (3)$$

The constraints (3) and $z_i \geq 0$ are rewritten by the hinge loss function:

$$z_i = \max\{0, 1 - y_i(w \cdot x_i)\} \quad (4)$$

Substituting for z from the constraint in terms of w into the objective function Ψ of the quadratic programming (1) yields an unconstrained problem (5):

$$\min \Psi(w, [x, y]) = \frac{\lambda}{2} \|w\|^2 + \frac{1}{m} \sum_{i=1}^m \max\{0, 1 - y_i(w \cdot x_i)\} \quad (5)$$

And then, [37], [38] proposed the stochastic gradient descent method to solve the unconstrained problem (5). The stochastic gradient descent for SVM (denoted by SVM-SGD) updates w on T epochs with a learning rate η . For each epoch t , the SVM-SGD uses a single randomly received datapoint (x_i, y_i) to compute the sub-gradient $\nabla_i \Psi(w, [x_i, y_i])$ and update w_{t+1} .

As mentioned in [37], [38], the SVM-SGD algorithm quickly converges to the optimal solution due to the fact that the unconstrained problem (5) is convex games on very large datasets. The algorithmic complexity of SVM-SGD is linear with the number of datapoints. An example of its effectiveness is given with the classification into two classes of 780 000 datapoints in 47 000-dimensional input space in 2 seconds on a PC and the test accuracy is similar to standard SVM one (ref. <http://leon.bottou.org/projects/sgd>). Therefore, we propose using SVM-SGD to perform oblique decision stumps.

3.2 Algorithms of Oblique Decision Stump Forests

Bagging of oblique decision stumps (illustrated in figure 3) constructs a collection of base learners in the same way of classical bagging [7]. An oblique decision stump (denoted by ODS) in a forest (Bagging, denoted by Bag-ODS) is independently learnt as follows:

- the training set is a bootstrap replica of m individuals, i.e. a random sampling with replacement from the original training set.
- at the root node, learning a linear SVM to perform a multi-variate splitting.
- the terminal nodes directly derived from the root node are labelled.

The classification of a new individual X uses a majority vote of oblique decision stumps.

We have also applied the Arcx4 algorithm [8] to ODS. Arcx4 of ODS (denoted by Arcx4-ODS, illustrated in figure 4) calls repeatedly a ODS learning algorithm k times so that each iterative step concentrates mostly on the errors produced by the previous steps. For achieving this goal, it needs to maintain a distribution weights over the training datapoints. Initially, all weights are set equally and at each iterative step the weights of misclassified datapoints are increased so that the ODS learner is forced to focus on the hard examples in the training set. Arcx4-ODS algorithm is described as follows:

- Initialize: distribution

$$d^0(i) = \frac{1}{\text{trainsize}}$$

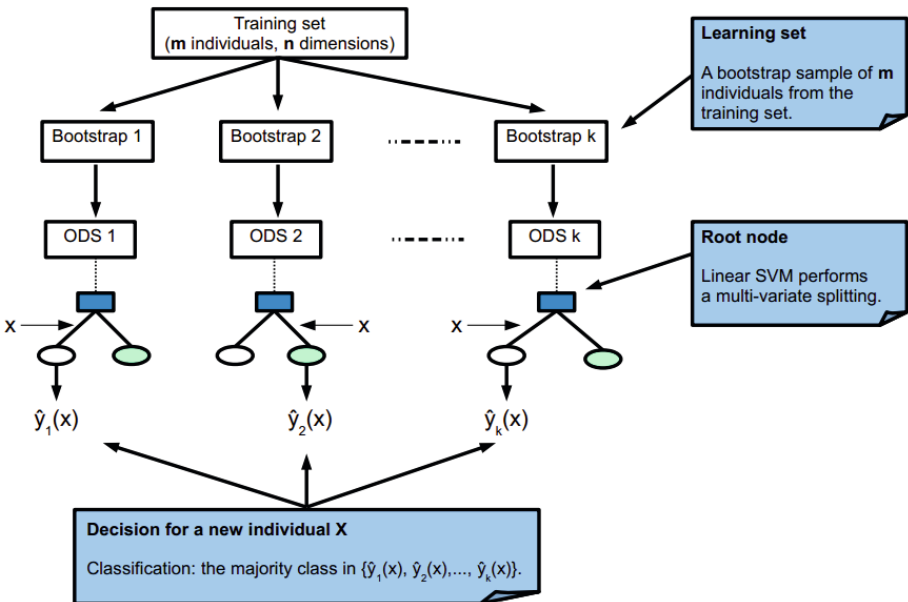


Fig. 3. Bagging of oblique decision stumps

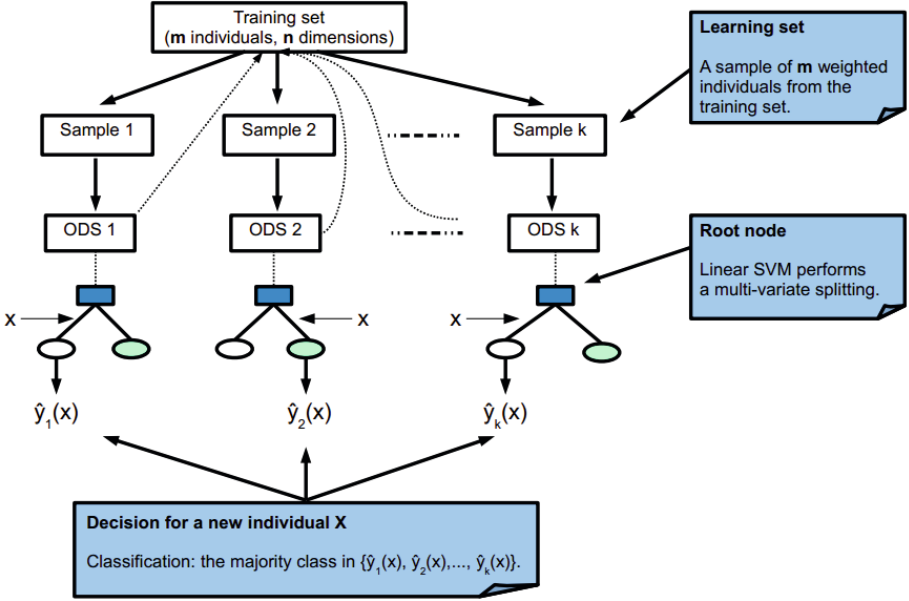


Fig. 4. Arcx4 of oblique decision stumps

- For k^{th} step
 1. sampling S^k is based on d^{k-1}
 2. training ODS^k model from S^k ,
 3. updating distribution

$$d^k(i) = \frac{(1 + m(i)^4)}{\sum_{p=1}^{trainsize} (1 + m(p)^4)}$$

where $m(i)$ is the number of misclassifications of the i^{th} datapoint by $ODS^1, ODS^2, \dots, ODS^k$

4 Evaluation

We are interested in the performance of the new oblique decision stump forests (Bag-ODS and Arcx4-ODS) for classifying large amounts of tweets. In order to compare performance for this Twitter sentiment classification, we have implemented Bag-ODS, Arcx4-ODS in C/C++ using the SGD library [38]. Due to baseline algorithms, we have implemented Maximum Entropy (denoted by MaxEnt) and Multinomial Naïve Bayes (denoted by MNB) in C/C++ and also use the highly efficient standard linear SVM algorithm LIBLINEAR [39].

Experiments are conducted with twittersentiment.appspot.com corpora collected by [11]. This corpora contains 1,600,000 tweets (800,000 tweets with positive emotions and 800,000 tweets with negative emotions). Pre-processing steps proposed by [1]

are used to perform the tweets cleaning, including convert Twitter usernames to token *USERNAME*, remove repeated letters. We use LibBoW [40] to build the BoW representation of this corpora. Without any feature selection, this yielded a dataset of 1,600,000 datapoints (tweets) in 244,895 dimensions (words). The dataset is randomly partitioned into a training set (1,066,667 datapoints, $\sim \frac{2}{3}$ full set) to learn the models and a testing set (533,333 datapoints, $\sim \frac{1}{3}$ full set) to test the models.

Our evaluation of the classification performance bases on the contingency table to compare the common measures: *F1* measure (Eq. 6), *Gmean* (Eq. 7) and *Accuracy* (Eq. 8) (global accuracy, positive accuracy *TP*. Rate and negative accuracy *TN*. Rate) [41]. The *Gmean* measure represents a trade-off between positive accuracy *TP*. Rate and negative accuracy *TN*. Rate.

$$F1 = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \quad (6)$$

$$Gmean = \sqrt{\frac{TP}{TP + FN} \cdot \frac{TN}{TN + FP}} \quad (7)$$

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \quad (8)$$

Where

- *TP* (true positive): the number of examples correctly labelled as belonging to the positive class;
- *FN* (false negative): the number of positive examples labelled as negative class;
- *TN* (true negative): the number of examples correctly labelled as belonging to the negative class;
- *FP* (false positive): the number of negative examples labelled as positive class.

Due to the BoW representation in very large number of dimensions, a linear SVM deals well with the classification task [29], [28], [42]. Our forest algorithms Bag-ODS, Arcx4-ODS build 50 oblique decision stumps² to perform classification on tweets. The main classification results are presented in table 2. The best results are bold faces and the second ones are underlined. The plot chart in figure 5 also visualises the classification results.

In comparison of classification results obtained by Bag-ODS, Arcx4-ODS facing the baseline algorithms, our Bag-ODS, Arcx4-ODS outperform MNB, SVM, MaxEnt, in terms of *F1* measure, *Gmean* and *Accuracy* because their classification models give a trade-off between positive accuracy *TP*. Rate and negative accuracy *TN*. Rate. However, SVM provides very competitive results in terms of *F1* measure, *Gmean* and *Accuracy*. MaxEnt gives the best result of positive accuracy *TP*. Rate while making many false positive errors (*FP*). In contrast to MaxEnt, MNB gives the best negative

² We remark that we tried to vary the number of oblique decision stumps from 10 to 500 for finding the best experimental results. And then, we obtained accurate models with 50 oblique decision stumps and it seems that the results are unchanged while increasing the number of oblique decision stumps over 50.

accuracy TN . Rate while making many false negative errors (FN). Therefore, MaxEnt and MNB are less accurate than other ones in terms of $F1$ measure, $Gmean$ and $Accuracy$.

Table 2. Classification results on twittersentiment.appspot.com corpora

Algorithm ↓	TP Rate	TN Rate	F1-measure	Gmean	Accuracy
MNB	76.88	74.57	75.83	75.72	75.72
SVM	78.18	73.67	76.28	75.89	75.91
MaxEnt	79.42	71.41	76.17	75.31	75.38
Bag-ODS	79.01	<u>73.90</u>	<u>76.86</u>	76.41	76.43
Arcx4-ODS	<u>79.18</u>	73.74	76.90	76.41	76.43

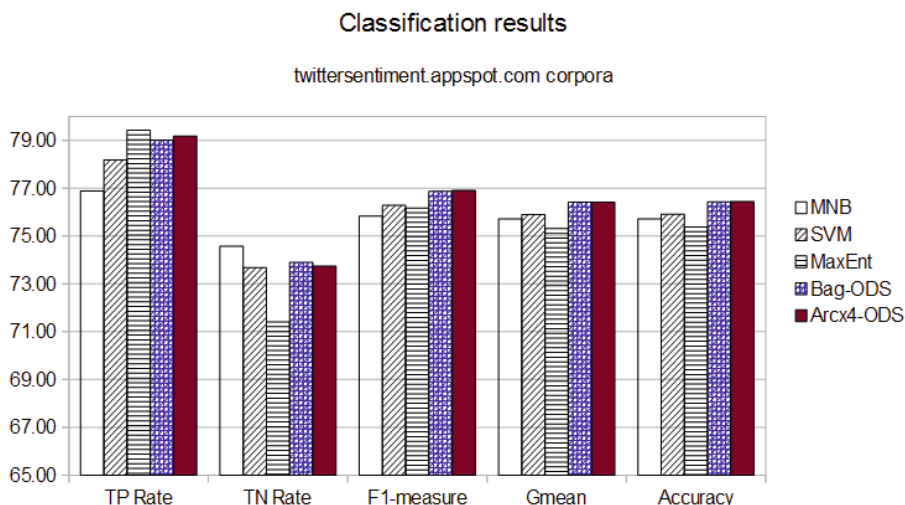


Fig. 5. Classification results on twittersentiment.appspot.com corpora

5 Conclusion and Future Works

We presented oblique decision stump forest algorithms that achieve high performances for classifying the sentiment in Twitter. The BoW representation of tweets leads to large amounts of very high dimensional datapoints. Therefore we propose oblique decision stump forests, Bag-ODS and Arcx4-ODS that is suitable for classifying large amounts of high dimensional datapoints. The numerical test results on twittersentiment.appspot.com corpora show that our forests are efficient compared to baseline algorithms, including MNB, MaxEnt and SVM.

The test results show that the effectiveness of Bag-ODS, Arcx4-ODS is not so large. The reason could be that the pre-processing steps (data cleaning and native BoW model) do not deal with the noisy in twittersentiment.appspot.com corpora. Another features

of tweets (e.g. like users, emotional icons) could be complementarily used to improve the classification performance. The drawbacks of the BoW model do not take into account the synonymy and the polysemy. These problems degrade the classification performance. In the future, we intend to use the latent Dirichlet allocation (LDA) model [43] or the lexical database WordNet [44] to deal with the synonymy and the polysemy arising from the BoW model. This can significantly improve the classification results of the sentiment in Twitter.

References

1. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. *Processing*, 1–6 (2009)
2. Barbosa, L., Junlan, F.: Robust sentiment detection on twitter from biased and noisy data. In: *Proceedings of the International Conference on Computational Linguistics, COLING 2010*. Association for Computational Linguistics (2010)
3. Pang, B., Lee, L.: *Opinion Mining and Sentiment Analysis*. Foundations and Trend. Now Publishers Inc. (July 2008)
4. Liu, B., Zhang, L.: A survey of opinion mining and sentiment analysis. In: Aggarwal, C.C., Zhai, C. (eds.) *Mining Text Data*, pp. 415–463. Springer US (January 2012)
5. Hassan, S.: *Sentiment analysis of microblogs mining the new world* (March 2012)
6. Breiman, L.: Random forests. *Machine Learning* 45(1), 5–32 (2001)
7. Breiman, L.: Bagging predictors. *Machine Learning* 24(2), 123–140 (1996)
8. Breiman, L.: Arcing classifiers. *The annals of statistics* 26(3), 801–849 (1998)
9. Wayne, I., Pat, L.: Minimizing the misclassification error rate using a surrogate convex loss. In: *Proceedings of the Ninth International Conference on Machine Learning, ICML 1992*, July 1-3, pp. 233–240. Morgan Kaufmann, CA (1992)
10. Vapnik, V.: *The Nature of Statistical Learning Theory*. Springer (1995)
11. Go, A., Bhayani, R., Huang, L.: Twitter sentiment, <http://help.sentiment140.com> (accessed date May 12, 2014)
12. Lewis, D.D., Gale, W.A.: A sequential algorithm for training text classifiers. In: *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 1994*, pp. 3–12. Springer-Verlag New York, Inc., New York (1994)
13. Berger, A.L., Pietra, V.J.D., Pietra, S.A.D.: A maximum entropy approach to natural language processing. *Computational Linguistics* 22(1), 39–71 (1996)
14. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In: *Proceedings of the ACL 2002 Conference on Empirical Methods in Natural Language Processing, EMNLP 2002*, vol. 10, pp. 79–86. Association for Computational Linguistics, Stroudsburg (2002)
15. Saif, H., He, Y., Alani, H.: Semantic sentiment analysis of twitter. In: Cudré-Mauroux, P., Heflin, J., Sirin, E., Tudorache, T., Euzenat, J., Hauswirth, M., Parreira, J.X., Hendler, J., Schreiber, G., Bernstein, A., Blomqvist, E. (eds.) *ISWC 2012, Part I. LNCS*, vol. 7649, pp. 508–524. Springer, Heidelberg (2012)
16. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2004*, pp. 168–177. ACM, New York (2004)
17. Read, J., Carroll, J.: Weakly supervised techniques for domain-independent sentiment classification. In: *Proceedings of the 1st International CIKM Workshop on Topic-sentiment Analysis for Mass Opinion, TSA 2009*, pp. 45–52. ACM, New York (2009)

18. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R.: Sentiment analysis of twitter data. In: Proceedings of the Workshop on Languages in Social Media, pp. 30–38. Association for Computational Linguistics, Stroudsburg (2011)
19. Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. In: Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D. (eds.) Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, Valletta, Malta, May 17–23, pp. 1320–1326. European Language Resources Association (2010)
20. Bifet, A., Frank, E.: Sentiment knowledge discovery in twitter streaming data. In: Pfahringer, B., Holmes, G., Hoffmann, A. (eds.) DS 2010. LNCS, vol. 6332, pp. 1–15. Springer, Heidelberg (2010)
21. Song, G., Ye, Y., Du, X., Huang, X., Bie, S.: Short text classification: A survey. *Journal of Multimedia* 9(5), 635–643 (May)
22. Do, T.-N., Moga, S., Lenca, P.: Random forest of oblique decision trees for ERP semi-automatic configuration. In: Sobecki, J., Boonjing, V., Chittayasothorn, S. (eds.) Advanced Approaches to Intelligent Information and Database Systems. SCI, vol. 551, pp. 25–34. Springer, Heidelberg (2014)
23. Harris, Z.S.: Distributional structure. *Word* 10, 146–162 (1954)
24. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. *Communications of the ACM* 18(11), 613–620 (1975)
25. Yang, Q., Wu, X.: 10 challenging problems in data mining research. *International Journal of Information Technology and Decision Making* 5(4), 597–604 (2006)
26. Sebastiani, F., Ricerche, C.N.D.: Machine learning in automated text categorization. *ACM Computing Surveys* 34, 1–47 (2002)
27. Manning, C.D., Raghavan, P., Schtze, H.: *Introduction to Information Retrieval*, 1st edn. Cambridge University Press (July 2008)
28. Joachims, T.: Text categorization with support vector machines: Learning with many relevant features. In: Nédellec, C., Rouveïrol, C. (eds.) ECML 1998. LNCS, vol. 1398, pp. 137–142. Springer, Heidelberg (1998)
29. Dumais, S., Platt, J., Heckerman, D., Sahami, M.: Inductive learning algorithms and representations for text categorization. In: Proceedings of the Seventh International Conference on Information and Knowledge Management, CIKM 1998, pp. 148–155. ACM, New York (1998)
30. Do, T.-N., Poulet, F.: Towards high dimensional data mining with boosting of PSVM and visualization tools. In: Proc. of 6th Intl. Conf. on Enterprise Information Systems, pp. 36–41 (2004)
31. Dietterich, T., Kong, E.B.: Machine learning bias, statistical bias, and statistical variance of decision tree algorithms. Technical report (1995), <http://datam.i2r.a-star.edu.sg/datasets/krbd/>
32. Freund, Y., Schapire, R.: A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence* 14(5), 771–780 (1999)
33. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: *Classification and Regression Trees*. Wadsworth International (1984)
34. Quinlan, J.R.: *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo (1993)
35. Murthy, S., Kasif, S., Salzberg, S., Beigel, R.: OC1: Randomized induction of oblique decision trees. In: Proceedings of the Eleventh National Conference on Artificial Intelligence, pp. 322–327 (1993)
36. Do, T.-N., Lenca, P., Lallich, S., Pham, N.-K.: Classifying Very-High-Dimensional Data with Random Forests of Oblique Decision Trees. In: Guillet, F., Ritschard, G., Zighed, D.A., Briand, H. (eds.) *Advances in Knowledge Discovery and Management*. SCI, vol. 292, pp. 39–55. Springer, Heidelberg (2010)

37. Shalev-Shwartz, S., Singer, Y., Srebro, N.: Pegasos: Primal estimated sub-GrAdient SOLver for SVM. In: Proceedings of the Twenty-Fourth International Conference Machine Learning, pp. 807–814. ACM (2007)
38. Bottou, L., Bousquet, O.: The tradeoffs of large scale learning. In: Platt, J.C., Koller, D., Singer, Y., Roweis, S. (eds.) Advances in Neural Information Processing Systems, vol. 20, pp. 161–168. NIPS Foundation (2008), <http://books.nips.cc>
39. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: Liblinear: A library for large linear classification. *Journal of Machine Learning Research* 9, 1871–1874 (2008)
40. McCallum, A.K.: Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering (1996), <http://www.cs.cmu.edu/~mccallum/bow>
41. Rijsbergen, C.J.V.: Information Retrieval, 2nd edn. Butterworth-Heinemann, Newton (1979)
42. Yuan, G.X., Ho, C.H., Lin, C.J.: Recent advances of large-scale linear classification. *Proceedings of the IEEE* 100(9), 2584–2603 (2012)
43. Blei, D., Ng, A., Michael, J.: Latent dirichlet allocation. *Journal of Machine Learning Research* 3, 993–1022 (2003)
44. Fellbaum, C.: WordNet: An Electronic Lexical Database (Language, Speech, and Communication). The MIT Press (May 1998)

Performance Evaluation of a Natural Language Processing Approach Applied in White Collar Crime Investigation

Maarten van Banerveld¹, Nhien-An Le-Khac², and M-Tahar Kechadi²

¹ Surinameweg 4, 2035 VA, Haarlem, The Netherlands
mj.van.barneveld@belastingdienst.nl

² School of Computer Science & Informatics, University College Dublin
Belfield, Dublin 4, Ireland
{an.lekhac, tahar.kechadi}@ucd.ie

Abstract. In today's world we are confronted with increasing amounts of information every day coming from a large variety of sources. People and corporations are producing data on a large scale, and since the rise of the internet, e-mail and social media the amount of produced data has grown exponentially. From a law enforcement perspective we have to deal with these huge amounts of data when a criminal investigation is launched against an individual or company. Relevant questions need to be answered like who committed the crime, who were involved, what happened and on what time, who were communicating and about what? Not only the amount of available data to investigate has increased enormously, but also the complexity of this data has increased. When these communication patterns need to be combined with for instance a seized financial administration or corporate document shares a complex investigation problem arises. Recently, criminal investigators face a huge challenge when evidence of a crime needs to be found in the Big Data environment where they have to deal with large and complex datasets especially in financial and fraud investigations. To tackle this problem, a financial and fraud investigation unit of a European country has developed a new tool named LES that uses Natural Language Processing (NLP) techniques to help criminal investigators handle large amounts of textual information in a more efficient and faster way. In this paper, we present briefly this tool and we focus on the evaluation its performance in terms of the requirements of forensic investigation: speed, smarter and easier for investigators. In order to evaluate this LES tool, we use different performance metrics. We also show experimental results of our evaluation with large and complex datasets from real-world application.

Keywords: Big data, natural language processing, financial and fraud investigation, Hadoop/MapReduce.

1 Introduction

Since the start of the digital information age to the rise of the Internet, the amount of digital data has dramatically increased. Indeed, we are dealing with many challenges

when it comes to data. Some data is structured and stored in a traditional relational database, while other data, including documents, customer service records, and even pictures and videos, is unstructured. Organizations also have to consider new sources of data generated by new devices such as sensors. Moreover, there are other new key data sources, such as social media, click-stream data generated from website interactions, etc. The availability and adoption of newer, more powerful mobile devices, coupled with ubiquitous access to global networks will drive the creation of more new sources for data. As a consequence, we are living in the Big Data era. Big Data can be defined as any kind of datasets that has three important characteristics: huge volumes, very high velocity and very wide variety of data. Obviously, handling and analysing large, complex, and velocity data have always offered the greatest challenges as well as benefits for organisations of all sizes. Global competitions, dynamic markets, and rapid development in the information and communication technologies are some of the major challenges in today's industry. Briefly, we have had a deluge of data from not only science fields but also industry, commerce and digital forensics fields. Although the amount of data available to us is constantly increasing, our ability to process it becomes more and more difficult. This is especially true for the criminal investigation today. For instance, a criminal investigation department CID of the Customs Force in a European country has to analyse around 3.5 Terabyte of data (per case) to combat fiscal, financial-economic and commodity fraud safeguards the integrity of the financial system and to combat also organized crime, especially its financial component.

Actually, CID staff focuses on the criminal prosecution of: Fiscal fraud (including VAT/carousel fraud, excise duty fraud or undisclosed foreign assets); Financial-economic fraud (insider trading, bankruptcy fraud, property fraud, money laundering, etc.); Fraud involving specific goods (strategic goods and sanctions, raw materials for drugs, intellectual property, etc.). Seizing the business accounts is usually the first step in the investigation. The fraud must be proved by means of the business accounts (among other things). Investigation officers not only investigate paper accounts, but also digital records such as computer hard disks or information on (corporate) networks. The fraud investigation unit uses special software to investigate these digital records. In this way we gain an insight into the fraud and how it was committed. Interviewing or interrogation of a suspect is an invariable part of the investigation. A suspect can make a statement, but may also refuse this. In any case, the suspect must be given the opportunity to explain the facts of which he is suspected. During their activities the investigation officers can rely on the Information-gathering teams for information and advice. These teams gather, process and distribute relevant information and conduct analyses. With digital investigations we respond to the rapid digitalization of society. This digitalization has led to new fraud patterns and methods, and all kinds of swindle via the Internet. In order to trace these fraudsters we use the same digital possibilities as they do.

As the CID handles around 450 criminal investigations every year, the amount of (digital-) data that is collected increases year over year. A specific point of attention is that the CID operates in another spectrum of investigations as 'regular' police departments. The types of crime that the CID needs to investigate mostly revolve around written facts. So the evidence that is collected by the CID by default contains of large amounts of textual data. One can imagine how much textual data a multinational firm

produces, and how many e-mails are being sent in such companies. A specific challenge for law enforcement departments that are involved with fraud investigations is: how can we find the evidence we need in these huge amounts of complex data. Because of the enormity of the large and complex data sets the CID seizes, it is necessary need to look for new techniques that make computers perform some analysis tasks, and ideally assist investigators by finding evidence. Recently, CID has developed a new investigation platform called LES. This tool is based on Natural Language Processing (NLP) techniques [1] such as Named Entity Extraction [2] and Information Retrieval (IR) [3] in combining with a visualization model to improve the analysis of a large and complex dataset.

In this paper, we evaluate the performance of LES tool because there are very few NLP tools that are being exploited to tackle very large and complex datasets in the context of investigation on white-collar crimes. Indeed, theoretical understanding of the techniques that are used is necessary. This theoretical review can help explain the usage of these new techniques in criminal investigations, and pinpoint what work needs to be done before the most effective implementation and usage is possible. The rest of this paper is organised as follows: Section 2 shows the background of this research including related work in this domain. We present briefly LES tool and evaluation methods in Section 3. We apply our method to analysis the performance of LES tool on a distributed platform in Section 4. Finally, we conclude and discuss on future work in Section 5.

2 Background

2.1 Natural Language Processing in Law Enforcement

NLP implemented techniques can be very useful in a law enforcement environment, especially when unstructured and large amounts of data need to be processed by criminal investigators. Already commonly used techniques like Optical Character Recognition (OCR) [4] and machine translations [5] can be successfully used in criminal investigations. For example OCR is used in fraud investigations to automatically transform unstructured invoices and other financial papers into searchable and aggregated spread sheets. In the past more difficult to implement techniques like automatic summarization of texts, information extraction, entity extraction and relationship extraction [1] are now coming into reach of law enforcement and intelligence agencies. This is mainly so because of the decline in cost per processing unit and the fact that these techniques need a large amount of processing power to be able to used effectively.

To zoom in on this a little further: for example the extraction of entities out of large amounts of text can be useful when it is unclear what persons or other entities are involved in a criminal investigation. Combined with a visual representation of the present relations between the extracted entities, this analysis can provide insight in the corresponding (social-) networks between certain entities. Indeed, the usage of NLP techniques to ‘predict’ criminality, for example grooming by possible paedophiles [6] or trying to determine when hit-and-run crimes may happen by analysing Twitter messages [7] is possible today. A movement from single available NLP

techniques like text summarization, text translation, information and relationship extraction towards more intelligent NLP based implementations for law enforcement like crime prediction, crime prevention, criminal intelligence gathering, (social-) network analysis and anomaly detection can be observed in literature. Also theoretical frameworks and models in the field of ‘forensic linguistics’ [8] are proposed which can be used behind the technical implementation of NLP techniques in criminal investigations.

When (commercial-) solutions using these techniques come available, this could lead to more extensive NLP based law enforcement systems that can handle Crime prediction, deliver automated intelligence on criminal activities, analyse the behaviour of subjects on social networks and detect anomalies in texts or other data. The output of these systems is ideally presented in a visual comprehensible way so the criminal investigator can quickly assess the data and take appropriate action.

2.2 Big Data in Criminal Investigations

No strict definition can be given for the concept Big Data [9] as such, but what can be concluded is that Big Data at least has some common elements and that Big does not necessarily mean large volumes. Complexity and the inner structure of the data are also very important to determine if a dataset belongs to the concept of Big Data or not. Another term that is commonly used when people talk about ‘Big Data’ is ‘Unstructured Data’. As law enforcement we are confronted with at least parts of the Big Data problem; for instance in fraud investigations *the fraud investigation unit* regularly seizes a complete company (network-) environment including cloud storage and all belonging data. Because this data for the *fraud investigation unit* as outsiders is unstructured, and from a variety of sources (computer images, servers, internal communication, wiretap data, databases etc.) these datasets fall under the definition, and elements, of Big Data in terms of volume and complexity (also known as variety of the data). But also a very large e-mail database containing millions of suspect e-mails can fall under the Big Data problem because of the complexity of this data set. Please note that in most descriptions Big Data is measured against three axes: Volume, Variety and Velocity. What we see is that in the *fraud investigation unit’s* types of investigation, the focus is mostly on the volume and variety of the large data set. Velocity is not really an issue as they are investigating a static data set. This is so because after seizure the data that needs to be investigated will not (rapidly) change anymore.

What can be said is that the existence of Big Data poses new and unique challenges for law enforcement when evidence needs to be found in an investigation with these characteristics. What also can be said is that not only the actual size of the total seized data matters, but also the rate of complexity of the data that determines if a case falls under a Big Data definition.

As an example, in a large carousel fraud case that the *fraud investigation unit* investigated in the past, the suspect was a bank that operated from *the fraud investigation unit’s* territory and several countries abroad. In this case investigation data was collected and seized from a lot of sources: internet wiretaps, forensic disc images from tens of workstations, user data from server systems, e-mail servers with literally millions of e-mails, company databases, webservers, and the complete banking

back-end systems containing all bank transactions. This investigation had the characteristics of Big Data on both levels, a high complexity of the data (the complete banking system had to be reconstructed and analysed) and a high amount of total data (the house searches were in 2006, and in that time a total of 15 terabyte of data was seized).

This paper is about the usage of NLP techniques in fraud investigations, there are specific characteristics for these types of investigations that determine why another approach towards Big Data investigation is necessary. In fact our *fraud investigation unit* mostly investigates White Collar Crime cases. Most police departments focus on other criminal offenses like murder cases, child abuse, threats, hacking, malware etc. The *fraud investigation unit* on the other hand acts on criminal cases like money laundering, terrorism funding, (tax-) fraud, etc. For the *fraud investigation unit* this focus on White Collar crime means that the *fraud investigation unit* has to be able to investigate: (i) Complex (unstructured-) datasets; (ii) Large datasets; (iii) Company networks; (iv) Complex communication networks between suspects; (v) Mostly text based evidence.

As you can see, this list shows that the *fraud investigation unit* will encounter Big Data problems *because* of the specific criminal investigation domain, and that the evidence the *fraud investigation unit* gathers is *mostly* text based. Before the introduction of NLP techniques running on the new *fraud investigation unit* platform LES, the *fraud investigation unit* had massive problems with handling the enormous amounts of data that are so specific for white-collar crime investigations. These problems can be summarized in:

- Time taken to process all data that was seized
- Forensic software not able to handle the huge amounts of data items coming from for instance e-mail databases
- Crashing software when querying the investigation tooling database, because of overload
- Unacceptable waiting time for investigators when performing a query on the data (up to 30 minutes per query)
- Too many search hits to make analysis of the evidence humanly possible in many cases
- Too much technical approach and interfacing for regular investigators by currently used tooling

What also can be observed is that most police cases can make use of the Digital Forensics methodology and tooling as is described in literature [10]. Unfortunately the *fraud investigation unit* has to use tooling that is best suitable for criminal investigations falling under Police Types of crime where evidence can be found in/from files, desktop/mobile devices, email, network/memory analysis, etc.

2.3 Related Work

There are very few researches of NLP in the context of Digital Forensics, especially to tackle the problem of Big Data of financial crimes. In the context of Digital Forensics, [11] used NLP techniques to classify of file fragments. In fact, they use support

vector machines [12] along with feature vectors consisted of the unigram and bigram counts of bytes in the fragment. The method proposed is efficient; it is however, not in the context of investigating documents related to financial crimes. In [13], authors proposed a corpus of text message data. This corpus can support NLP techniques in investigating data located on mobile devices. This corpus is very useful in analysing short text but it is not for long, complex documents such as MS word document, presentations, spread sheets, etc. Related to the forensics financial crimes, [14] proposed a semantic search based on text mining and information retrieval. Authors however focus on documents from collaboration platform such as e-mail, forum as well as in social networks. Their main objective is how to optimise the searching queries.

3 LES Tool and Method of Evaluation

In this section, we present briefly LES, a NLP based tool that has been developed to study the possibilities and benefits the usage of NLP techniques can provide in complex fraud investigations. Next, we describe the investigating process where we apply LES tool to analyse evidence files. Finally we present methods we used to evaluate this tool.

3.1 LES Tool

Because of the problems of handling Big Data investigations mentioned earlier, our fraud investigation unit decided to develop tooling in-house that would be able to handle these specific types of investigations. The three most important requirements for the new tool are:

- Improving the data processing time, to handle large amounts of data
- Improving the data analysis time needed, to handle complex datasets
- Enable end users to perform complex tasks with a very simple interface

This tool was called LES (Figure 1) and its main characteristics are:

- Running on an Apache Hadoop platform [15]
- Ability to handle large amounts of data
- Use NLP techniques to improve evidence finding
- Visualisation of found (possible-) evidence
- A simple web based GUI with advanced search capabilities

In house developed software components allow investigators to rapidly access forensic disk images or copied out single files. MapReduce [16] jobs are then executed over the data to make parallel processing possible over multiple server nodes. Other MapReduce jobs are built in LES tool for text extraction and text indexing. At this moment the following NLP techniques are implemented in LES:

- Information extraction
- Named Entity Recognition (NER)
- Relationship Extraction

The Information and NER extraction process uses a combination of techniques to extract useful information: tabular extraction (for lists of known and described entities), regular expression extraction, and the Stanford NER library also known as CRFClassifier [17]. The relationships between entities are arbitrarily determined by the distance between these entities. If a distance is smaller than a threshold, a relationship between two entities is stored in the LES system (a Hadoop cluster of computers running LES tool). This implementation of relationship extraction is based on co-reference between words, which in system tests appears to perform quite well.

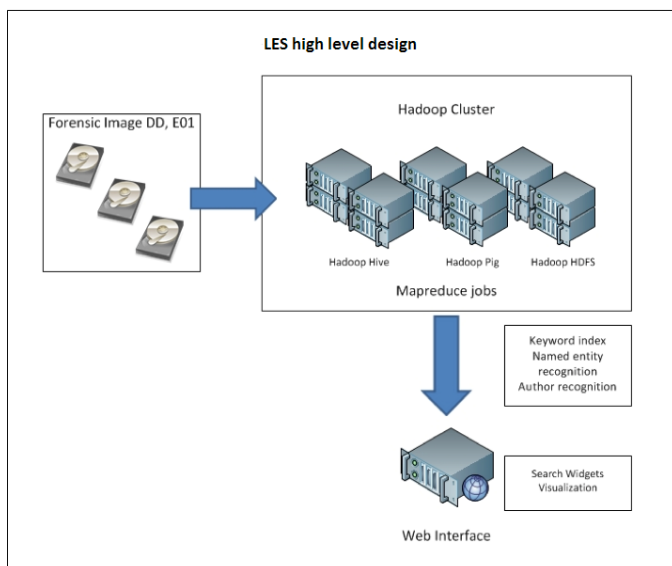


Fig. 1. High level design of LES tool

3.2 Investigation Process

Evidence files are imported into the LES system by running specific MapReduce jobs in a predefined sequence:

- 1) Prepare Evidence
- 2) Extraction phase
- 3) Indexing phase
- 4) NER extraction phase
- 5) Relationship Extraction
- 6) Analysing

The evidence acquired during house-searches by the digital investigators is mainly recorded in a forensic format like raw *dd* files or Encase format. During the preparation phase the evidence containers are mounted and integrity is checked. Then by default only the most relevant files are extracted for further processing. At this moment these files are the most common document and textual types and e-mail data.

All files that need to be investigated by LES tool are placed in a so-called binary ‘blob’ or data stream on the Hadoop cluster. Pointers to the original files are recorded in the file index on the cluster. This makes later viewing and retrieval of the original file much easier. When all extracted files are present in the data stream the indexing job is run. Next, the NER and RE phase are performed and finally all results are imported in the LES Hadoop Elastic search environment.

3.3 Methodology

Our evaluation method is based on the combination of Microsoft’s MSDN performance testing methodology [18], TMap NEXT [19] from Sogeti and some custom evaluation items (quality characteristics). The combination of different methodologies has led to the following concrete test case parameters that were evaluated:

- Test data set processing time, split in time to generate NER, extract relations, generate keyword index
- Test data set query response times
- Accuracy of the data retrieval: cross referencing with standard tooling
- Data controllability: completeness of the data, evidence integrity and chain of evidence reproducibility

4 Experiments and Analysis of Results

In this section, we describe firstly the dataset we used in our experiments. We also show the platform where we performed our tests. Finally, we present and analyse the results of these experiments.

4.1 Dataset

The dataset that was used is applicable for the two dimensions Volume and Variety (Complexity) of Big Data. Velocity is not an issue for our experiments at this stage. The data set that is used contains historical data from the period 2006 – 2012.

The testing dataset consisted of:

- Total size of dataset: 375GB
- Disk images (Encase E01): 292 disk images
- Microsoft Exchange mail databases: 96GB
- Office documents: 481.000
- E-mails: 1.585.500
- Total size of documents to investigate: 156GB
- Total size of extracted textual data: 21GB

As we are looking for evidence in a fraud case we can expect that most incriminating content can be found in textual data, coming from documents, e-mails etc. LES will automatically extract all files containing textual information out of file containers like Encase images, Exchange databases, zip files etc.

Next, from these files all text is extracted leading to a total size of pure flat text of 21 GB out of a total dataset of 375 GB. As this investigation was performed in the past, we today know that finding and processing the evidence that was needed, took a total of six years investigation. Because the amount of total items to investigate, and the complexity of this dataset, the time needed for this investigation took a lot longer than was thought of at the start. Some statistics can be found as follows:

- Evidence items found in dataset: 2.718
- Total textual items in test dataset: 2.144.254
- Percentage of evidence found versus total textual items: $0,126\% (2718 / 2.144.254) \times 100 = 0,126\%$

As we can see, the percentage of usable evidence for this case was only 0,126 percent. This indicates the needle in the haystack problem we are facing for these types of investigations.

4.2 Testing Platform

The testing system is a cluster consists of 14 physical servers with the following roles:

- 2x Hadoop Namenode (for redundancy purposes)
- 6x Hadoop Datanode (to store the data on)
- 1x Hadoop Edgenode (for cluster management)
- 4x Index nodes (to process the data)
- 1x webserver (for the end-user GUI)

Hadoop processing and storage:

- 18TB storage
- 24 cores Intel Xeon E5504

Index nodes processing and storage:

- 12TB storage
- 12 cores Intel Xeon E5504

Total cluster internal memory is 256GB. The cluster has been build and configured according to the Hadoop recommendations for building a Hadoop cluster.

4.3 Result Description and Analysis

We evaluate first of all the performance perspective of LES tool. We compare the processing time between Forensic Toolkit (FTK) [20] and LES tool on the same testing dataset. FTK has been configured in such a way that it approached the LES way of processing data the most. That means that all images and container items were read in FTK, but all extra options were disabled to make comparison fairer (Figure 2). As you can see, FTK has been configured to not perform Entropy test, carving, OCR. Indeed, only possible textual files were added as evidence to the case (documents, spreadsheets, e-mail messages). Only from this selection FTK was allowed to make a keyword based index.

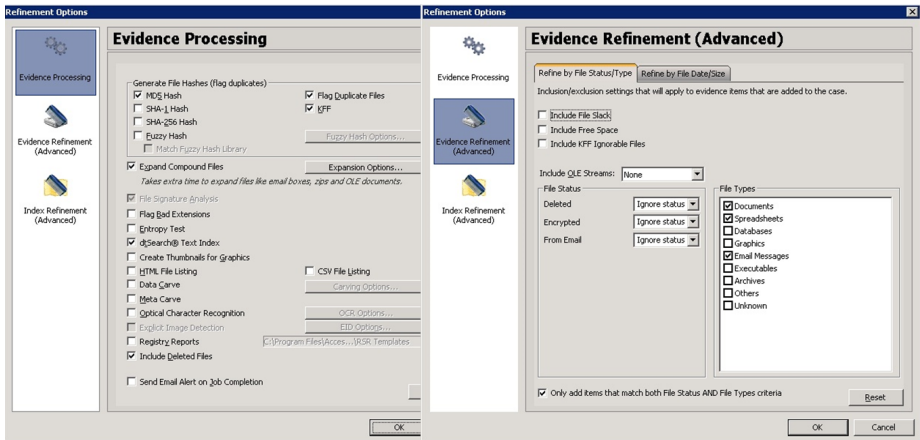


Fig. 2. FTK case configuration

According to the FTK processing log, the FTK processing time of the testing dataset, with the criteria shown above is 10 hours and 54 minutes. For LES tool, the total processing time is 1 hour and 24 minutes including 34 minutes of text extraction, 38 minutes of generating NLP NER databases and 12 minutes of generating the searchable keyword index based on the testing datasets.

In fact, the LES tool was evaluated by running various experiments on the testing datasets. As we can see, the overall processing time of this tool is 6 times faster than FTK with the same testing datasets. Furthermore, when LES is configured to only create a keyword based index, similar to Forensic Toolkit, the LES tool is even 11 times faster than FTK running on the same datasets. LES does however need extra processing time to perform the needed NLP calculations, but enhances the ease of finding evidence by running these NLP processes.

Besides, response times are significantly better when LES is used to search through the test data by using single keywords. Table 1 shows the response time of keyword searching. FTK shows some remarkable slow response times when using single keywords to search for, whereas LES in most cases is a factor 1000 or faster when searching through the data.

Furthermore, the retrieval times in LES tool are always under 0.5 seconds, but cannot be shown in a counter, and therefore difficult to give an exact count in milliseconds.

However, this is not a very efficient approach to find unique evidence items, most of the time using only one keyword leads to long lists of results. Also, it is very difficult to make up the best fitting keyword to find the evidence. Normally a combination of keywords or multiple search iterations is used, but in practice our investigators start hypothesis building by trying single keywords and see what comes back as a result. What can be seen from the FTK evaluation is that when using single keywords when trying to pinpoint evidence, the retrieval time can be very long. When using single keywords only and keywords are not chosen well or are not unique enough, waiting time becomes unacceptable long from an end-user perspective when we choose a response time of 20 seconds maximum. Of course investigators also need to be

trained to perform smart search actions when using FTK. It is essential to choose keyword combinations well. Next, we use the five known evidence items to locate and use an AND combination of keywords to evaluate response time and retrieval time of evidence items. Table 2 shows the results of this experiment.

Table 1. Response time of single keyword search

<i>Document</i>	<i>Keywords</i>	<i>FTK</i>		<i>LES</i>	
		<i>Response Time(s)</i>	<i>Retrieve Time(s)</i>	<i>Response Time(s)</i>	<i>Retrieve Time(s)</i>
<i>D-195</i> (.msg)	discontinue	6	27	0.005	< 0.5
	fraud	3	41	0.078	< 0.5
	revenue	6	287	0.119	< 0.5
<i>D-550</i> (.ppt)	scrubbing	6	17	0.081	< 0.5
	blacklists	5	14	0.099	< 0.5
	violation	6	365	0.069	< 0.5
<i>D-718</i> (.xls)	[NAME1]	18	295	0.138	< 0.5
	training	6	383	0.143	< 0.5
	crimecontrol	7	4	0.195	< 0.5
<i>D-735</i> (.msg)	[NAME2]	3	22	0.006	< 0.5
	[NAME3]	5	52	0.023	< 0.5
	NewYork	5	195	0.141	< 0.5
<i>D-805</i> (.txt)	[NAME4]	5	4	0.038	< 0.5
	Bermuda	5	1190	0.091	< 0.5
	[NAME5]	5	33	0.020	< 0.5

Table 2. Response time of combined keyword search

<i>Document</i>	<i>Keywords</i>	<i>FTK</i>		<i>LES</i>		
		<i>Response Time(s)</i>	<i>Retrieve Time(s)</i>	<i>Response Time(s)</i>	<i>Retrieve Time(s)</i>	
<i>D-195</i> (.msg)	discontinue, fraud, revenue	28	7.1	0.006	< 0.5	
	<i>D-550</i> (.ppt)	scrubbing, blacklists, violation	24	5.2	0.138	< 0.5
		<i>D-718</i> (.xls)	[NAME1], training crimecontrol	10	18	0.195
<i>D-735</i> (.msg)	[NAME2], [NAME3], NewYork		11	5	0.232	< 0.5
	<i>D-805</i> (.txt)	[NAME4], Bermuda, [NAME5]	16	4	0.004	< 0.5

When a combination of keywords is used we see that the response times for FTK are worse than single keyword search. On the other hand, using multiple keywords in LES the response time is also milliseconds for the performed evaluations. From an end-user perspective a response in milliseconds is more or less instant and thus leading to a better investigation experience. When the various NLP techniques are used to search for evidence, the LES tool response times are also in milliseconds. For instance the selection of named entities and the drawing of a relation diagram are performed very fast by the system.

Next, we evaluate the total amount of processed evidence items per file type analysing system and processing log files for FTK and LES (Table 3).

Table 3. Total number of processed evidence items per file type

<i>Document Type</i>	Number of items per file type	
	<i>FTK</i>	<i>LES</i>
<i>Email</i>	1.585.500	1.641.063
<i>Word documents</i>	44.105	44.837
<i>Spreadsheets</i>	68.101	38.580
<i>Presentations</i>	6.548	2620

As we are not sure how FTK counts evidence items, and which types are counted and which are not, it is difficult to draw a conclusion from these figures. But what we do know is that FTK counts for instance every OLE object item as a unique evidence item for Microsoft Office documents. So that increases the count for FTK significantly. However, since we have found all our randomly selected evidence items in both FTK and LES we can be carefully positive that no essential data is lost in LES.

Looking at the functionality perspective, LES has more possible search paths towards an evidence item; this could mean that evidence can be found faster using LES, because an investigator has more chance ‘hitting’ a useful search path. This coincides with the fact that in LES evidence can be found in more ways, because more search methods are implemented. These search methods increase the ways investigators can search for evidence. Especially the implemented NLP entity selection in combination with other search methods creates new evidence finding possibilities that previously were not possible. When looking at the data presentation parts of the software evaluation we can see that LES has more ways of presenting data to the investigator; the visualization view of found evidence can help investigators finding new leads. Another important functional part is the integrity and chain of evidence of the data. What can be seen is that FTK has better data control embedded, thus in FTK the chain of evidence is maintained more thoroughly. Also, FTK has better file control embedded; tracing back a file to its originating location is better implemented in FTK than in LES, thus the chain of evidence is maintained better. A big advantage of LES is that LES has been developed with the end-user in mind, in this case a financial and fraud investigator who needs to investigate a Big Data set. Specifically the LES query interface is very flexible and helps analyzing complex and large data sets, especially the possibility to add query windows (widgets) and refine searches by doing that is very powerful.

Specific evaluation requirements that were mostly focused on the implementation and usage of NLP techniques show that the implemented NLP techniques can help investigators finding evidence in another way, possibly faster and more efficient. At the minimum a new view towards complex data is presented for investigators. LES requires less search iterations to find evidence, because of the implementation of NLP NER and visualisation of evidence. On the other hand, FTK's keyword based search requires investigators to work through more data and refine search queries a lot of times.

Indeed, some noteworthy points that also came up during the evaluation were for instance that it was difficult to find literature that evaluates Accessdata forensic tool-kit on a performance and data controllability level. It looks like this tooling has not been evaluated very thoroughly yet by a respectable authority. For Hadoop/ MapReduce techniques we found that the usage of a Hadoop cluster seems to be very efficient when one needs to process large amounts of textual data. However, the programming paradigm of Hadoop/MapReduce are more complex than regular programming problems because of the distributed and multi-processing nature of the Hadoop cluster. The issues that we found during the evaluation were that a (too-) large edges and nodes file leads to graphical representation problems. Too much named entities and extracted relations leads to information overload for the end-user. The forensic chain of evidence is more difficult to maintain in LES. This is because of the nature of LES' inner workings, and the fact that it extracts textual information out of forensic images.

At organization level, we found that the CID will need to explain the difference between forensic computer investigation and analysis of Big Data. When to use what tool all depends on the type of investigation, the needed evidence, and the amount and complexity of the data. As the CID mainly has large fraud cases, a logical choice would be to use LES as the preferred tool for these kinds of investigations. One remark that must be made is that all data found in LES must be verified using a (forensic-) tool until LES has a proven track record in court of law.

As a conclusion, the usage of LES tool that uses NLP as key enabler to handle very large and complex data investigations. This means LES tool improves the 'white collar crime' investigation process in terms of speed and efficiency.

5 Conclusions and Future Work

In this paper, we present and evaluate LES tool that is based on NLP techniques to help criminal investigators handle large amounts of textual information. In fact, we evaluate different perspectives of LES tools. In terms of speed: the proposed solution is significantly faster in handling complex (textual) data sets in less time compared to traditional forensics approach. In terms of efficiency: the proposed solution is optimized for the fraud investigation process. The usage of NLP techniques helps in optimizing the investigation process. Investigators have more possibilities finding evidence in very large and complex dataset, aided by smart NLP based techniques. This greatly improves fraud investigation efficiency.

Some topics for further scientific and practical research is coming up. In terms of LES tool, more functions have being added such as automatic summarization of texts, author recognition, detection of cover language, detection of communication patterns, language detection, adding fraud domain knowledge to a NLP language corpus, visualisation of searching results, etc. Therefore, we will also evaluate the performance of these upcoming features.

References

1. Liddy Elizabeth, D.: Natural Language Processing, 2nd edn. Marcel Decker, Inc., NY (2001); In Encyclopedia of Library and Information Science
2. Tjong, K.S., Erik, F.: Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In: Proc. Conference on Natural Language Learning, Edmonton, Canada (June 2003)
3. Rijsbergen, C.J.: Information Retrieval, 2nd edn. Butterworths (1979)
4. Lais S.: Quick Study: Optical Character Recognition. Computer World (June 25, 2014), http://www.computerworld.com/s/article/73023/Optical_Character_Recognition
5. (June 25, 2014), <http://www.systransoft.com/systran/corporate-profile/translation-technology/what-is-machine-translation/>
6. Jurafsky, D., Martin James, H.: Speech and Language Processing - An Introduction to Natural Language Processing, 2nd edn. Pearson Prentice Hall, Stanford University (2009)
7. Fromkin, V., Rodman, R., Hyam, N.: An Introduction to language, 9th edn. Wadsworth (2011)
8. Rafferty, A.N., de Marneffe, M.-C., Manning, C.D.: Finding Contradictions in Text. In: ACL 2008 (2008), <http://nlp.stanford.edu/pubs/contradiction-acl08.pdf> (June 25, 2014)
9. Sokol, L., Ames, R.: Analytics in a Big Data Environment. IBM Redbooks (2012)
10. Innis Tasha, R., et al.: Towards Applying Text Mining and Natural Language Processing for Biomedical Ontology Acquisition. In: TMBIO 2006 Proceedings of the 1st International Workshop on Text Mining in Bioinformatics, pp. 7–14 (2006)
11. Fitzgerald, S., et al.: Using NLP techniques for file fragment classification. Digital Investigation (9) (2012)
12. Scholkopf, B.: A short tutorial on kernels. Microsoft Research, Tech Rep: MSR-TR-200-6t (2000)
13. O'Day, D.R., Calix, R.A.: Text Message Corpus: Applying Natural Language Processing to Mobile Device Forensics. In: IEEE International Conference on Multimedia and Expo, San Jose, USA, July 15 - 19 (2013)
14. Van Dijk, D., Henseler, H.: Semantic Search in E-Discovery: An Interdisciplinary Approach. In: Workshop on Standards for Using Predictive Coding, Machine Learning, and Other Advanced Search and Review Methods in E-Discovery, ICAIL 2013 (2013)
15. (June 25, 2014), <http://hadoop.apache.org>
16. Dean, J., Ghemawat, S.: MapReduce: Simplified Data Processing on Large Clusters. In: OSDI 2004: Sixth Symposium on Operating System Design and Implementation, San Francisco, CA (December 2004)

17. Popowitch, F.: Using text mining and natural language processing for health care claims processing. ACM SIGKDD Explorations Newsletter - Natural Language Processing and Text Mining 7(1), 59–66 (2005)
18. Meier, J.D., et al.: Microsoft Performance Testing Guidance for Web Applications. Redmond (2007),
<http://msdn.microsoft.com/en-us/library/bb924375.aspx>
19. Buist, A.H., Kraaij, W., Raaijmakers, S.: Automatic Summarization of Meeting Data: A Feasibility Study. In: Proceedings of the 15th CLIN Conference (2005)
20. (June 25, 2014),
<http://www.accessdata.com/solutions/digital-forensics/ftk>

An Efficient Similarity Search in Large Data Collections with MapReduce

Trong Nhan Phan¹, Josef Küng¹, and Tran Khanh Dang²

¹ FAW Institute, Johannes Kepler University Linz, Austria
{nphan, jkueng}@faw.jku.at

² HCMC University of Technology, Ho Chi Minh City, Vietnam
khanh@cse.hcmut.edu.vn

Abstract. The era of big data has been calling for many innovations on improving similarity search computing. Such unstoppable large amounts of data threaten both processing capacity and performance of existing information systems. Joining the challenges on scalability, we propose an efficient similarity search in large data collections with MapReduce. In addition, we make the best use of the proposed scheme for widespread similarity search cases including pairwise similarity, search by example, range query, and k-Nearest Neighbor query. Moreover, collaborative strategic refinements are utilized to effectively eliminate unnecessary computations and efficiently speed up the whole process. Last but not least, our methods are enhanced by experiments, along with a previous work, on real large datasets, which shows how well these methods are verified.

Keywords: Similarity search, large datasets, MapReduce, Cosine, Hadoop.

1 Introduction

Similarity search has spread so many real-world applications whose core objective is to find objects that are similar to one another. It is, however, a time-consuming process in order to successfully achieve the goal. Considering the inceptive pairwise similarity search whose demand is to do the self-join of all possible pairs gives an exponential complexity. Such a high cost hardly makes itself response a query in time as well as meet a user's needs. The issue is especially getting harder and harder when data keeps rapidly growing up. When the digital world and its technologies are enormously developed, data come from anywhere; from world-wide web, history, and social networks to automation processes, mobile devices, and sensors. Dealing with a large amount of data not only puts a serious challenge on similarity search but also impose a potential risk toward traditional processing mechanisms.

We cannot deny the important role of similarity search in a wide range of applications. Its relevant issues have, therefore, gained much attentions world-wide. Different kinds of indexes or approximate but efficient ways are showed to tackle these issues [5][6]. Moreover, state-of-the-arts take the advantage of parallel mechanism either by optimizing parallel algorithms [1] or by deploying computations on a novel parallel paradigm like MapReduce [1][4][8][12][14][15] to improve large-scale similarity search when data size keeps growing. Engaging in the new trend where

the era of big data has come, we propose an efficient similarity search in large data collections with MapReduce. Our main contributions are summed up as followings:

1. A general similarity search scheme toward scalability is presented.
2. Collaborative strategic refinements, which reduce a large amount of candidate size leading to eliminating unnecessary computing and costs, are effectively integrated into MapReduce phases.
3. The proposed scheme flexibly adapts itself to well-known similarity searches including pairwise similarity, search by example, range search, and k-Nearest Neighbor (kNN) search.
4. These methods are consolidated by experiments with real datasets from DBLP [7] and Apache Hadoop Framework [3]. Furthermore, these methods are compared to the work in [10], which shows how much beneficial they might get while data size never stops increasing.

The rest of the paper is organized as follows: Section 2 shows related work that is pretty close to our approach. Section 3 introduces basic concepts associated with our current work. Next, we propose the general scheme in section 4 and how the scheme is applicable to diverse similarity search cases in section 5. Then, relevant experiments and analytics are given in section 6 before we make our final remarks and the future work in section 7.

2 Related Work

The traditional similarity search like pairwise similarity calls for full computations for all possible candidate pairs in the corpus, which brings about the complexity as $O(n^2)$. Such a high cost is not suitable for either real-time processing or data-intensive applications. Much work focuses on this issue and tries to reduce the search space by various filtering-based or approximate approaches. In [11], the authors introduce a method aiming at fast similarity search in very large string sets called State Set Index. The index is interpreted as a nondeterministic finite automaton in which each character of a string is processed to map with a state, and the last character defines the accepting state. Nevertheless, the whole process is sequentially taking place. The authors from the work in [16] make use of the positional filtering principle, which is combined with prefix and suffix filtering to efficiently do similarity joins for near duplicate detection. Again, no parallelism is taken into account. Another work in [17] shows the combination between index structure based method, k-means clustering tree, for prune strategy and a hashing based method, hamming distance, for fast distance computation. However, these methods are step-by-step done without any parallel mechanism. Moreover, only k-Nearest Neighbor query is taken into consideration.

Meanwhile, the authors in [15] propose a 3-stage MapReduce approach for self-join among records. However, computing the similarity score is not clearly showed. Besides, the work presented in [10] employs MapReduce to compute pairwise similarity scores. The goal is to accumulate the inner product of term frequencies between a pair of documents through two MapReduce phases as follows: (1) building a standard inverted index where each term is associated with a list describing the document it belongs to and its corresponding term frequency; and (2) calculating and summing all of the individual

values of a pair to generate its final similarity score. The approach looks like using Cosine measure but without normalization and strategic filtering. Another approach with MapReduce comes from the work in [12]. The basic idea is to build a word frequency dictionary, and then inputs referenced to the dictionary are converted into vector texts. Finally, prefixes of each vector text are calculated and stored in a PLT inverted file, which has the form of <word, textid, length, threshold value>. When there is a query text search, the query text will be transformed into vector texts which have been later on processed for their prefixes. In the end, words in each prefix will be searched from the PLT inverted file to find the text pairs that satisfy the given similarity threshold. This approach consumes, however, lots of computations and large amounts of prefixes, which easily leads to slowing down the whole system due to massive datasets. In [1], the authors present a hybrid combination between forward and inverted indexing but a mapper-only scheme. They develop a partitioning method for static filtering which assure dissimilar pairs are at different partitions. The authors then use the circular assignment, together with a hybrid indexing, to assign tasks that compute the all-pair similarity between these partitions. Cosine measure is exploited in this work, but its normalization is assumed to be already done before further computing. The work in [8] also uses Cosine measure and a filtering strategy to eliminate terms based on an upper-bounding pivot and a threshold. Nevertheless, how to normalize the weights, again, is not mentioned.

3 Preliminaries

3.1 Concepts

A workset Ω consists of a set of N documents D_i , which is represented as $\Omega = \{D_1, D_2, D_3, \dots, D_n\}$, and each document D_i composes of a set of words as term_k , which is shown as $D_i = \{\text{term}_1, \text{term}_2, \text{term}_3, \dots, \text{term}_k\}$. In general, each document D_i has the probability to share its terms with others, and we define common terms as those contained in all the considering documents in the workset Ω . Meanwhile, each term_k has its own term frequency tf_{ik} , which is described as the number of times the term_k occurs in the document D_i . The inverse document frequency idf_{ik} shows how much popular a term_k of a document D_i is across all the documents. In addition, the sign $[,]$ indicates a list, the sign $[[,], [,]]$ demonstrates a list of lists, and the sign $[,]_{\text{ord}}$ denotes an ordered list.

In this paper, we utilize the Cosine measure, which is popular and employed by the work in [1][4][10][16], to compute the similarity between a pair of documents D_i and D_j , whose formulae are defined as follows:

$$\text{sim}(D_i, D_j) = \sum_{k=1}^t W_{ik} * W_{jk} \quad (1)$$

$$\text{Where } W_{ik} = \frac{tf_{ik} * \log \frac{N}{n_k}}{\sqrt{\sum_{k=1}^t [(tf_{ik})^2 * (\log \frac{N}{n_k})^2]}} \quad (2)$$

From the equations (1) and (2), n_k represents the total number of documents sharing the same term $_k$, and idf_{ik} is computed as $\log(N/n_k)$. All of the documents, however, have to be normalized before being further processed. We call W_{ik} the normalized weight of term $_k$ in the document D_i , which is done by the equation (2). The purpose of normalization integration is to avoid the much affection of large documents to small ones and make the similarity scores fall into the interval $[0, 1]$, which is easily visualized to humans. Besides, bringing the normalization into the processing makes sense in reality and not an assumption in the context of big data because of its computation costs. Last but not least, we also exploit an inverted index, which maps a term $_k$ to the document D_i to which it originally belongs, to speed up the processing then.

3.2 MapReduce Paradigm

MapReduce is a parallel programming paradigm which aims at many large-scale computing problems [9]. The basic idea is to divide a large problem into independent sub-problems which are then tackled in parallel by two operations known as Map and Reduce. Its mechanism is deployed in commodity machines in that one is in charge of a master node and the others are responsible for worker nodes. The master delivers m Map jobs and r Reduce jobs to workers. Those which are assigned Map jobs are called mappers whilst those which are assigned Reduce jobs are called reducers. In addition, Map jobs are specified by a Map function and Reduce jobs are defined by a Reduce function. The single flow of MapReduce can be shortly described as follows: (1) The input is partitioned in a distributed file system (e.g., Hadoop Distributed File System – HDFS) [3], which produces a key-value pairs of the form $[\text{key}_1, \text{value}_1]$; (2) Mappers execute the Map function to generate intermediate key-value pairs of the form $[\text{key}_2, \text{value}_2]$; (3) The shuffling process groups these pairs into $[\text{key}_2, [\text{value}_2]]$ according to the keys; (4) Reducers execute the Reduce function to output the result; and (5) The result is finally written back into the distributed file system.

4 The Proposed Scheme

In this section, we propose an overview scheme that derives similarity scores between pairs of documents with MapReduce. From a general point of view and for simplicity, we firstly show the scheme as in the traditional self-join case without any query parameters in that we want to find pairwise similarity. Other specific cases following the scheme are presented in section 5 of the paper. As illustrated in Fig. 1, the whole process consists of four MapReduce phases as follows: (1) Building the customized inverted index; (2) Normalizing candidate pairs; (3) Building the normalized inverted index; and (4) Computing similarity pairs. Moreover, each phase is equipped with filtering strategies in order to eliminate dissimilar pairs and reduce overheads including storage, communication, and computing costs.

At the first MapReduce phase, sets of documents known as worksets are inputs to build the customized inverted index. Prior Filter is applied to discard common words. The reason is that they contribute nothing to the final similarity score but give a burden to

the whole process. Next, the customized inverted index will be normalized at the second MapReduce phase. In parallel, Query Term Filtering and Lonely Term Filtering are applied to filter those which only exist in a single document or those which are not in the given query document, respectively. In addition, the key-value pairs are descendingly ranked, which is according to their values. These key-value pairs are then fed to the third MapReduce phase so that the normalized inverted index is generated. Besides, Pre-pruning-1 will be done to reduce the candidate size when given a query document. Finally, the normalized inverted index is employed to filter candidate pairs again according to specific similarity queries like range or k-Nearest Neighbor (k-NN) queries before outputting similarity pairs. Last but not least, Pre-pruning-2 will be utilized to reduce candidate size at the Map task of this phase. More details of each phase are given in section 5 of the paper, which depends on specific similarity search cases.

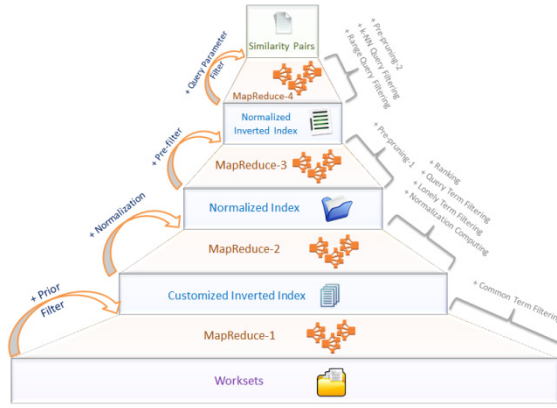


Fig. 1. The overview scheme

In general, let D_i be the i^{th} document of the workset, $term_k$ be the k^{th} word of the whole workset, tf_{ik} be the term frequency of the $term_k$ in the document D_i , idf_{ik} be the inverse document frequency of the $term_k$ in the document D_i , W_{ik} be the normalized weight of the $term_k$ in the document D_i , M_i be the total weight of all the terms in the document D_i , W_i be the largest weight of the $term_k$ in the document D_i , and $sim(D_i, D_j)$ be the similarity score between a document pair. A special character, e.g., @, is employed to semantically separate the sub-values in the values of a pair. The overall MapReduce operations can be summarized as follows:

MAP-1:	$[workset]$	$\rightarrow [term_k, D_i]$
REDUCE-1:	$[term_k, D_i]$	$\rightarrow [term_k, [D_i@tf_{ik}@idf_{ik}]]$
MAP-2:	$[term_k, [D_i@tf_{ik}@idf_{ik}]]$	$\rightarrow [D_i, term_k@tf_{ik}@idf_{ik}]$
REDUCE-2:	$[D_i, term_k@tf_{ik}@idf_{ik}]$	$\rightarrow [D_i, term_k@W_{ik}]_{ord}$
MAP-3:	$[D_i, term_k@W_{ik}]_{ord}$	$\rightarrow [term_k, D_i@M_i@W_i@W_{ik}]$
REDUCE-3:	$[term_k, [D_i@M_i@W_i@W_{ik}]]$	$\rightarrow [term_k, [D_i@M_i@W_i@W_{ik}]_{ord}]$
MAP-4:	$[term_k, [D_i@M_i@W_i@W_{ik}]_{ord}]$	$\rightarrow [D_{ij}, inner - product]$
REDUCE-4:	$[D_{ij}, inner - product]$	$\rightarrow [D_{ij}, sim(D_i, D_j)]$

5 Similarity Search Cases

The proposed scheme is applicable not only to popular similarity searches like pairwise similarity and search by example but also to those with query strategies such as range search and k-NN search. In each sub section below, we show in detail how it gets insight on the specific similarity searches.

5.1 Pairwise Similarity

Pairwise similarity search is the case in that we want to find out all possible similar pairs. In other words, one is bound to every other to give their similarity. Following the scheme, worksets are initially passed to mappers at Map-1 method, which produces intermediate key-value pairs of the form $[\text{term}_k, D_i]$. They are then retrieved by reducers at Reducer-1 method to output the key-value pairs of the form $[\text{term}_k, [D_i@tf_{ik}@idf_{ik}]]$, where tf_{ik} and idf_{ik} are derived. At this step, common words which have idf_{ik} equal to 0 are discarded by the Prior Filter. For example, assuming that there are three documents named $D_1, D_2,$ and $D_3,$ and each document contains its corresponding words as the input illustrated in Fig. 2. After Map-1 method, we have a list of intermediate key-value pairs $[[A, D_1], [B, D_1], [B, D_1], [C, D_1], [A, D_1], [E, D_1], [C, D_2], [A, D_2], [D, D_3], [B, D_3], [A, D_3], [E, D_3]]$. The list is then accessed by reducers at Reduce-1 method. The common word A is ignored by the Common Term Filtering while the lonely word D is marked as Terms Not Proceeded- $\{TNP\}$. The reason why the lonely word D is not discarded right away but marked as a special sign at this phase is that it should be kept joining the normalization step later on even though it does not contribute to any similarity scores in the end. Therefore, we have the output list as follows $[[B, [D_1@2@0.176, D_3@1@0.176]], [C, [D_1@1@0.176, D_2@1@0.176]], [\{TNP\}, D_3@1@0.477]], [E, [D_1@1@0.176, D_3@1@0.176]]]$.

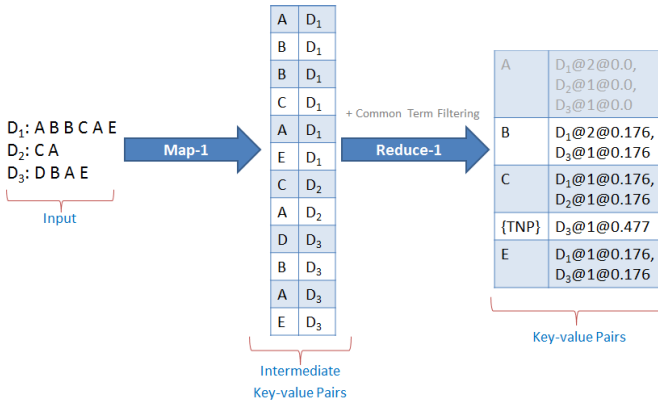


Fig. 2. MapReduce-1 operation

Next, the key-value pairs from the first MapReduce are normalized at the second MapReduce. The intermediate key-value pairs after Map-2 method have the form of

$[D_i, \text{term}_k @ \text{tf}_{ik} @ \text{idf}_{ik}]$. The Reduce-2 method normalizes these pairs into an ordered list of the form $[D_i, [\text{term}_k @ W_{ik}]]$. The values are sorted by their sizes and then by their W_{ik} . Fig. 3 shows the ongoing example at the second MapReduce. The mappers at Map-2 method output the intermediate key-value pairs as the list $[[D_1, B @ 2 @ 0.176], [D_3, B @ 1 @ 0.176], [D_1, C @ 1 @ 0.176], [D_2, C @ 1 @ 0.176], [D_3, \{TNP\} @ 1 @ 0.477], [D_1, E @ 1 @ 0.176], [D_3, E @ 1 @ 0.176]]$. These pairs are later normalized by the reducers at Reduce-2 method which gives us the normalized and ordered output list as following $[[D_1, [B @ 0.8165, C @ 0.4082, E @ 0.4082]], [D_3, [B @ 0.3271, E @ 0.3271]], [D_2, [C @ 0.1760]]]$. It is worth noting that the lonely term $\{TNP\}$ is filtered by Lonely Term Filtering at Reduce-2 method.

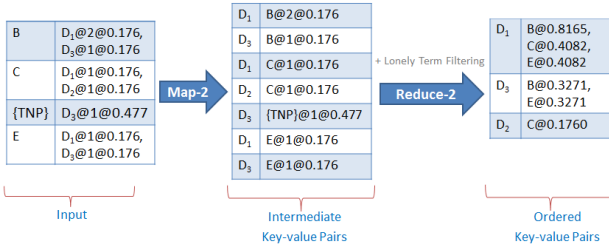


Fig. 3. MapReduce-2 operation

After the normalization, the third MapReduce takes the normalized inverted index into account. The mappers at Map-3 method emit the intermediate key-value pairs of the form $[\text{term}_k, D_i @ M_i @ W_i @ W_{ik}]$. The reducers at Reduce-3 method output the ordered key-value pairs of the form $[\text{term}_k, [D_i @ M_i @ W_i @ W_{ik}]]$. Fig. 4 presents the ongoing example at this phase. We have the list after Map-3 method as follows:

$[[B, D_3 @ 0.6542 @ 0.3271 @ 0.3271], [E, D_3 @ 0.6542 @ 0.3271 @ 0.3271],$
 $[B, D_1 @ 1.6329 @ 0.8165 @ 0.8165], [C, D_1 @ 1.6329 @ 0.8165 @ 0.4082],$
 $[E, D_1 @ 1.6329 @ 0.8165 @ 0.4082], [C, D_2 @ 0.1760 @ 0.1760 @ 0.1760]]]$

And we have the list after Reduce-3 method as follows:

$[[B, [D_1 @ 1.6329 @ 0.8165 @ 0.8165, D_3 @ 0.6542 @ 0.3271 @ 0.3271]],$
 $[E, [D_1 @ 1.6329 @ 0.8165 @ 0.4082, D_3 @ 0.6542 @ 0.3271 @ 0.3271]],$
 $[C, [D_1 @ 1.6329 @ 0.8165 @ 0.4082, D_2 @ 0.1760 @ 0.1760 @ 0.1760]]]$.

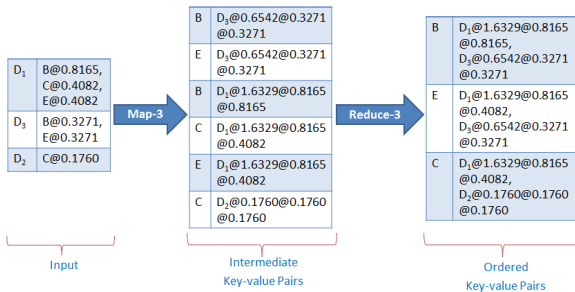


Fig. 4. MapReduce-3 operation

Finally, the fourth MapReduce computes the partial product of each corresponding term of a pair, which has the form $[D_{ij}, \text{inner-product}]$, at Map-4 method and leads to the final similarity score of each pair, which has the form $[D_{ij}, \text{sim}(D_i, D_j)]$. The running example is closed at this phase from Fig. 5. The intermediate key-value pairs $[[D_{13}, 0.2881], [D_{12}, 0.0718], [D_{13}, 0.1440]]$ after Map-4 method are aggregated to the final similarity scores $[[D_{13}, 0.4321], [D_{12}, 0.0718]]$ at Reduce-4 method. Last but not least, Query Parameter Filtering is optionally applied to obtain closer results when query parameters are given.

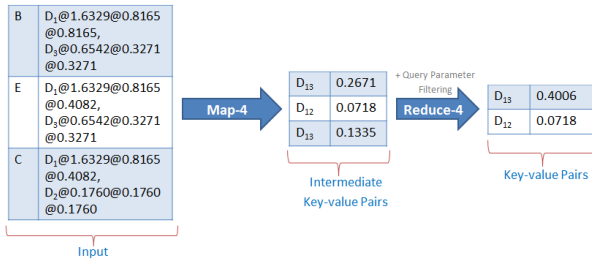


Fig. 5. MapReduce-4 operation

5.2 Search by Example

Search by example is a well-known similarity search case when given a pivot object as an example for the search. The goal is to find the most similar objects according to the pivot. Once it is the case, not only are lonely words in the pivot discarded but also those which do not exist in the pivot are ignored by Lonely Term Filtering and Query Word Filtering at Reduce-2 method. The reason is that they do not contribute to the similarity between a pair but make the process bulky. Doing so significantly contributes to the reduction of overheads such as storage, communication, and computing costs through the whole process of MapReduce jobs.

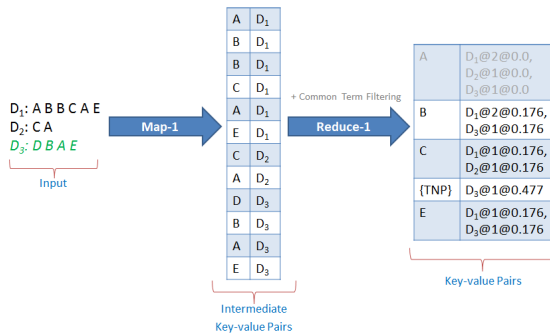


Fig. 6. MapReduce-1 operation when given the pivot

Let us come to the example as illustrated in Fig. 6, and at this time, the document D₃ is considered as the pivot. The MapReduce operations conform to the proposed

scheme. The difference here is that term C in D_1 , term C in D_2 , and {TNP} are discarded in advance. Furthermore, search by example can be leveraged by query strategies presented in section 5.4, which shows how soon candidate pairs are filtered to reduce the candidate size and fit the query.

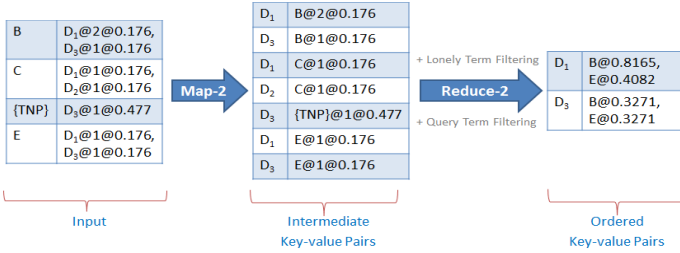


Fig. 7. MapReduce-2 operation when given the pivot

5.3 Query Strategies

Most similarity searches are also accompanied with search query strategies such as range search or k-NN search. The range search adds the similarity threshold ϵ so that those pairs whose similarity is greater or equal to the threshold should be returned as the final result. Meanwhile, the k-NN search looks for the k most similar objects from the candidate sets. As a consequence, the parameters ϵ and k are utilized to filter objects so that the final result, on the one hand, is as close as users' needs and the search process, on the other hand, is significantly improved. In order to exploit them for the proposed scheme, both Pre-pruning-1, for the case a query document is given, and Pre-pruning-2, for other cases, are attached but not mutually exclusive.

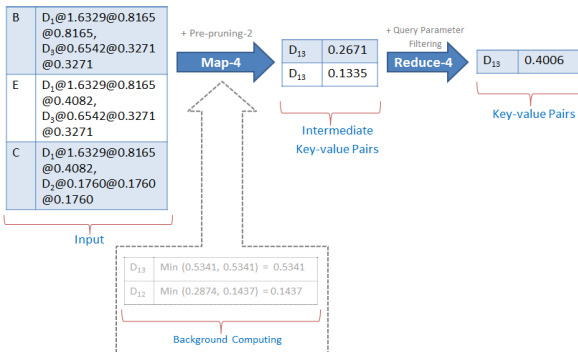


Fig. 8. MapReduce-4 operation with Pre-pruning-2

In the case of pairwise similarity, we do not actually want to find all-pair similarity due to the fact that it is rarely used in a specific range of applications whereas its entire result is not completely utilized. Moreover, such a big process consumes much time and resources, which is not really suitable for most application scenarios, especially for real-time intensive ones. Thus, the threshold ϵ is provided to filter necessary

pairs from the total candidates to meet certain needs. Pre-pruning-2 at Map-4 method catches this line of thought. It employs the two below inequalities with the latter adopted in [1] to do its task as candidate filtering:

$$\text{sim}(D_i, D_j) = \sum_{k=1}^t W_{ik} * W_{jk} \geq \varepsilon \quad (3)$$

$$\text{sim}(D_i, D_j) \leq \min(M_i * W_j, M_j * W_i) = \sigma \quad (4)$$

From the inequalities (3) and (4), the filtering rule is to find those whose σ is greater or equal to the threshold ε . Let us back to the example of pairwise similarity in section 5.1. At Map-4 method as illustrated in Fig. 8, the pair D_1 and D_3 has their σ as 0.5341 whilst the pair D_1 and D_2 has their σ as 0.1437. Assuming that the threshold ε has the value 0.4, the pair D_1 and D_2 is early discarded. Meanwhile, Pre-pruning-1 is able to sooner get rid of unnecessary pairs when given a query object, and this supporting process takes place at Reduce-3 method. It is worth noting that the key-value pairs at this phase have the form [term_k, D_i@M_i@W_i@W_{ik}], so the above filtering rule can be shortly derived. From the instance of search by example in section 5.2, the Pre-pruning-1 method indicated in Fig. 9 estimates candidate pairs whether σ is greater or equal to ε . The value of σ is computed as 0.4006, which is the minimum between 0.4006 and 0.5342. Assuming the threshold ε has the value 0.4, the pair D_1 and D_3 is, therefore, further processed to get their final similarity.

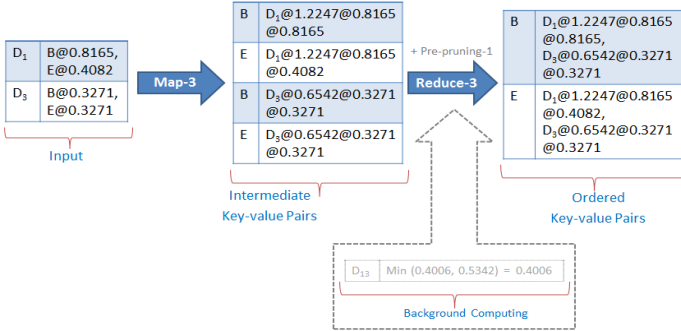


Fig. 9. MapReduce-3 operation with Pre-pruning-1

On the other hand, k-NN query is also attached together with a query object. Pre-pruning-1 takes the k parameter into account to filter objects before their similarity is computed. In other words, each mapper at Map-3 method approximately emits top-k key-value pairs whose size is according to the total number of running mappers as the equation (5) below:

$$\text{top - k pairs for each mapper} = \max_{k \in \mathbb{N}} \left(\left\lfloor \frac{k}{\sum \text{Mappers}} \right\rfloor, 1 \right) \quad (5)$$

It is totally possible because the key-value pair input of Map-3 method has been ordered by its size and normalized weights from the second MapReduce operation.

Moreover, the probability a pair is the most similar is high when each combined object has its largest size and normalized weights. As a consequence, the equation (5) helps reduce unnecessary computing and the candidate size.

6 Experiments

6.1 Environment Settings

In order to do our experiments with MapReduce, we employ the stable version 1.2.1 of Hadoop and DBLP dataset [7], which is used to do similarity search on the title of publications. The Hadoop framework is installed in the cluster of commodity machines called Alex, which has 48 nodes and 8 CPU cores and either 96 or 48 GB RAM for each node [2]. In general, we leave Hadoop configurations in default mode as much as possible, for we want to keep the most initial settings which a commodity machine may get even though some parameters could be tuned or optimized to fit the Alex cluster. The configured capacity is set to 5GB per node, so the 48-node cluster totally has 240GB. The number of reducers for a reduce operation is set to 168. In addition, the replication factor is set to 47. The possible heap size of the cluster is about 629 MB, and each HDFS file has 64MB Block Size. It is worth noting that Alex has suffered the overhead of other coordinating parallel tasks, i.e., these nodes are not exclusively for the experiments. In addition, they are diskless nodes, but in the background data are located on a storage area network. Last but not least, each benchmark has its fresh running. In other words, data from the old benchmark are removed before the new benchmark starts. All the experiments for one type of query are consecutively run so that their environments are close as much as possible.

6.2 Empirical Evaluations

In this section, we perform some performance measurements for examining methods. The measuring time is bound since the time MapReduce jobs start running to the time they finish writing the result to the HDFS. First, Fig. 10 shows Pairwise similarity case among the naïve approach (i.e., the approach without applying any filtering), the filtering approach, and the work in 2008 [10] and search by example. Apart from the work in 2008, the other approaches are based on our proposed scheme in section 4. Besides, we also compare the search by example case with the pairwise similarity case. The dataset size is increased turn by turn from 50MB to 500MB. From Fig. 10a, the result shows that our proposed methods outperform the work in 2008 in terms of query processing time. More concretely on the average, the naïve approach is 68.38% faster than the work in 2008, the filtering approach is 69.41% faster than the work in 2008, and the search by example is 73.03% faster than the work in 2008. The main reason is that the work in 2008 finds the term frequency right away at mappers instead of reducers whose main goal is to perform reduced computations. In other words, the functionality of mappers is mistakenly used from the beginning. Moreover, the work in 2008 computes all possible candidates without filtering whilst our approach does. On the other hand, there is no big difference among the naïve approach, the filtering approach, and search by example while the dataset size is still small, or to say, under a

specific threshold. The reason is due to the operation cost of the whole system. Once the dataset size is significantly increased, a big gap among them emerges. On the average, the naïve approach consumes 3.5% more CPU time than the filtering approach and 15.1% more CPU time than search by example.

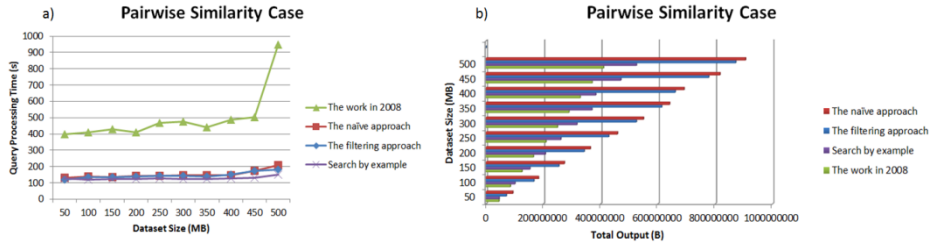


Fig. 10. Pairwise similarity between the naïve approach, the filtering approach, and the work in 2008 and search by example; (a) Query processing time; and (b) The saved data volume

In terms of data volumes, Fig. 10b shows the correlation of data quantity among the approaches during MapReduce operations. The work in 2008 has fewer amounts of data output in the end. More specifically, the work in 2008 produces 54.01% less data than the naïve approach, 50.01% less data than the filtering approach, and 17.68% less data than search by example, respectively on the average. The reason is that the proposed scheme, on the one hand, needs to normalize inputs before computing the similarity whilst no filtering is accompanied. On the other hand, it is worth noticing that the work in 2008 computes the similarity score between two documents by summing the inner products of the term frequencies, which are not normalized yet. Normalization is essential because weight terms should be high if they are frequent in relevant documents but infrequent in the collection as a whole. If normalization is taken into account, the work in 2008 suffers more computations and data volumes. Nevertheless, we implement it as the original version, i.e., without normalization. Furthermore, the result indicates how much important the refinements are applied in order for the filtering approach to save 4% data quantity and for search by example to save 32.33% data quantity, on the average, when compared to the naïve approach. Last but not least, the amount of data output from MapReduce-2 operation to MapReduce-4 operation, when filtering is applied, just gets 0.04% data proportion on the average compared to the whole data output in the case of search by example itself. As a consequence, search by example has 43.97% less data than the naïve approach. In summary, the data output volume without filtering is nearly double in comparison with the data input size due to normalization. In addition, MapReduce mechanism always writes down intermediate outputs into HDFS, whose disk access costs are too expensive. Filtering strategies are, therefore, essential to reduce the candidate size and related computing costs as well.

On the other side, we conduct experiments with query strategies when the dataset size is step-by-step increasing from 300MB to 700MB, which are shown in Fig. 11. The data values from Fig. 11a indicate that there is no big difference in terms of query processing among range queries where the similarity thresholds are set to 90%, 70%, and 50%. Likewise, the values from Fig. 11b point out the same evaluation for k-NN

queries where the values of parameter k are set to 100, 300, and 500, respectively. Moreover, the two kinds of query strategies mostly have the same performance. In other words, either the parameter ϵ for range queries or the parameter k for k -NN queries does not give a gap between them. Last but not least, the two kinds of query strategies perform 2.67% to 4% faster than search by example without Pre-pruning.

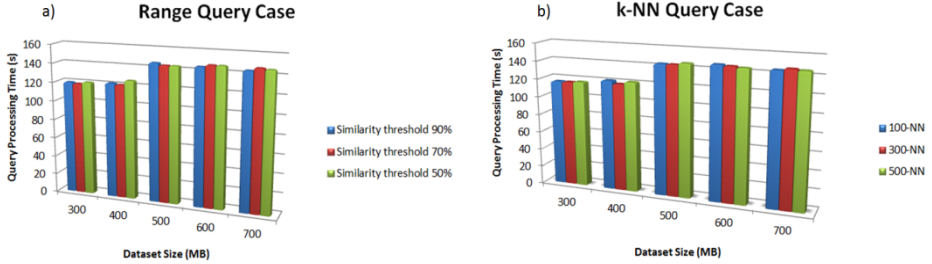


Fig. 11. Query strategies; (a) Range query case; and (b) k -NN query case

To additionally wrap up by these experiments, we further describe the important factors that most matter to achieving high performance with MapReduce as followings: (1) The MapReduce operations should not be too complex due to limited computing resources; (2) The less computations the similarity measure is, the high efficiency the whole system gets [13]; (3) The ways of programming for Map and Reduce functions also affect the entire system; and (4) Depending on the characteristics of commodity machines, the environment settings can be further optimized to improve the overall performance.

7 Conclusion and Future Work

Aiming at scalability, we propose an efficient similarity search in large data collections with MapReduce, which is flexibly adaptable to popular similarity search cases such as pairwise similarity, search by example, range query, and k -NN query. In addition, collaborative strategic refinements associated with the proposed scheme not only foster the prospective scalability of MapReduce but also eliminate unnecessary computations as well as diminish candidate sizes. Furthermore, our methods are verified by experiments on real massive datasets and Hadoop framework, which is deployed and experienced in the commodity machines. For the future work, we will enhance the model with distinct n -grams instead of words in order to minimize the intermediate data outputs. Besides, our proposed approach is possibly extended to the case of incremental similarity search, which gives best support to both offline and high-frequency incoming data processing. Last but not least, we will consider and take care of other challenges under the context of big data in order to strengthen our methods supporting data-intensive applications.

Acknowledgements. We would like to give our thanks to Mr. Faruk Kujundžić, Information Management team, Johannes Kepler University Linz, for kindly supporting us in Alex Cluster.

References

1. Alabduljalil, M.A., Tang, X., Yang, T.: Optimizing Parallel Algorithms for All Pairs Similarity Search. In: Proceedings of the 6th ACM International Conference on Web Search and Data Mining, USA, pp. 203–212 (2013)
2. Alex cluster, <http://www.jku.at/content/e213/e174/e167/e186534> (referenced on February 4, 2014)
3. Apache Software Foundation. Hadoop: A Framework for Running Applications on Large Clusters Built of Commodity Hardware (2006)
4. Baraglia, R., De Francisci Morales, G., Lucchese, C.: Document Similarity Self-Join with MapReduce. In: Proceedings of the 10th IEEE International Conference on Data Mining, pp. 731–736 (2010)
5. Dang, T.K.: Solving Approximate Similarity Queries. *International Journal of Computer Systems Science and Engineering* 22(1-2), 71–89 (2007)
6. Dang, T.K., Küng, J.: The SH-tree: A Super Hybrid Index Structure for Multidimensional Data. In: Mayr, H.C., Lazanský, J., Quirchmayr, G., Vogel, P. (eds.) DEXA 2001. LNCS, vol. 2113, pp. 340–349. Springer, Heidelberg (2001)
7. DBLP data set, <http://dblp.uni-trier.de/xml/> (referenced on March 8, 2014)
8. De Francisci Morales, G., Lucchese, C., Baraglia, R.: Scaling Out All Pairs Similarity Search with MapReduce. In: Proceedings of the 8th Workshop on Large-Scale Distributed Systems for Information Retrieval, pp. 25–30 (2010)
9. Dean, J., Ghemawat, S.: MapReduce: Simplified Data Processing on Large Clusters. In: Proceedings of the 6th Symposium on Operating Systems Design and Implementation, pp. 137–150. USENIX Association (2004)
10. Elsayed, T., Lin, J., Oard, D.W.: Pairwise Document Similarity in Large Collections with MapReduce. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies, Companion Volume, Columbus, Ohio, pp. 265–268 (2008)
11. Fenz, D., Lange, D., Rheinländer, A., Naumann, F., Leser, U.: Efficient Similarity Search in Very Large String Sets. In: Ailamaki, A., Bowers, S. (eds.) SSDBM 2012. LNCS, vol. 7338, pp. 262–279. Springer, Heidelberg (2012)
12. Li, R., Ju, L., Peng, Z., Yu, Z., Wang, C.: Batch Text Similarity Search with MapReduce. In: Du, X., Fan, W., Wang, J., Peng, Z., Sharaf, M.A. (eds.) APWeb 2011. LNCS, vol. 6612, pp. 412–423. Springer, Heidelberg (2011)
13. Phan, T.N., Küng, J., Dang, T.K.: An Elastic Approximate Similarity Search in Very Large Datasets with Mapreduce. In: Hameurlain, A., Dang, T.K., Morvan, F. (eds.) Globe 2014. LNCS, vol. 8648, pp. 49–60. Springer, Heidelberg (2014)
14. Szmit, R.: Locality Sensitive Hashing for Similarity Search Using MapReduce on Large Scale Data. In: Kłopotek, M.A., Koronacki, J., Marciniak, M., Mykowiecka, A., Wierchoń, S.T. (eds.) IIS 2013. LNCS, vol. 7912, pp. 171–178. Springer, Heidelberg (2013)
15. Vernica, R., Carey, M.J., Li, C.: Efficient Parallel Set-similarity Joins Using MapReduce. In: Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data, USA, pp. 495–506 (2010)
16. Xiao, C., Wang, W., Lin, X., Yu, J.X.: Efficient Similarity Joins for Near Duplicate Detection. In: Proceedings of the 17th Int'l World Wide Web Conference, pp. 131–140 (2008)
17. Zhang, D., Yang, G., Hu, Y., Jin, Z., Cai, D., He, X.: A Unified Approximate Nearest Neighbor Search Scheme by Combining Data Structure and Hashing. In: Proceedings of the 23rd International Joint Conference on Artificial Intelligence, pp. 681–687 (2013)

Memory-Based Multi-pattern Signature Scanning for ClamAV Antivirus

Nguyen Kim Dien, Tran Trung Hieu, and Tran Ngoc Think

Faculty of Computer Science and Engineering
HCMC University of Technology
Ho Chi Minh City, Vietnam
diennguyenkim@gmail.com,
{hieutt, tnthink}@cse.hcmut.edu.vn

Abstract. Signature scanning plays an important role on modern security application such as virus scanners, intrusion detection/prevention systems, and firewalls. High demand of scanning throughput gives rise to recent efforts on hardware-based matching engine. In this paper, we proposed an efficient architecture for matching Clam Antivirus (ClamAV) signatures on reconfigurable platform (FPGA). We utilize Bloom filter technique for filtering input data and Bloomier filter technique for one round check suspect data. Our proposed approach supports up to 256 byte length signatures and can handle both basic and regular expression signatures. Our prototype on NetFPGA platform could handle up to 16K regular expression signatures and 64K basic signatures.

Keywords: Antivirus, bloom filter, fpga, pattern matching.

1 Introduction

Nowadays, when the use of internet is increasing rapidly and globally, information security has become more and more critical. One of the most important solutions for tightening system security is antivirus. Antivirus application exploits pattern matching as a main mechanism for detecting illegal programs such as computer viruses, malwares, worms... Although there are many improvements in pattern matching algorithms, this process still occupies a significant amount of resources and slowdown system performance. In addition, due to the growing number and the complexity of virus signatures, it has become a bottle-neck task in software-based antivirus program.

The limitation in software-based antivirus program has led to a high demand of hardware solution to speed up this process. One of the most common hardware technologies in this field is Field Programmable Gate Array (FPGA) because of its capability of parallelism and flexibility in changing application. There are various FPGA-based approaches have been proposed, but those systems [1, 2, 3, 4] need quite large on-chip memory or consume lots of logic elements [5]. Moreover, the problem of regular expression (regex) signatures has not been resolve completely. In this paper, we proposed a high-speed FPGA architecture for scanning Clam Antivirus (ClamAV) signatures. We exploit Bloom Filter and Bloomier Filter (BBF) techniques that result in low resource consumption and proposed approaches for handling regex

signatures. Our approach is mainly based on memory, so that the system update could be simple and fast.

The remaining of this paper is organized as follows. In section 2, we introduce ClamAV database and discuss related works including our previous work on ClamAV static pattern matching engine. Section 3 discuss our signature preprocessing process. This is an import part which helps clarify the design of your proposed architecture. Section 4 describes hardware architecture of our proposed virus matching engine. The experimental results of our system are presented in section 5. Finally, conclusions and future work are given in Section 6.

2 Background and Related Works

2.1 ClamAV Database

Clam Antivirus (ClamAV) is an open-source cross-platform antivirus software that is mainly used on server-side for email virus scanning. ClamAV database has been used as experimental signature set in many hardware-based virus scanning solutions [6, 7]. As of September 2013, the entire ClamAV database contains over 2,400,000 virus signatures. The *main* database (released with the software) and the *daily* database (regularly updated) are the two parts of ClamAV database.

Virus signatures in ClamAV, in Table 1, can be divided to three main types: MD5 checksums, basic patterns, and regular expression (regex) patterns. In this paper, we only use the *main* database from the lasted stable version 0.98.4 (Sep, 2013). As depicted in the Table 1, the MD5 signatures take the largest number of virus signatures (93.96%). Basic and regex signatures are next with 3.67% and 0.40% respectively. Despite of largest proportion of signatures, MD5 signatures only account for around 8% of scanning time. Therefore, we ignore this kind of signature and focus on basic and regex types.

Table 1. Analysis on ClamAV database (Sep 2013)

	Basic	Regex	MD5	Other	Total
No. Signatures	88885	9670	2277804	47866	2424225
No. Sig. ratio (%)	3.67%	0.40%	93.96%	1.97%	100%
Scan time ratio (%)	63.86%	27.94%	8.2%		100%

The basic signature is a continuous byte string. A regex signature is an extension of the basic signature with various wildcards. Table 2 summarizes the most used wildcards in ClamAV. For detecting polymorphic viruses, regex signature support is necessary. Example of two kind signatures is given in Fig. 1. The signature for *Trivial-348* virus is defined in form of hex string, while the signature for *DOS.Lawine* virus is defined with some wildcards ($\{n-m\}, ??$). Each sign “??” can be replaced with any 8 bit data and sign “ $\{n-m\}$ ” can be replaced with within in range from n to m number of bytes. When searching for this regex signature, the matching engine must be able to match separate patterns “*8bd866b9b2080000bb*”, “*2e8037*”, “*f943fce2f7*”, and verify the distant constraints between these patterns.

```
Trivial-348 (Clam)=cd21b74093ba0001b11ecd21c32a2e2a00
DOS.Lawine (Clam)=8bd866b9b2080000bb{1-10}2e8037??f943fce2f7
```

Fig. 1. Example of virus signature

Table 2. Common wildcard in regular expression signature

Wildcard	Distance constraints
??	1 byte
{n}	n bytes
{-n}	less than n bytes
{n-}	more than n bytes
{n-m}	more than n bytes but less than m bytes
*	any number of byte

2.2 Related Works

In pattern matching, three most known approaches in FPGA-based solutions are shift-and-compare [5,10], state machine [11,17] and hashing method [9,12,16]. Systems which are implemented by using shift-and-compare or state machine method work intensively on logic cells rather than on-chip memory. As a result, the number of logic cells utilized by these systems depends on the number of patterns as well as length of patterns and the process of updating patterns requires reconfiguration of entire system. On the contrary, hashing approach employs on-chip memory to store patterns. Consequently, the resource consumption of logic cells of hash-based systems relies partially on patterns set and the process of modifying patterns set in these systems is simply changing their memory content. Our system based on two hashing technique: Bloom and Bloomier filter.

A Bloom filter [9] is a compact probabilistic data structure which is used an index table to determine if an element might be part in pre-defined set. The base data structure of a Bloom filter is a bit vector. Initially, all entries in the bit-vector index table are set to 0. To add an element to the index table, we use k hash functions to hash element to k position in the table, and set the bits in the bit vector at the index of those hashes to 1. This process is repeated until all members of pre-defined set are hashed to index table.

To test for membership, we simply hash the element with the same k hash functions, then see if those values are set in the bit vector table. If one of these entries is 0, this string is not member of the set, otherwise, the existence of this string in the set is uncertain and further test is required. This uncertainty is caused by "false positive" problem in hash-based system. "False positive" probability is calculated by equation, with m is length of bit vector table, n is size of set and k is the number of hash functions.

$$f = (1 - e^{-\frac{nk}{m}})^k \tag{1}$$

Due to "false positive" problem, this hash method is not efficient with applications that require accurate searching. In order to archive an acceptable probability, a large

amount of memory must be used. But Bloom filter can creating index table process very quick and index tables can be created directly by a module on FPGA.

In Bloom Filter's worst case, the entire pre-defined set is scanned to confirm the uncertain match result from hashing operation. [13] introduces better approach by using secondary hash function and index table to reduce the scope of matching but it still has to scan more than one pattern. This task consumes lots of time because the patterns are usually stored in low speed off-chip memory.

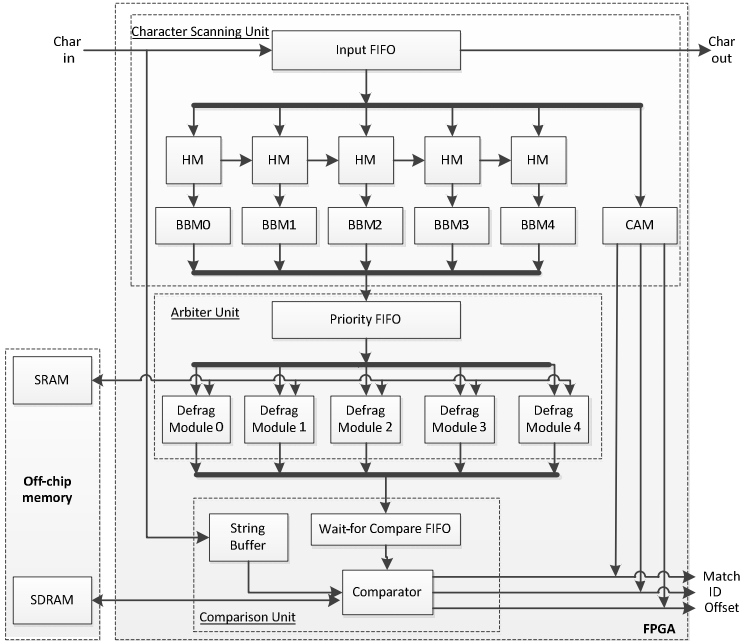


Fig. 2. Bloom-Bloomier Filter Pattern Matching Engine architecture

Bloomier Filter [14] is developed to solve this weakness. It can show exactly which pattern in the set is the best match with the searched string so the query time is constant. Bloomier Filter's algorithm is similar to Bloom Filter, but its index table is constructed in a different method. Instead of using one bit for each index table entry, Bloomier Filter stores more information in one entry, as a result, size of each entry depends on which information is encoded. Because of this extra information, Bloomier's index table is built in a more complex way as compared with Bloom Filter [15]. The advantage of this hashing method is just compare with a pattern.

2.3 Bloom-Bloomier Filter Pattern Matching System

Our previous work [8], BBF-Engine is a pattern matching engine which bases on the combination of Bloom filter and Bloomier Filter. Besides providing a simple match/no-match answer, BBF-Engine also indicates which pattern resulted in a match, speeding up exact matching to verify against with low number of false positives.

BBF-Engine in Fig. 2 includes a Character Scanning Unit, an Arbiter Unit, a Comparison Unit and an off-chip memory to store original patterns. The Character Scanning Unit includes a series of Hash Modules to calculate hash value for each possible string of which length is between 1 and 128 characters. There are also 5 Bloom-Bloomier Modules (BBM) for the corresponding string lengths: 8, 16, 32, 64 and 128. The Arbiter Unit repeatedly fetches bloom-match records in Bloom-match FIFOs for processing. Based on these records, the Arbiter Unit looks through record's extra pre-processed information in SRAM to determine whether this is a single-fragment pattern or two-fragment pattern and the correlated original pattern's address in SDRAM. After defragmenting fragments, the Arbiter Unit put suspected pattern id into Wait-for-Compare FIFO for Comparison Unit. The Comparison Unit uses those suspected pattern id to compare suspected string stored in the String Buffer against original pattern in SDRAM and produces a match together with pattern ID if the suspected string and the original pattern are identical.

The remaining problem of BBF-Engine is that it only support basic signatures of ClamAV. In this paper, we extend our BBF-Engine and propose a comprehensive architecture which supports both basic and regex signatures.

3 Pattern Analyzer

3.1 Preprocessing of Virus Signature

In order to be deployed on FPGA, all ClamAV signatures need to be preprocessed. This process is repeated whenever new signatures are updated. Fig. 3 illustrates the workflow of our signature preprocessing process. First, we extract virus signatures from ClamAV database and classify them into two separate groups: basic signature group and regex signature group. In the second step, for each signature in regex group, we extract all wildcards and split it in to multiple sub-patterns. A sub-pattern is a continuous byte string and could be considered as a basic signature. We keep reference information with each sub-pattern in Pattern Metadata for later reconstruction of original signature. All extracted sub-patterns and basic signatures are then merged into a deploying signature set. In the third step, we split each signature into fragments in order to meet pre-defined pattern lengths of our BBF-Engine.

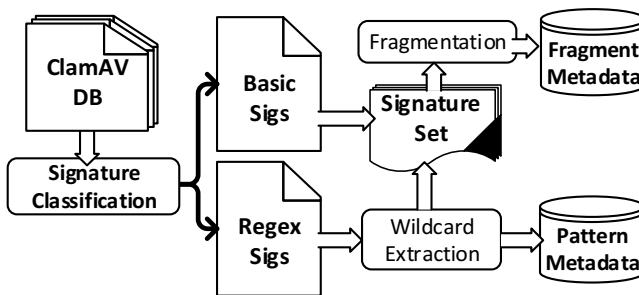


Fig. 3. The flow diagram of pre-processing virus signatures

3.2 Wildcard Extraction

Wildcard defines a gap or displacement of a sub-pattern to another sub-pattern in order. Therefore, we propose to match regex signature by matching each sub-pattern and verify its displacement constrains to the others. Our software will preprocess each regex signature to extract all sub-patterns and create metadata for it.

Example 1: Given pattern $P=abcdefghij\{10\}def$.

Pattern P contains the displacement operator $\{10\}$, which means a gap of 10 arbitrary bytes, and would be divided into two sub-patterns:

Sub-pattern 1: $P1=abcdefghij$

Sub-pattern 2: $P2=def$.

During the wildcard extraction process, we convert all wildcards to the same form of $\{m,n\}$. By this way, the algorithm for reassembly of sub-pattern is uniform and simpler. Table 3 show values of m and n for each conversion of wildcard. ∞ denotes the maximum length that our engine could support. The exact value of ∞ depends on deployed signature set that will be discussed in section 5.

Table 3. Wildcard conversion

Original	At-least(n)	Within(m)
??	length of wildcards	length of wildcards
*	0	0
{n}	n	n
{n-}	n	∞
{-m}	∞	m
{n-m}	n	m

3.3 Signature Fragmentation

Length of sub-patterns considerably varies to hundreds of characters, for that reason, we cannot implement Bloom-Bloomier Filter (BBF) for every sub-pattern length. This consumes too many resources and wastes memory because there are some groups of sub-patterns of the same length which only have a few sub-patterns. Therefore, we break sub-patterns into fragments, put fragments having the same length into distinct groups and apply BBFs for those groups.

To generalize the system, we do not analyze the sub-pattern database thoroughly to fragment sub-patterns, we implement fixed number of BBF for the fragment length of 8, 16, 32, 64 and 128 characters. The group of 8-character fragments should cover all sub-patterns which have length between 8 and 15, similarly, group of 16-character fragments covers 16-to-31 character sub-patterns, and so on. All of sub-pattern's fragments should be at the same length to simplify the process of reconstructing fragmented patterns. As a result, each sub-pattern only has at most 2 same-length, overlapped fragments, we would not have enough memory to store

hash tables, since the size of hash tables roughly depends on number of fragments[9]. The fragmentation produce is rather easy and fast.

Example 2: According example 1, sub-pattern $P1=abcdefghij$. The length of sub-pattern P1 is 10, hence we split this sub-pattern into two 8-character-fragments:

Fragment 0: $F0=abcdefgh$

Fragment 1: $F1=cdefghij$

The distance of 2 fragments is so then $10 - 8 = 2$. The distance means the number of characters must be counted starting from the time fragment 0 is reported match to the next match of fragment 1.

We follow three linked list method as [16] to reconstruct a pattern from its fragments. There are three pieces of information are encoded in first fragment record: Distance to second fragment, hash value of that second fragment and pattern id. When a first fragment is matched, our matching engine stores its information for comparison with the upcoming match of second fragment. If next fragment length, its on-arriving relative distance as well as hash value are all equivalent to any correspondence previous record, we can assume these two fragments belong to one pattern.

4 System Architecture

4.1 Overview

The overall architecture of our virus signature scanning system is illustrated as in Fig. 4. The system is divided into two main parts. The first part, Pattern Matching Module, is for matching single pattern and the second part, Pattern Reassembly Module, is for reassembling separate sub-patterns into a complete multi-pattern signature. As discussed in previous section, there are two types of virus signatures that are supported by our system: single pattern signature and multi-pattern signature. All patterns that are either single pattern signatures or sub-patterns of a regex signatures are handled by the Pattern Matching Module. Only patterns that are a part of regex signature are processed by Pattern Reassembly Module.

The processing flow of our proposed system is as follows. Scanning data that are read from file or network data buffer feed Pattern Matching Module (PMM) in form of byte stream. The PMM, as discuss in background section, buffers scanning bytes and looks for the appearance of virus signature pattern. After the successful matching of a pattern, PMM send an event to Arbiter Module. An event include a pattern identify number (PID) and a byte index (BI). The PID is a unique number for identifying a pattern in our system and the BI is the matching position of pattern in scanning byte stream. If the matched pattern is a single pattern signature, the Arbiter Module will trigger a virus matching alert and return a sig ID as the identification of found virus. Otherwise, if the matched pattern belongs to a regex signature, it will forward this event to Pattern Reassembly Module (PRM) for further processing. As a regex signature is formed by several sub-patterns. In order to conclude a successful match, all sub-pattern must be appeared in scanning data in a correct order and their matched positions must satisfy all wildcard condition. The role of PRM is verifying these constraints on matched sub-pattern events.

If all constraints of a regex signature are satisfied, the PRM will send an event to Arbitrer Module as indication of a found virus signature.

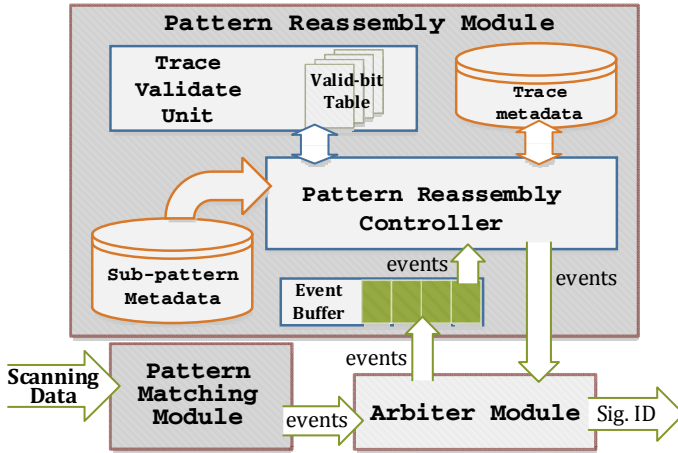


Fig. 4. Overview architecture of Wildcard Processing Module

Pattern Reassembly Controller (PRC) sequentially reads and processes event from Event Buffer. For each event, PRC looks into Sub-pattern Metadata for its metadata. The metadata, as discussed below, consists of information about the type of sub-pattern, the order of sub-pattern, and wildcard condition to the next sub-pattern. If the matched pattern is first sub-pattern of a regex signature, PRC will create a trace for this signature and keep track of this trace with the information stored in Valid-bit Table and Trace Metadata. The Valid-bit Table indicate the active trace and Trace Metadata keep information of the next sub-pattern of a signature. If a trace is fail to satisfy the constraint on a sub-pattern, the trace will be marked as invalid and must be started again. Following sections will discussed each component of PRM.

4.2 Sub-pattern Metadata and Trace Metadata

The metadata associates with each sub-pattern is showed in Table 4. *sig_id* entry is required by PRC for tracking a sub-pattern. PRC uses *sig_id* for checking valid trace in *Trace Validate Unit* (TVU). If the trace corresponding to *sig_id* is not valid and the sub-pattern is not the first one (order = 1), PRC can ignore this sub-pattern and continue to serve the next event from Event Buffer. *next_id* entry indicates the next sub-pattern that follows current pattern. The *m* and *n* entries store the number of bytes from current *byte_index* to the first and the last position for looking for the next sub-pattern. For each matched sub-pattern belonging to active trace, PRC will verify its metadata with the metadata of the active trace, as showed in Table 5, stored in *Trace Metadata*. If pattern id (*pid*) of current sub-pattern and its order in the signature equal to figures in trace metadata and its current matching position (*byte_index*) lies between *first_pos* and *last_post*, the current sub-pattern is the satisfy trace condition.

Its metadata will be later update to the Trace Metadata. The operation of PRC for tracking regex signature is described by pseudo code in Fig.5.

Table 4. Metadata of a sub-pattern

Metadata	Description
sig_id	Signature ID
next_id	ID for next sub-pattern
m	maximum distance from the previous pattern
n	minimum distance from the previous pattern
order	order of sub-pattern

Table 5. Metadata of a trace

Metadata	Description
next_id	ID for next sub-pattern
first_pos	first expected position of next sub-pattern
last_pos	last expected position of next sub-pattern
order	order of sub-pattern

```

Get metadata of sub-pattern pid from Sub-Pattern Metadata
if (Trace_Validate(pid.metadata.sig_id) == true):
    Get metadata of trace sig_id from Trace Metadata
    if(sig_id.metadata.next_id == pid
        and sig_id.metadata.order == pid.metadata.order
        and byte_index >= sig_id.metadata.first_pos
        and byte_index <= sig_id.metadata.last_pos):
        if(pid.metadata.next_ID == INVALID_PID):
            //clear trace
            Trace_Validate(pid.metadata.sig_id) = false
            Sent event to Arbiter Module
        else:
            //update trace
            sig_id.metadata.next_id = pid.metadata.next_id
            sig_id.metadata.order = pid.metadata.order + 1
            sig_id.metadata.first_pos = pid.metadata.m + byte_index
            sig_id.metadata.last_pos = pid.metadata.n + byte_index
    if(byte_index > sig_id.metadata.last_pos):
        //clear trace
        Trace_Validate(pid.metadata.sig_id) = false
else:
    if(pid.order == 0):
        //set trace valid
        Trace_Validate(pid.metadata.sig_id) = true
        //update trace metadata
        sig_id.metadata.next_id = pid.metadata.next_id
        sig_id.metadata.order = pid.metadata.order + 1
        sig_id.metadata.first_pos = pid.metadata.m + byte_index
        sig_id.metadata.last_pos = pid.metadata.n + byte_index

```

Fig. 5. Pseudo code for tracking regex signature

4.3 Trace Validate Unit

Trace Validate Unit (TVU) can keep track of status of a trace by simply storing its valid bit. The 1 and 0 value of valid bit indicate the active and inactive state respectively. However, these valid bits are only dedicated to a file or a network buffer that is currently scanned. When byte stream of a new file is sent to our system, all valid bits become invalid and they must be reset to 0. For software implementation of this model, valid bits can be reset by a loop through all traces with the time of $O(n)$, where n is the number of traces. For hardware implementation, there are two approaches for doing this task. The first is that valid bits are implemented using Flip-Flops. By this way, all valid bit can be reset within a single clock cycle. However, this approach is not scalable. The hardware resource will soon become unacceptable when the number of trace increases. The second is that valid bits are stored in onchip-memory. The onchip-memory is less expensive than Flip-Flop and available in all modern re-configurable devices. However, the problem with this approach is that the valid bits must be sequentially reset by replacement of memory content. It will stall the system for n clock cycles before new scanning data is accepted.

In order to overcome the problem of second approach, we propose to use multiple instances of valid bit table. When new scanning data is sent to our system, TVU will change to a new instance of valid bit table. While byte stream is scanned, all other instances can be reset.

5 Evaluation

Our system is implemented using Verilog hardware description language. All design and configuration target to NetFPGA 10G platform [18] with Xilinx 240K logic cell Virtex-5 FPGA device. We use Xilinx ISE 13.4 for synthesis, simulation, placing and routing and mapping.

Table 6. Sub-pattern and Trace Metadata Field Size

	Sub-pattern Metadata					Trace Metadata		
Field	sig_id	next_id	m	n	order	next_id	last_pos	order
Size	14	16	10	10	4	16	31	4

Since the size of sub-pattern metadata is large, we decide to use off-chip memory for keeping them. NetFPGA 10G board has three CY7C1515KV18 SRAM chips and each chip has 72Mbit storage. We use one SRAM for sub-pattern metadata and the other two SRAMs for Fragment-metadata of Pattern Matching Module.

Table 5 shows the field size of Sub-pattern Metadata and Trace Metadata of our implementation. Our analysis of ClamAV virus signature database shows that more than 99% regex signatures have less than 16 sub-patterns, as showed in Fig. 6. Therefore, we restrict to support signatures with the maximum of 16 sub-patterns and configure order field with only 4 bits. Because there are about 9K regex virus signatures and 33K sub-patterns extracted from them, the sig_id field and next_id field are configured to 14 and 16 respectively. With this configuration our system could support up to 16K regex signatures with 64K sub-patterns. The minimum and maximum

byte displacement (m,n) are set to 10 bits since our statistic shows that it could satisfy more than 98% wildcard conditions of regex signatures, as depicted in Fig. 7. The next_id and order fields of Trace Metadata are copied from Sub-pattern Metadata so that they have the same field size. Since the last_pos field keep the byte_index of matched pattern in byte stream, it is needed to be long enough for supporting large files and content. We configure it with 31 bits for capability of tracing maximum 2GB file.

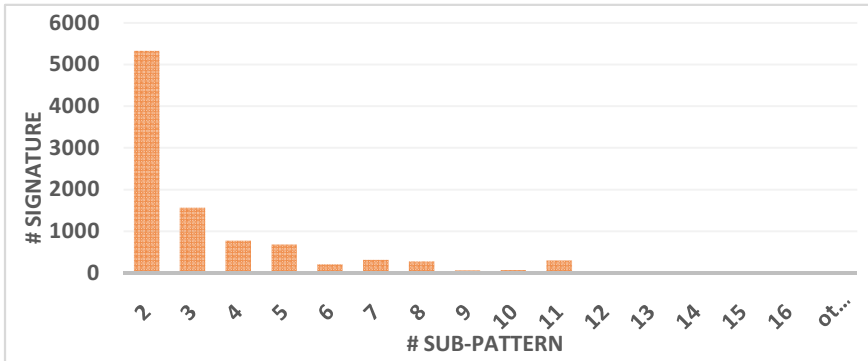


Fig. 6. Statistic of sub-pattern size

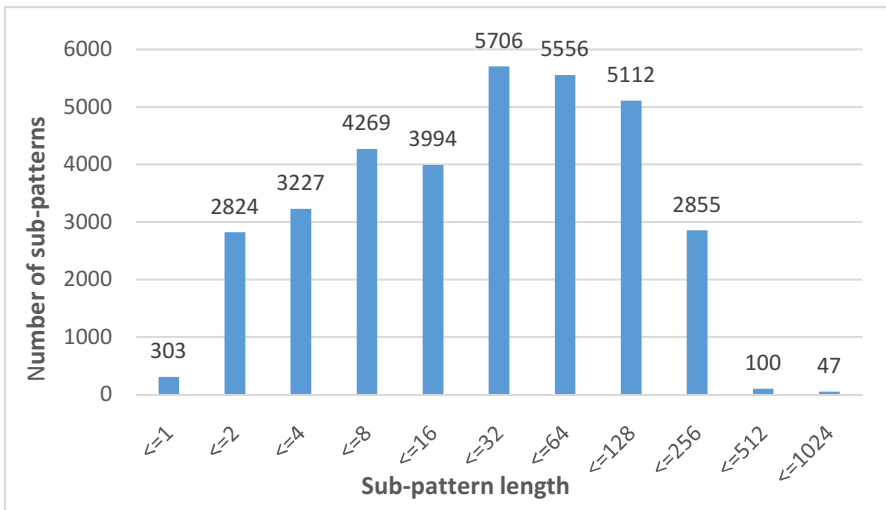


Fig. 7. Distribution of sub-pattern length on ClamAV Signature database

With this configuration, the Sub-pattern Metadata requires 448Kbyte SRAM memory and Trace Metadata requires 816Kbit on-chip memory. The current implemented virus signature utilize only 53% SRAM and 61% Onchip memory.

6 Conclusion and Future Works

In this paper, we proposed a virus signature matching architecture on FPGA. Our system based on Bloom and Bloomier filter approaches and could support both basic and regex virus signatures. Our design mainly bases on memory and could be easily updated for new virus signature set.

In near future, we plan to extend the system to support all ClamAV signature database. The system will be integrated as a part of a hybrid system which is a combination of hardware and software. The hybrid system could exploit the high matching speed of FPGA-based engine while maintain the flexibility and user interface of software application.

Acknowledgments. This research is funded by The Department of Science and Technology in Ho Chi Minh City under grand number 170/2013/HĐ-SKHCN.

References

1. Sourdis, I., Pnevmatikatos, D., Wong, S., Vassiliadis, S.: A reconfigurable perfect-hashing scheme for packet inspection. In: Proceeding of FPL, pp. 644–647 (2005)
2. Thinh, T.N., Kittitornkun, S., Tomiyama, S.: - Applying cuckoo hashing for FPGA-based pattern matching in NIDS/NIPS. In: Proceeding of ICFPT, pp. 121–128 (2007)
3. van Lunteren, J.: High-performance pattern-matching for intrusion detection. In: Proceeding of IEEE Int'l. Conf. on Comp. Comm., pp. 1–13 (2006)
4. Papadopoulos, G., Pnevmatikatos, D.: - Hashing + memory = low cost, exact pattern matching. In: Proceeding of FPL, pp. 39–44 (2005)
5. Thinh, T.N., Kittitornkun, S.: Systolic Array for String Matching in NIDS. In: Proceeding of 4th IASTED Asian Conference Communication System and Networks, April 2-4 (2007)
6. Zhou, X., Xu, B., Qi, Y., Li, J.: MRSI: A fast pattern matching algorithm for anti-virus applications. In: Int'l. Conf. on Networking, pp. 256–261 (2008)
7. Ho, J.T.L., Lemieux, G.G.F.: PERG: A Scalable FPGA-based Pattern-matching Engine with Consolidated Bloomier Filters. In: ICECE Technology, FPT 2008 (2008)
8. Tuan, N.D.A., Hieu, B.T., Thinh, T.N.: High Performance Pattern Matching using Bloom Bloomier Filter. In: The 7th IEEE International Conference Electrical Engineering/ Electronics, Computer, Telecommunications and Information Technology (2010)
9. Bloom, B.: Space/Time Tradeoffs in Hash Coding with Allowance Errors. *Comm. ACM* 13(7), 422–426 (1970)
10. Ehtesham Rafiq, A.N.M., El-Kharashi, M.W., Gebali, F.: Systolic Arraybased String Matching Unit for Spam Blocking. In: Proceeding of 9th IDEAS (2005)
11. Aho, A.V., Corasick, M.J.: - Efficient string matching: an aid to bibliographic search. *Communications of the ACM* 18, 333–340 (1975)
12. Pnevmatikatos, D.N., Arelakis, A.: Variable-length hashing for Exact Pattern Matching. In: Proceeding of FPL 2006, pp. 1–6 (2006)
13. Song, H., Dharmapurikar, S., Turner, J., Lockwood, J.: Fast Hash Table Lookup Using Extended Bloom Filter: An Aid to Network Processing. *ACM SIGCOMM* 35(4), 181–192 (2005)

14. Chazelle, B., Kilian, J., Robinfeld, R., Tal, A.: - The Bloomier Filter: an Efficient Data Structure for Static Support Lookup Table, pp. 30–39. Society for Industrial and Applied Mathematics (2004)
15. Hasan, J., Cadambi, S., Jakkula, V., Chakradhar, S.: Chisel: A Storage efficient, Collision-free Hash-based Network Processing Architecture. In: Proceeding of 33rd International Symposium on Computer Architecture, pp. 203–215
16. Think, T.N., Surin, K., Shigenori, T.: PAMELA: Pattern Matching Engine with Limited-Time Update for NIDS/NIPS. IEICE(E92-D) (5), 1049–1061 (May 2009)
17. Hieu, T.T., Think, T.N., Shigenori, T.: ENREM: An efficient NFA-based regular expression matching engine on reconfigurable hardware for NIDS. Journal of Systems Architecture 59(4), 202–212 (2013)
18. NetFPGA. Netfpga platform technical specifications (2012)

Constructing Private Indexes on Encrypted Data for Outsourced Databases

Yi Tang^{1,2}, Ji Zhang^{1,2}, and Xiaolei Zhang^{1,2}

¹ School of Mathematics and Information Science
Guangzhou University, Guangzhou 510006, China
ytang@gzhu.edu.cn

² Key Laboratory of Mathematics and Interdisciplinary Sciences
of Guangdong Higher Education Institutes
Guangzhou University, Guangzhou 510006, China

Abstract. Data privacy and query performance are two closely linked and inconsistent challenges for outsourced databases. Using mixed encryption methods on data attributes can partially reach a trade-off between the two challenges. However, encryption cannot always hide the correlations between attribute values. When the data tuples are accessed selectively, inferences based on comparing encrypted values could be launched, and some sensitive values may be disclosed. In this paper, we explore the intra-attribute based and inter-attribute based inferences in mixed encrypted databases. We develop a method to construct private indexes on encrypted values to defend against those inferences while supporting efficient selective access to encrypted data. We have conducted some experiments to validate our proposed method.

1 Introduction

Encryption is an essential secure technique for outsourced databases. Executing selective queries securely and efficiently over those encrypted data is a main concern in database research community. Early efforts are focused on translating plain queries at trusted client side into corresponding encrypted queries on untrusted server side and assume that each tuple is encrypted with a single key [5][8]. This implies that a certain user may access any encrypted tuples if he gets the decryption key.

Access control provides users selective restriction of access to data resources. The selective encryption methods use different keys to encrypt different data portions such as tuples or attributes [9]. To avoid users from managing too many keys, the keys can be derived from user hierarchy [6] and the traditional ciphers can be changed into some asymmetric methods such as the attribute-based encryption (ABE) method [17]. Although these methods provide an effective and possible way to combine encryption with access control, the fulfillment of access control depends on the precision of locating encrypted data and the comprehensibility of decrypted data. This means that some decryption efforts on client side are wasted. It needs purified techniques to locate encrypted tuples on server side as precisely as possible.

The granularity of data encryption in outsourced databases can be a tuple or an attribute. When the encryption is on a tuple, it generally needs to introduce auxiliary attributes to index encrypted values [8], and the constructed indexes can be value-based or bucket-based [14]. It is obviously that value-based indexing method provides more accuracies in locating encrypted tuples than the bucket-based method. Encrypting a tuple as a whole may limit the efficiency of data operation because it needs decryption before performing further data operations. Compared to the encryption in tuple, encrypting a single attribute may provide more flexibilities in data management. To meet the requirements with different data operations, an attribute value may be encrypted by mixed encryption methods. For example, in order to meet the various query requirements, an attribute value may be simultaneously encrypted by a symmetric encryption, an order preserving encryption, and a homomorphic encryption, respectively [11]. When an attribute is encrypted by mixed encryption methods, it may require extra attributes to represent related encrypted values [13]. Note that in this case, the attribute index is value-based which can be built directly on the encrypted values.

In the scenario of outsourced databases, on one hand, the database service provider is not required to guarantee a strict separation among data portions available to different users. On the other hand, if encrypted values on some attributes cannot be distinguished on the tuples with different access control lists, the equality relations on the plain values among those tuples will be demonstrated implicitly. This may lead an adversary user draw inferences on those tuples although he has no right to access them.

In this paper, we consider the case of data encryption granularity in attributes. We will address the issue of defend against inference attacks by mitigating the explicit equality relations among attributes. We try to make a trade-off between efficient querying and selective encryption for access controls. The contributions of this paper can be enumerated as follows.

1. We explore the inferences of sensitive information due to the equality relations between encrypted attribute values.
2. We argue that the inference attacks could be not only launched on the equality relations between values of a same attribute but also on the equality relations between different attributes.
3. We introduce an encryption key constructing method which is not only depended on the key materials shared in users and data owners but also depended on the attribute related access control policy.

The rest of this paper is structured as follows. In Section 2, we first overview some encryption methods addressed in mixed encrypted databases, and then give an example to show the cases of combining access control lists with encrypted attributes. We also demonstrate possible inference attacks in those cases. In Section 3, we discuss the intra-attribute inference attack and the inter-attribute attack and introduce random salt to defend against the inference attacks. In Section 4, we discuss the execution of SQL queries in mixed encrypted database via

some auxiliary attributes and tuples. In Section 5, we conduct some experiments to validate our proposed method. And finally, the conclusion is drawn in Section 6.

2 Background

As illustrated in Fig. 1, three parties are involved in the outsourced database scenario where the users and the proxy are on client side and a provider on server side provides outsourced database services. When user u initiates a query q_u , a proxy at client side will translate the plain q_u into an encrypted version q_u^s and send it to the remote server. After executing the query q_u^s over encrypted data on server side, the query result T_u^s will be returned to the proxy as replies. The proxy will decrypt tuples in T_u^s , perform computations over the decrypted tuples according to the conditions in q_u , and finally return results to user u . The specified user, *Administrator* (shortened as a), acts as the owner of the data. He maintains the access control lists and shares some secrets with users. He also has his own private secrets to deal with data tuples.

2.1 Encryption Methods

Two encryption methods, the symmetric key and the asymmetric key, are often adopted in real applications. The symmetric key encryption is traditional and uses a same key for both encryption and decryption. This key is uniquely associated with one or more users and should be made private. The AES algorithm is a typical symmetric key encryption algorithm. When using AES cipher to encrypt a message, the ciphertext is depending on the encryption key and the initialization vector (IV). It implies that the encrypted results is deterministic when the encryption key and IV are given. In this case, the mapping between plaintext and ciphertext is injective. We label this kind of encryption as *DT*.

Order-preserving encryption (*OP*) is a kind of symmetric encryption scheme that preserves numerical ordering of plaintext. It was first suggested in [1], and was deeply studied in [2][3]. According to this method, $x < y \Leftrightarrow OP(x) < OP(y)$

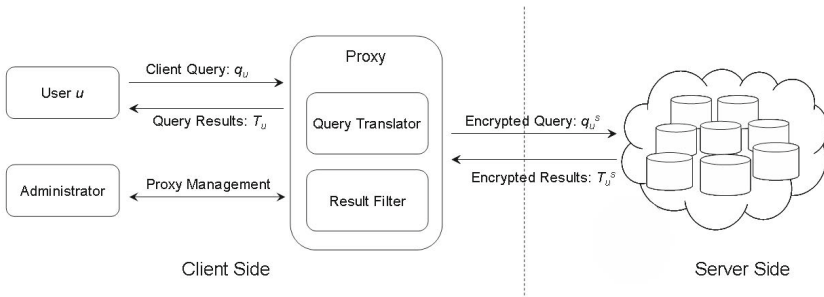


Fig. 1. The outsourced databases scenario

where x and y are plain numbers, and $OP(x)$ and $OP(y)$ are corresponding OP-encrypted version, respectively. This special feature makes it possible to perform inequality comparisons on encrypted data while not decrypting them.

The asymmetric key encryption is another class of encryption algorithms whose keys are in pairs. This method is also known as public key cryptography, since each user will create a related key pair, and make one public (*the public key*) and the other secret (*the private key*). When encrypting a message with an asymmetric key encryption algorithm, a random nonce is often introduced. This implies that the encrypted results are variable in different encryption procedures.

Running a database application often requires some computations on data attributes. In the case of outsourced databases, the ideal solution is to perform computations over encrypted data. Since homomorphic encryption (HE) enables an equivalent relation exists between one operation performed on the plaintext and another operation on the ciphertext, it is considered as an effective solution to this issue. According to the relations supported, the homomorphic encryption methods can be partially (*PH*) [10] or fully (*FH*) [7]. Considering that the current *FH* methods need huge space and computing cost, they are impractical in real applications. It seems that the *PH* methods are more practical in applications although the computations are still expensive. For example, the Paillier cryptosystem, a kind of *PH* method, is an additive homomorphic cryptosystem [10] which is considered as a probabilistic asymmetric encryption algorithm.

2.2 The Assumptions

We assume that the proxy is trust and secure, and the user can access any outsourced data based on his access rights. The service provider is honest but curious, sometimes a bit greedy. This means that the provider can provide the service he claims to be able to provide but he may leak some stored encrypted tuples out to others for curiosity or benefits. When no ambiguity is possible, we also call the service provider as the server.

We also assume that the encryption algorithms used in data attributes are limited in three methods, i.e., the deterministic symmetric encryption *DT*, the order-preserving encryption method *OP*, and the partially homomorphic encryption method *PH*.

2.3 An Outsourced Database Example

Considering an original relation R with an access control list (ACL) demonstrated in Table 1(a), three users, u , v , and w , are associated with this table, and the encryption granularity is in attributes. Suppose that the attributes *Sales* and *Inventory* must be kept privacy and the *ShopID* can be in plain. The following four SQL queries are needed to execute over the relation R .

q_1 : **select** ShopID **from** R **where** Sales = 60

q_2 : **select** ShopID **from** R **where** Sales < 65

q_3 : **select sum(Sales) from R**
 q_4 : **select ShopID from R where Inventory = 60**

Table 1. A Relation in plain and encrypted with *ACL*

(a) Original Relation with *ACL*

	<i>ACL</i>	Sales	Inventory	ShopID
t_1	u	80	50	3
t_2	u	60	40	2
t_3	u, v	60	80	1
t_4	v	50	40	5
t_5	v, w	60	50	4

(b) Encrypted Relation with ACL-specified Mixed Encrypted Version

	tid	Sal_DT	Sal_OP	Sal_PH	Inv_DT	ShopID
t_1^s	1	$DT_u(80)$	$OP_u(80)$	$PH_u(80)$	$DT_u(50)$	3
t_2^s	2	$DT_u(60)$	$OP_u(60)$	$PH_u(60)$	$DT_u(40)$	2
t_3^s	3	$DT_{uv}(60)$	$OP_{uv}(60)$	$PH_{uv}(60)$	$DT_{uv}(80)$	1
t_4^s	4	$DT_v(50)$	$OP_v(50)$	$PH_v(50)$	$DT_v(40)$	5
t_5^s	5	$DT_{vw}(60)$	$OP_{vw}(60)$	$PH_{vw}(60)$	$DT_{vw}(50)$	4

(c) Encrypted Relation with user-specified Mixed Encrypted Version

	tid	Sal_DT	Sal_OP	Sal_PH	Inv_DT	ShopID
t_1^s	1	$DT_u(80)$	$OP_u(80)$	$PH_u(80)$	$DT_u(50)$	3
t_2^s	2	$DT_u(60)$	$OP_u(60)$	$PH_u(60)$	$DT_u(40)$	2
t_3^s	3	$DT_u(60)$	$OP_u(60)$	$PH_u(60)$	$DT_u(80)$	1
		$DT_v(60)$	$OP_v(60)$	$PH_v(60)$	$DT_v(80)$	
t_4^s	4	$DT_v(50)$	$OP_v(50)$	$PH_v(50)$	$DT_v(50)$	5
t_5^s	5	$DT_v(60)$	$OP_v(60)$	$PH_v(60)$	$DT_v(50)$	4
		$DT_w(60)$	$OP_w(60)$	$PH_w(60)$	$DT_w(50)$	

To perform these queries effectively and efficiently over encrypted data, we need three encryption methods, *DT*, *OP*, and *PH* to encrypt the attribute Sales and Inventory. For example, the comparison and summation are performed on attribute Sales, and thus it needs the three encryption methods to encrypt the Sales values, respectively. Note that we also need to define four extra attributes, Sal_DT, Sal_OP, Sal_PH, and Inv_DT, to support queries on server side. To simplify the key management at client side, although it needs different keys for users to encrypt attributes selectively because of the access control list, it is generally assumed that an authorized user will encrypt all attributes of a tuple with a same key.

An intuitive key assignment method is ACL-specified, i.e., the key for encrypting Sales and Inventory in each tuple depends on the corresponding ACL list. As shown in Table 1(b), the keys are associated with the ACL lists. However, the ACL-specified key may overload the number of keys for each user. For example, when user u issues the query q_1 , the condition Sales = 60 will be translated into

$\text{Sal_DT in } \{DT_u(60), DT_{uv}(60), DT_{uw}(60), DT_{uvw}(60)\}$ where $\{DT_u(60), DT_{uv}(60), DT_{uw}(60), DT_{uvw}(60)\}$ is called as the matching set. Note that the size of matching set for a database with n users is $\sum_{i=0}^{n-1} C_{n-1}^i$, the ACL-specified key assignment method obviously increases computation costs either in client or in server side.

We consider the user-specified key assignment method. It means that each user has his own key to encrypt attribution values, and hence private indexes could be constructed on those encrypted values. Table 1(c) demonstrates the encrypted relation with user-specified key assignments. In this case, the query q_1 issued from user u , the condition $\text{Sales} = 60$ will be translated into $\text{Sal_DT in } \{DT_u(60)\}$. It is obviously concise and efficient when comparing to the case of ACL-specified key assignment.

It seems perfect when executing query over encrypted data with user-specified mixed encryption methods. However, the service provider is a pure storage service provider, he has no obligation to design appropriate storage constraints to separate tuple sets on access rights. A set of encrypted tuples may be leaked intentionally or unintentionally. This means that an adversary user could potentially get some encrypted tuples that he cannot access. Though the adversary cannot take the plain values by decryption, the same encrypted values could open a door to draw inferences on those tuples and thus the inference attack could be launched.

There are two kinds of inferences which could be launched to the Table 1(c).

- *The Intra-attribute-based Inference* Considering the tuples t_2^s and t_3^s , user u can access t_2^s and t_3^s , and user v can only access t_3^s according to the *ACL* lists. However, v can realize that the decrypted value of $t_2^s.\text{Sal_DT}$ is 60 because he finds that the value of $t_2^s.\text{Sal_DT}$ is appeared in $t_3^s.\text{Sal_DT}$ and he knows that $t_3^s.\text{Sales}$ is 60. Note that in this case, v neither has the right to access t_2 nor has the key associated with user u .
- *The Inter-attribute-based Inference* Considering the tuples t_4^s and t_5^s , user v can access t_4^s and t_5^s , and user w can only access t_5^s according to the *ACL* lists. However, w can realize that the decrypted value of $t_4^s.\text{Sal_DT}$ is 50 because he finds that the value of $t_4^s.\text{Sal_DT}$ is appeared in $t_5^s.\text{Inv_DT}$ and he knows that $t_5^s.\text{Inventory}$ is 50. Note that in this case, w neither has the right to access t_4 nor has the key associated with user v .

The reason why these attacks could be launched is because of the conflicts introduced by the inconsistent relationships between the equal encrypted attribute values and the unequal access control lists in some tuples. For example, the tuples t_2 and t_3 are such tuples that are conflicting over attribute *Sales*. This inference can be prevented if the equality relation between encrypted values is destroyed.

3 The Inference Attacks and Defences

In this section, we will explore the inference attack which is based on the equality relations among encrypted attribute values.

3.1 The Intra-attribute Based and Inter-attribute Based Inferences

Definition 1. [14] *The tuples t_i and t_j are called intra-attribute conflicting tuples over attribute A , denoted by $t_i \sim_A t_j$, if the condition, $t_i.A = t_j.A \wedge t_i.ACL \neq t_j.ACL \wedge t_i.ACL \cap t_j.ACL \neq \phi$, is satisfied.*

For the tuples in Table 1(a), we have $t_2 \sim_{\text{Sales}} t_3$ because of the satisfied condition $t_2.\text{Sales} = t_3.\text{Sales} \wedge t_2.ACL \neq t_3.ACL \wedge t_2.ACL \cap t_3.ACL \neq \phi$. It means that t_2 and t_3 are intra-attribute conflicting over attribute Sales.

Definition 2. *The tuples t_i and t_j are called inter-attribute conflicting tuples over attributes A and B , denoted by $t_i \sim_{A,B} t_j$, if the condition, $t_i.A = t_j.B \wedge t_i.ACL \neq t_j.ACL \wedge t_i.ACL \cap t_j.ACL \neq \phi$, is satisfied.*

For the tuples in Table 1(a), we have $t_4 \sim_{\text{Sales, Inventory}} t_5$. This is because that the condition $t_4.\text{Sales} = t_5.\text{Inventory} \wedge t_4.ACL \neq t_5.ACL \wedge t_4.ACL \cap t_5.ACL \neq \phi$ is satisfied, and we say t_4 and t_5 are inter-attribute conflicting over attributes Sales and Inventory.

Definition 3. *The encryption method $enc : X \rightarrow Y$ is equality-preserved if $\forall x_1, x_2 \in X$ with $x_1 = x_2$, we have $enc(key, x_1) = enc(key, x_2)$ where key is an encryption key.*

For the encryption methods we discussed previously, both *DT* and *OP* methods are equality-preserved, but the *PH* method is not equality-preserved. Furthermore, both *DT* and *OP* are also injective.

The logic behind the two discussed inference attacks lies in the observation that the fact of two equal images of an equality-preserved injective mapping implies that the corresponding preimages are equal. Recall the case in Table 1(c), if the attribute values are encrypted by an equality-preserved injective function and an adversary user obtains some encrypted tuples he has no rights to access, he could infer some plain attribute values in those tuples via intra-attribute conflicting or inter-attribute conflicting tuples which he is involved.

As inference instances, if user v can obtain encrypted tuples t_2^s in Table 1(c), he can infer that $t_2^s.\text{Sal-DT}$ is the encrypted version of value 60 because of the intra-conflicting relationship over attribute Sales between t_2 and t_3 . Similarly, if user w can obtain encrypted tuples t_4^s in Table 1(c), he can infer that $t_4^s.\text{Sal-DT}$ is the encrypted version of value 50 because of the inter-conflicting relationship over attributes Sales and Inventory between t_4 and t_5 .

Definition 4. *Let \mathcal{RA} be the attribute set in relation R . A function f is conflict-free if $\forall t_i, t_j \in R$ and $\forall A, B \in \mathcal{RA}$, $A \neq B$,*

1. *if $t_i \sim_A t_j$, $\forall u \in t_i.ACL \cap t_j.ACL$, $f(t_i.A) \neq f(t_j.A)$;*
2. *if $t_i \sim_{A,B} t_j$, $\forall u \in t_i.ACL \cap t_j.ACL$, $f(t_i.A) \neq f(t_j.B)$.*

According to the above definition, if the encryption method is conflict-free, both inferences could be blocked. To construct a conflict-free function, we can change the equality-preserved function into a kind of piecewise function to destroy the characteristics of equality-preserved in conflicting tuples.

3.2 Constructing Conflict-Free Partition

Definition 5. Let \mathcal{A} be a set of attributes in relation R . A conflict-free partition $\mathcal{C}_{\mathcal{A}}$ with size m is a set of tuple sets $\{C_1, C_2, \dots, C_m\}$ such that $\cup_{i=1}^m C_i = R$, $C_i \cap C_j = \phi$ where $i \neq j$ and $1 \leq i, j \leq m$, and $\forall C \in \mathcal{C}_{\mathcal{A}} : \forall t_{i'}, t_{j'} \in C : \forall A, B \in \mathcal{A}, t_{i'} \approx_{\mathcal{A}} t_{j'} \wedge t_{i'} \approx_{A,B} t_{j'}$.

Considering the last condition in Definition 5, if only the $t_{i'} \approx_{\mathcal{A}} t_{j'}$ is satisfied, the partition $\mathcal{C}_{\mathcal{A}}$ is intra-attribute conflict-free, meanwhile, while if only the $t_{i'} \approx_{A,B} t_{j'}$ is satisfied, the partition $\mathcal{C}_{\mathcal{A}}$ is inter-attribute conflict-free.

Lemma 1. [14] Finding an intra-attribute conflict-free partition with minimum size is NP-hard.

Definition 6. Let $\mathcal{A} = \{A_1, A_2, \dots, A_n\}$, the extended attribute $\text{Ext}\mathcal{A}$ over \mathcal{A} is the attribute vector (A_1, A_2, \dots, A_n) .

Definition 7. Let t_i and t_j be two tuples over $\mathcal{A} = \{A_1, A_2, \dots, A_n\}$ and $\text{Ext}\mathcal{A}$ is the extended attributes over \mathcal{A} , $t_i.\text{Ext}\mathcal{A} =_{\text{ext}} t_j.\text{Ext}\mathcal{A}$ if $\exists k : 1 \leq k \leq n : t_i.A_k = t_j.A_k \vee \exists k_1, k_2 : 1 \leq k_1 < k_2 \leq n, t_i.A_{k_1} = t_j.A_{k_2}$.

Definition 8. The tuples t_i and t_j are called conflicting tuples over extended attribute $\text{Ext}\mathcal{A}$, denoted by $t_i \sim_{\text{Ext}\mathcal{A}} t_j$, if the condition, $t_i.\text{Ext}\mathcal{A} =_{\text{ext}} t_j.\text{Ext}\mathcal{A} \wedge t_i.ACL \neq t_j.ACL \wedge t_i.ACL \cap t_j.ACL \neq \phi$, is satisfied.

For the example in Table 1(a), we can define the extended attribute SI as $\text{SI} = (\text{Sales}, \text{Inventory})$. And then we have $t_2.\text{SI} =_{\text{ext}} t_3.\text{SI}$ because of $t_2.\text{Sales} = t_3.\text{Sales}$. Since $t_4.\text{Sales} = t_5.\text{Inventory}$, we also have $t_4.\text{SI} =_{\text{ext}} t_5.\text{SI}$. Considering the ACL relationships among tuples t_2, t_3, t_4 , and t_5 , we have $t_2 \sim_{\text{SI}} t_3$ and $t_4 \sim_{\text{SI}} t_5$.

Note that the conflict-free partition over attribute set \mathcal{A} can be viewed as the intra-attribute conflict-free partition over attribute $\text{ext}\mathcal{A}$, we have the following lemma.

Lemma 2. Finding a conflict-free partition is equivalent to finding an intra-attribute conflict free partition.

With previously discussed two lemmas, we have the following theorem.

Theorem 1. Finding a conflict-free partition with minimum size is NP-hard.

3.3 The Algorithms

If we can find a conflict-free partition and define a conflict-free function over tuple sets according to the conflict-free partition, we can find a solution to prevent the inference attacks. A simple and direct strategy is combining random salts with user-specified keys when encrypting data attributes. If two tuples in a same partition are accessible to user u , the same salt can be used to construct a key encrypt attribute values in both tuples, i.e., the encryption key is the same when user u performs attribute encryptions on the two tuples. On the other side, if two tuples

in different partitions, the corresponding salts are different. It means that different keys are used by user u when he encrypts tuples in different partitions. The number of salts depends on the size of partition. However, as described in Theorem 1, finding a conflict-free partition with minimum size is NP-hard. There are two methods to construct a conflict-free partition. The attribute level method defines partitions on each single attribute while the relation level method defines partitions on a set of attributes. The attribute level method may introduce smaller sizes of conflict-free partitions but may also lead to intricate computations. For example, it needs to determine the number of salts for each attributes. Considering that we will perform analytical computations on encrypted data, the relation level method is a better choice. We will adopt a relation level heuristic method in Algorithm 1 to find conflict-free partition $\mathcal{C}_{\mathcal{A}}$ over attribute set \mathcal{A} .

Algorithm 1. Constructing Conflict-free Partition $\mathcal{C}_{\mathcal{A}}$ over Attribute Set \mathcal{A}

```

1:  $\mathcal{C}_{\mathcal{A}} = \phi$ 
2: for each  $t \in R_{\mathcal{A}}$  do
3:   if  $\mathcal{C}_{\mathcal{A}} == \phi$  then
4:      $C = \{t\}$ 
5:      $\mathcal{C}_{\mathcal{A}} = \mathcal{C}_{\mathcal{A}} \cup \{C\}$ 
6:   else
7:      $CandC = \{C | C \in \mathcal{C}_{\mathcal{A}} : \forall t' \in C : \forall A, B \in \mathcal{A} : t' \approx_A t \wedge t' \approx_{A,B} t\}$ 
8:     if  $CandC \neq \phi$  then
9:       random pick  $C_r \in CandC$ 
10:       $C_r = C_r \cup \{t\}$ 
11:     else
12:       $C = \{t\}$ 
13:       $\mathcal{C}_{\mathcal{A}} = \mathcal{C}_{\mathcal{A}} \cup \{C\}$ 
14:     end if
15:   end if
16: end for.

```

According to this algorithm, we initialize the conflict-free partition $\mathcal{C}_{\mathcal{A}}$ as an empty set. When given a tuple t , we will distribute it into a chosen partition $C \in \mathcal{C}_{\mathcal{A}}$ such that t does not conflict with any tuples in C . To achieve this, we first construct a set of candidate partitions $CandC$ where each partition C in $CandC$ is the possible partition that t will be distributed. If the set $CandC$ is empty, we create a new partition, distribute t into it, and append it into $\mathcal{C}_{\mathcal{A}}$. Otherwise, we random choose a partition $C \in CandC$ and distribute t into C . This procedure will be continued until all tuples are distributed, and finally, we can obtain a conflict-free partition $\mathcal{C}_{\mathcal{A}}$.

We use the notation $t.u.salt$ to represent the salt that user u used in tuple t . The process of salt assignment is demonstrated in Algorithm 2.

As in Algorithm 2, we use the conflict-free partition constructed by Algorithm 1 to assign salts user by user. For each user u and each partition $C \in \mathcal{C}_{\mathcal{A}}$,

we extract tuple set $C_u, C_u \subseteq C$, where u is accessible to each tuple in C_u . We assign a random salt $salt_u$ for u to construct his private encrypted attributes in $t \in C_u$.

Algorithm 2. Assigning *salt* via the Conflict-free Partition \mathcal{C}_A

```

1: for each user  $u$  do
2:   for each  $C \in \mathcal{C}_A$  do
3:      $C_u = \{t | t \in C \wedge u \in t.ACL\}$ 
4:      $salt_u = ChooseSalt_u(C_u)$ 
5:     for each  $t \in C_u$  do
6:        $t.u.salt = salt_u$ 
7:     end for
8:   end for
9: end for.
```

4 Extra Attributes for Supporting SQL Queries

With the user-specified mixed encryption, multiple encrypted value-index pairs (one for each authorized user) may be defined for the same encrypted attribute. It means that the attribute is a set type. However, current SQL database implementations do not support this kind of attribute. We will adopt the duplicating-tid strategy to support the representation of sets of values. As an example, after finishing the process of conflict-free partition and the salt assignment, the table in Table 1(c) is translated into the table in Table 2.

As an example for encrypted attributes in Table 1(c), we analyze the notation $DT_u(u_1, 80)$ for attribute *Sal_DT*. This notation denotes an encrypted value of 80. Firstly, user u combine salt u_1 with the secret shared with the data owner to generate a private key. And then, this key is used in *DT* encryption to encrypt the value 80. Finally, the encrypted value is obtained and denoted by $DT_u(u_1, 80)$. Other encrypted attributes can be analyzed in similar way except the attribute *Sal_PH*. As an example, the notation $PH_u(80)$ denotes the value 80 is encrypted with a *PH* method and the encryption key is from the secret shared between user u and the data owner.

To support SQL query over encrypted data on our proposed indexes, we add two attributes, the attribute *tid* is used to distinguish the duplicate tuples in original tables and the attribute *sid* is used to represent the salt used in this tuple. We use user-specified function $sid()$ to generate corresponding *sid* values. For example, we can define $sid()$ with a user specified cryptological hash function.

With the duplicating-tid strategy, we also simplify the access control list for each tuples, i.e., each tuple is accessible to a single user except the *Administrator*. We follow the basic notations in [4] and refer to an access control policy as an authorization policy. The notation $t.uid = u$ means that user u can access tuple t . It implies that the SQL queries initiated by user u implicitly includes a condition $uid = u$.

Table 2. Conflict-free Encrypted Relation with Mixed Encrypted Relation

tid	sid	Sal_DT	Sal_OP	Sal_PH	Inv_DT	ShopID
1	$sid_u(u_1)$	$DT_u(u_1, 80)$	$OP_u(u_1, 80)$	$PH_u(80)$	$DT_u(u_1, 50)$	3
2	$sid_u(u_1)$	$DT_u(u_1, 60)$	$OP_u(u_1, 60)$	$PH_u(60)$	$DT_u(u_1, 40)$	2
3	$sid_u(u_2)$	$DT_u(u_2, 60)$	$OP_u(u_2, 60)$	$PH_u(60)$	$DT_u(u_2, 80)$	1
3	$sid_v(v_1)$	$DT_v(v_1, 60)$	$OP_v(v_1, 60)$	$PH_v(60)$	$DT_v(v_1, 80)$	1
4	$sid_v(v_1)$	$DT_v(v_1, 50)$	$OP_v(v_1, 50)$	$PH_v(50)$	$DT_v(v_1, 40)$	5
5	$sid_v(v_2)$	$DT_v(v_2, 60)$	$OP_v(v_2, 60)$	$PH_v(60)$	$DT_v(v_2, 50)$	4
5	$sid_w(w_1)$	$DT_w(w_1, 60)$	$OP_w(w_1, 60)$	$PH_w(60)$	$DT_w(w_1, 50)$	4

To supported the SQL queries over the proposed conflict-free mixed encrypted database, each user u has the knowledge of: (1)the maximum number of random salts for tuples that he can access; (2)the salt generation function used by the data owner to generate; (3)the secret shared by the data owner to construct encryption key.

Recall the queries we discussed in Section 2

- q_1 : **select ShopID from R where Sales = 60**
 q_2 : **select ShopID from R where Sales < 65**
 q_3 : **select sum(Sales) from R**
 q_4 : **select ShopID from R where Inventory = 60**

The translation of the query q_1 is direct. When user u issues this query, q_1 will be translated into $q_{u,1}^s$: **select ShopID from R^s where Sal_DT in $\{DT_u(u_1, 60), DT_u(u_2, 60)\}$** . It is something different when user a , the *Administrator*, issues this query, it will be translated into $q_{a,1}^s$: **select distinct ShopID from R^s where Sal_DT in $\{DT_u(u_1, 60), DT_u(u_2, 60), DT_v(v_1, 60), DT_v(v_2, 60), DT_w(w_1, 60)\}$** . Similar methods could be adopted to the query q_4 .

For the query q_2 , the introduction of salt destroy the preserved order in attribute Sal_OP for each user, we add an auxiliary attribute, sid, into encrypted relation to distinguished different salts. For example, when user u issues this query, it will be translated into $q_{u,2}^s$: **select ShopID from R^s where ((sid == $sid_u(u_1)$ and Sal_OP < $OP_u(u_1, 65)$) or (sid == $sid_u(u_2)$ and Sal_OP < $OP_u(u_2, 65)$))**.

For the query q_3 , the computation is different. The **sum** can be directly computed over encrypted data because of the additive homomorphic encryption. When user u issues this query, the server side query $q_{u,3}^s$ will be as **select sum(Sal_PH) from R^s where sid in $\{sid_u(u_1), sid_u(u_2)\}$** . However, when user *Administrator*, the query will be translated into following sequences:

1. **select distinct tid, sid, Sal_PH into $TempR^s$ from R^s**
2. **select @sum $_u$ = sum(Sal_PH) from $TempR^s$ where sid in $\{sid_u(u_1), sid_u(u_2)\}$**
3. **select @sum $_v$ = sum(Sal_PH) from $TempR^s$ where sid in $\{sid_v(v_1), sid_v(v_2)\}$**
4. **select @sum $_w$ = sum(Sal_PH) from $TempR^s$ where sid in $\{sid_w(w_1)\}$**
5. **@sum = sum $_u$ + sum $_v$ + sum $_w$**

5 Experiments and Discussion

5.1 The Datasets

To evaluate the behavior of our proposed method, we need two types of materials for experiments, the data tuples and the authorized users for tuples.

For the data tuples, we first generate a relational table with 800000 tuples following the TPC-H benchmark specifications, and then randomly select 3000, 8000, 13000, and 18000 tuples to construct tables *Data3k*, *Data8k*, *Data13k*, and *Data18k*, respectively. Each table contain the same three attributes, including 10000, 9999, and 1000 distinct integers, respectively.

For the authorized users for tuples, we extract the authors coauthored with Professor Xuemin Shen from the DBLP repository. In particular, we extract the top m most productive authors and construct authors set of size n from the repository. We view the constructed authors set as the authorized users set, i.e., the *ACL* lists for tuples. In our experiments, we set m as 40, 90, 140, and 190, respectively, and correspondingly, we set n as 60, 124, 204, and 297, respectively. We denote our constructed *ACL* lists as *ACL1*, *ACL2*, *ACL3*, and *ACL4*, respectively.

5.2 The Results

We construct the conflict-free partitions for each instance table with each constructed *ACL* list and compute the maximum/median number of salts assigned to each user. We repeat the computation 100 times and compute the average maximum/median number of salts per user. Our experiments are focused on counting the number of salts assigned to users. This is because that the number of salts per user assigned determines the extra computation costs when a user executes SQL queries at client side.

Figure 2 demonstrates the average number of salts per user on different datasets where Figure 2(a) shows the average maximum salt number and Figure 2(b) shows the average median salt number. We find that as the number of tuples increasing, both maximum number and median number are also increasing. For the average maximum number of salts a user could be used is limited in the interval [4,12]. This means that when translating client side SQL queries into server side query versions, the average number of test index values is at most 12. Comparing with the *ACL*-specified indexes, we can achieve the same query results with much smaller computation overloads in both client side and server side.

On the other hand, as demonstrated in Figure 3 with Figure 3(a) shows the average maximum salt number and Figure 3(b) shows the average median salt number per user, respectively, comparing to the user numbers. We find that the number of salts is decreased as the number of users is increased when given a certain dataset. This is because the increased number of users will decrease the possibilities of conflicting tuples.

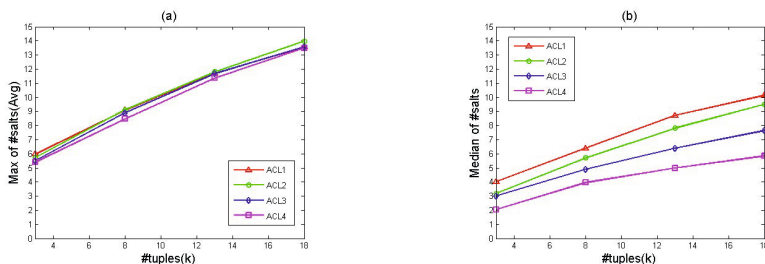


Fig. 2. Average number of salts per user on tuples

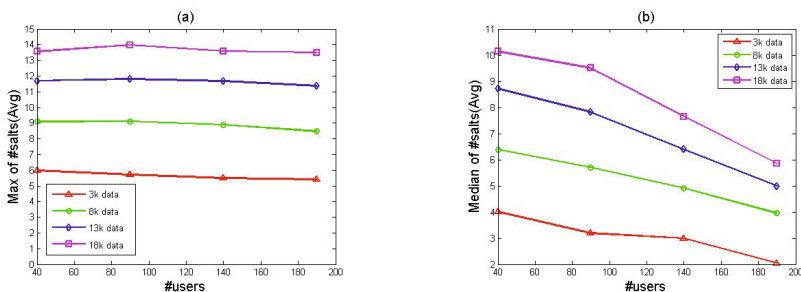


Fig. 3. Average number of salts per user on users

5.3 Related Work

Outsourcing data to third parties out of the control of data owners requires storing data encrypted on remote servers. To avoid storing many different encrypted versions of a same tuple on servers, encrypting each tuple with a single key is a common knowledge. Since the early efforts on outsourced databases [5][8] are focused on how to translate the client-side plain queries into corresponding server-side encrypted versions, they assume that all the tuples are encrypted by a same key. It implies that a certain user may have the full rights to access any encrypted tuples if he gets the decryption key.

The selective encryption methods use different keys to encrypt different data portions such as tuples or attributes [9]. To avoid users from managing too many keys, the keys can be derived from user hierarchy [6]. And also, the traditional ciphers are replaced with the attribute-based encryption (ABE) method to encrypt data [17]. However, the access controls provided by these methods depend on the readability of decrypted data. This means that some decryption efforts on client side are wasteful.

Other efforts are on developing new ciphers for keyword searching on encrypted data. However, either the symmetric encryption scheme [12] or the asymmetric encryption scheme [16] cannot prevent the curious service provider locating the positions with the same method. We note that locating encrypted tuples implies execute comparison operations over encrypted data on server without decryption.

The partially [10] or fully [7] homomorphic encryption methods can be used to perform the comparison. But, as mentioned previously, if the comparison results could be distinguished on server, the curious service provider could also manipulate in the same way to obtain the results of comparison.

To improve the speed of encrypted data retrieval operations on server, several index techniques are proposed. The CryptDB scheme [11] defines layers of encryption for different types of database queries. For executing a specific query, layers of encryption are removed by decrypting to an appropriate layer and the tuple index is directly on the encrypted data. This method may lead many sensitive values be stored to the level defined by the weakest encryption scheme. No inference attacks are considered in this scheme. The DAS (Database as a Service) model [8] proposes a bucketization method to construct the index. This index is defined on an auxiliary attribute which is associated with the corresponding original attribute. However, there is no formal security analysis about this kind of index. Value-based index is discussed in [5]. Comparing with the bucketization index, the value-based index locate encrypted data in high accuracy but also disclose many other useful information such as the data distribution.

To our knowledge, the authors in [14] firstly address the inferences of encrypted data in outsourced databases. They discuss a kind of inference attack introduced by the value-based index. The addressed inferences are on the explicit equality relations among tuples with different access rights which we call intra-attribute based inference in this paper. But they do not address the inter-attribute based inference on the equality relations between attributes.

6 Conclusion

Ensuring data privacy and improving query performance are two closely linked challenges for outsourced databases. Using different encryption methods to data attributes can reach an explicit trade-off between these two challenges. However, encryption cannot always conceal relations between attribute values. When the data tuples are accessed selectively, inference attacks by comparing encrypted values could be launched. In this paper, we explore the intra-attribute based and inter-attribute based inferences in mixed encrypted databases. We develop a method to construct private indexes on user-specified encrypted values to defend against the inferences while supporting efficient selective access to encrypted data. Possible future work may include the data consistency management in various encrypted versions and the defence of inference attacks which is introduced by the collusion between other users and the service providers.

Acknowledgments. This paper was partially supported by the Natural Science Foundation of Guangdong Province under grant S2012040007370.

References

1. Agrawal, R., Kiernan, J., Srikant, R., Xu, Y.: Order-preserving encryption for numeric data. In: Proceedings of SIGMOD 2004, pp. 563–574 (2004)
2. Boldyreva, A., Chenette, N., Lee, Y., O’Neill, A.: Order-preserving symmetric encryption. In: Joux, A. (ed.) EUROCRYPT 2009. LNCS, vol. 5479, pp. 224–241. Springer, Heidelberg (2009)
3. Boldyreva, A., Chenette, N., O’Neill, A.: Order-Preserving Encryption Revisited: Improved Security Analysis and Alternative Solutions. In: Rogaway, P. (ed.) CRYPTO 2011. LNCS, vol. 6841, pp. 578–595. Springer, Heidelberg (2011)
4. Chaudhuri, S., Kaushik, R., Ramamurthy, R.: Database Access Control and Privacy: Is there a common ground? In: Proceedings of CIDR 2011, pp. 96–103 (2011)
5. Damiani, E., Vimercati, S., Jajodia, S., Paraboschi, S., Samarati, P.: Balancing Confidentiality and Efficiency in Untrusted Relational DBMSs. In: Proceedings of ACM CCS 2003, pp. 93–102 (2003)
6. Damiani, E., Vimercati, S., Foresti, S., Jajodia, S., Paraboschi, S., Samarati, P.: Key Management for Multi-user Encrypted Databases. In: Proceedings of StorageSS 2005, pp. 74–83 (2005)
7. Gentry, C.: Fully Homomorphic Encryption Using Ideal Lattices. In: Proceedings of STOC 2009, pp. 169–178 (2009)
8. Hacigumus, H., Iyer, B., Li, C., Mehrotra, S.: Executing SQL over Encrypted Data in the Database-Service-Provider Model. In: Proceedings of ACM SIGMOD 2002, pp. 216–227 (2002)
9. Miklau, G., Suciu, D.: Controlling Access to Published Data Using Cryptography. In: Proceedings of VLDB 2003, pp. 898–909 (2003)
10. Paillier, P.: Public-Key Cryptosystems Based on Composite Degree Residuosity Classes. In: Stern, J. (ed.) EUROCRYPT 1999. LNCS, vol. 1592, pp. 223–238. Springer, Heidelberg (1999)
11. Popa, R., Redfield, C., Zeldovich, N., Balakrishnan, H.: CryptDB: Protecting Confidentiality with Encrypted Query Processing. In: Proceedings of SOSP 2001, pp. 85–100 (2011)
12. Song, D., Wagner, D., Perrig, A.: Practical Techniques for Searches on Encrypted Data. In: Proceedings of IEEE S&P 2000, pp. 44–55 (2000)
13. Tu, S., Kaashoek, M.F., Madden, S.: Zeldovich.Processing Analytical Queries over Encrypted Data. In: Proceedings of VLDB 2013 (2013)
14. Vimercati, S., Foresti, S., Jajodia, S., Paraboschi, S., Samarati, P.: Private Data Indexes for Selective Access to Outsourced Data. In: Proceedings of WPES 2011, pp. 69–80 (2011)
15. Wang, H., Lakshmanan, L.: Efficient Secure Query Evaluation over Encrypted XML Databases. In: Proceedings of VLDB 2006, pp. 127–138 (2006)
16. Yang, G., Tan, C.H., Huang, Q., Wong, D.S.: Probabilistic Public Key Encryption with Equality Test. In: Pieprzyk, J. (ed.) CT-RSA 2010. LNCS, vol. 5985, pp. 119–131. Springer, Heidelberg (2010)
17. Yu, S., Wang, C., Ren, K., Lou, W.: Achieving Secure, Scalable, and Fine-grained Data Access Control in Cloud Computing. In: Proceedings of INFOCOM 2010, pp. 534–542 (2010)

An Extensible Framework for Web Application Vulnerabilities Visualization and Analysis

Tran Tri Dang and Tran Khanh Dang

Faculty of Computer Science and Engineering,
Ho Chi Minh City University of Technology, VNU-HCM, Vietnam
{tridang, khanh}@cse.hcmut.edu.vn

Abstract. The popularity of web-based applications makes them interesting targets of cyber attacks. To deal with that threat, discovering existing vulnerabilities is a proactive step. Although there are many web application scanners designed for this task, they lack visual analysis capability and do not collaborate well together. In this paper, we propose a novel visualization technique and a flexible framework to solve the two problems mentioned above. We also develop a prototype based on the proposal and use it to experiment with virtual websites. Experiment results indicate the unique benefits our work offers. But more importantly, it shows that not only improving the visualization technique from a technical viewpoint is needed, but also improving it from a human cognitive viewpoint should be placed at a higher priority.

Keywords: Web vulnerability visualization, security visualization, web application security analysis.

1 Introduction

The widespread use of web applications in different areas makes them interesting targets for attackers. In fact, the 2014 Internet Security Threat Report from Symantec highlights that web-based attacks are up 23% and 1 in 8 legitimate websites has a critical vulnerability [1]. Furthermore, the ease of web development, together with the wide availability of programming libraries, also contributes to a large degree of poor quality web applications written by amateur programmers.

To overcome the above problem, many approaches are proposed by the security researchers and practitioners. Among them, one promising approach is to develop scanning tools that are able to discover the web vulnerabilities automatically. The advantage of these tools is undeniable: they can save a lot of security analysts' time and effort in securing web applications. However, there are some limitations that these tools cannot solve alone. The first limitation is that the report interface of these scanners is not very effective in communicating scanning results to security analysts. In our survey of some selected tools, we find that the generated reports are either too detailed or too general, and it is not easy switching from one level of abstraction to another level without losing the perception of the previous view. The second limitation is that although these tools have the same goal, it is difficult to use them together

to complement each other. There is a reason behind the need to make them work together seamlessly: each tool has its own strengths and weaknesses, and by combining them properly, we can amplify the strengths as well as reduce the weaknesses.

In this paper, we describe our work in solving the 2 limitations of web application vulnerability scanning tools mentioned above. For the first limitation, we propose a novel visualization that can display scanning results aggregated at many levels of detail on a same screen at a same time. In addition, the visualization is enhanced by user interaction techniques to provide more specific information. For the second limitation, we design an extensible framework to transform scanning results of different tools into a common format that can be used later by the visualization component. We also implement a prototype of this framework and use it to demonstrate the benefits our work can provide to security analysts.

The rest of this paper is structured as follow: in Section 2, we review the related works; we describe the framework architecture in Section 3; Section 4 is used to detail the visualization design and user interaction techniques; the implementation details and its test results are explained in Section 5; and finally, we use Section 6 to conclude the paper and suggest some future works.

2 Related Works

Generally speaking, there are two approaches that scanning tools used to look for vulnerabilities in web applications: white-box testing and black-box testing. In white-box testing, a web application is assessed on how well it handles user inputs, data validation, sensitive information communication, database queries, etc. based on existing source code, compiled code, or bytecode [2-5]. As a result, white-box testing tools are language-specific and platform-specific. Therefore, a web application developed in a particular environment can only be tested by compatible tools. This limitation, together with the complex nature of web application code these days, makes white-box testing tools not as popular as black-box testing tools, at least in an industrial setting. On the other hand, in black-box testing, web applications are checked against standard HTTP requests and responses for probable vulnerabilities. To do so, black-box scanners send virtual attacks to target web applications and process respective responses, looking for HTTP error codes or abnormal strings that may indicate the existence of vulnerabilities [6-9]. Because only HTTP requests and responses are used, black-box testing tools are independent of development environments and are widely used in both academic and industrial settings. Based on the differences between the two approaches mentioned above, in this work, we focus on visualizing scanning results from black-box scanners to make our work as popular as possible.

Applying information visualization techniques to solve computer and information system security problems is a rather young research field called security visualization with its own conference organized each year for practitioners and researchers in information visualization and security to exchange their works [10]. One noteworthy feature that differentiates studies in this field from other security research works is the role of human users. While other security studies treat human users as an external entity, this field considers human users as an internal component that can affect the whole system security by their good or bad decisions. To support users in making

effectively and efficiently decisions, security visualization techniques provide tools for situational awareness [11], security data exploration [12], and visual analysis [13]. The level of support is partly determined by the type of data source used in visualization. Data source at a low level of abstraction allows users to do more detailed analysis while data source at a high level of abstraction often brings them overview information more easily. Some popular data source types used in security visualization include network packet [14], IDS log [15], firewall configuration [16], etc. Although information visualization is used in many security domains, it does not attract much interests from web application security researchers. To the best of our knowledge, there are only a few works on web application security visualization that are published until now [17-18].

Web applications have a complex structure with pages and links between them. To support human users in understanding this structure efficiently, visualization techniques are developed to display as much of it as possible while keeping the visible parts clear. However, before the visualization can take place, a modeling step is needed to transform web applications to structures more appropriate to be handled by computing systems. Intuitively, a directed graph is one of the best matched candidates to model a web application because of their similarities: pages are viewed as nodes, and links are viewed as edges. Then, layout algorithms are used to position nodes/edges with regard to some aesthetic criteria [19]. Some widely used algorithms to draw graphs are spring embedder layout [20], force-directed layout [21], and their variations. However, with a middle or large website, the number of pages and links can make its representative graph illegible for users to comprehend. So, more simple models are also used with some trade-offs. Among them, tree is one of the most popular structures used to visualize websites. There are two widely used approaches to draw trees: space-filling [22] and node-link diagram [23-24], each has its own advantages and disadvantages.

3 System Architecture

The system architecture of our proposed framework is depicted in Fig. 1.

In designing this framework, our expectation is that there is no limit on the number of vulnerability scanners that can be used. To realize this feature, we separate the scanning result of each tool from each other and assign an individual Result normalization component to each of the results. The task of the Result normalization components is to convert scanning results generated by their respective scanners into a common format that can be used later by the Statistics visualization component. However, because each scanner has its own unique data, types, and ranges, the normalization process is far from complete. Therefore, to make the framework support as many scanning tools as possible, we extract a minimum set of data that should exist in all scanners' results, such as: vulnerable URLs, severity levels, full requests and responses. For tools' specific data, we provide access to them by linking user interactions to respective tools' interface.

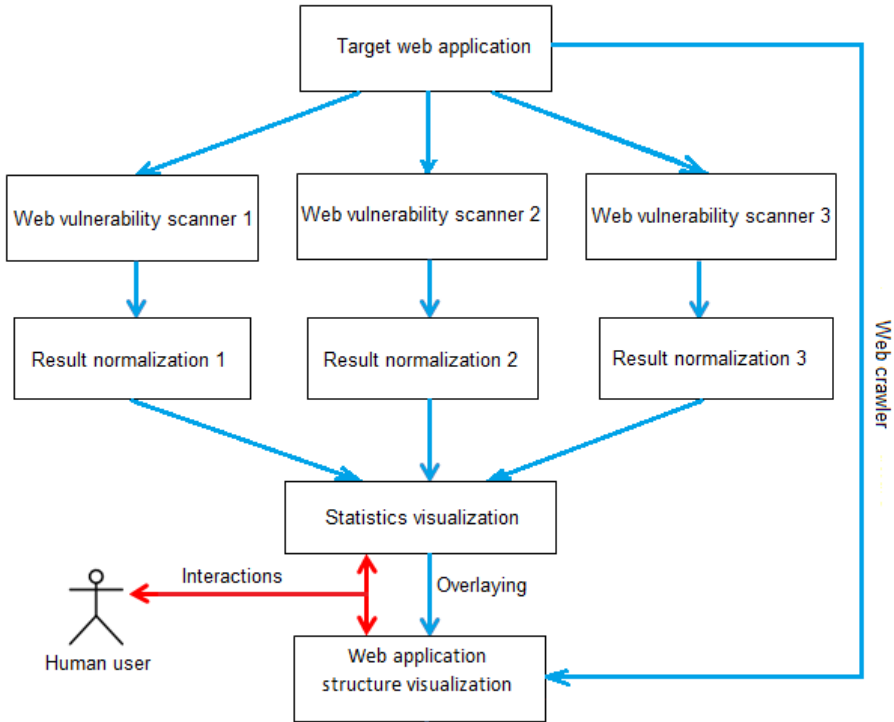


Fig. 1. Our proposed framework's system architecture

The Web application structure visualization component is responsible for presenting the target web application in an organized and easy to understand manner. In this work, we focus on the vulnerabilities discovered by scanners on pages, therefore we model target websites as rooted trees. This results to a more simple visualization and is appropriate to medium to large websites. Particularly, we use the Tree pattern of Polar plot [25] to emphasize the hierarchical nature of target web applications. The structural data this component uses for visualization is provided by a web crawler.

The task of the Statistics visualization component is to calculate basic statistics from raw result data obtained by scanning tools and overlay them on the drawing of target web application structure. More specifically, for each web page, we compute the total numbers of vulnerabilities discovered by each tool, grouped by their severity levels, and display these numbers as a stack chart on the respective web page node. Although simple, this statistics overlaying technique provides some analytical capabilities that are not easy to acquire by traditional tools: comparing the tools' effectiveness at various places; seeing the similarities and differences between nodes; and knowing the overall vulnerability distribution quickly.

Another component in this framework is the Interactions. Human users interact with this component to adjust the appearances of the Web application structure visualization and the Statistics visualization components. Usually, this component is used

for more focused analysis, happen after users having an overview picture and determining an interesting area. In this implementation, the framework supports basic interactions like node collapsing/expanding, filtering, zooming, and details-on-demand.

4 Visualization and Interaction Design

4.1 Web Application Structure Visualization

We model each target web application as a rooted tree with the home page as the root node. Each page with a unique URL is modeled as a tree node. Parent-child relationships between nodes are determined based on URL values of their respective pages. For example, a node with its page URL as “homepage/something” is the parent of a different node with its page URL as “homepage/something/child”, but not the parent of another node with its page URL as “homepage/somethingelse/child”. Two nodes with their page URLs as “homepage/child1” and “homepage/child2” are sibling nodes, and the common parent of them is the root node. From this construction, it is easy to see that each node will have no more than one parent, and the constructed graph is indeed a tree. Furthermore, the size of each node is decided by the number of vulnerabilities found on the related page of that node. We use node size to encode the number of vulnerabilities to make pages with many security issues stand out from the rest. We then apply a Polar plot-like on the constructed tree to create the needed visualization result. Compared to top-down layout, this radial layout has more room for the nodes. Fig. 2 depicts a sample website visualized by our method.

4.2 Statistics Visualization

For each page in the target web application, we extract the details of the vulnerabilities found and use it to create the Statistics visualization. In particular, we calculate the total numbers of vulnerabilities each scanner found and further divide these numbers into groups by common severity levels of vulnerabilities in those groups. These numbers are then used to draw stack charts with each color for each scanner’s result, and each transparent level for each severity level (Fig. 3, left). In this prototype version, we use three values: high, medium, and low to present severity levels. For scanning tools that output three severity levels, their results can be used directly. But for scanning tools that output a different number, a manual normalization rule is required at setup time. After all charts are built, they are overlaid on their respective tree nodes (Fig. 3, right).

4.3 User Interactions

We implement some basic user interactions to provide more focused analysis when needed. These interactions include: node collapsing/expanding, filtering, zooming, and details-on-demand. When a node collapses, its subtrees disappear and the vulnerabilities on these subtrees are totaled for the collapsed node. Vice versa, when a node

expands, its subtrees become visible with stack charts on them. While node collapsing is useful for summarizing data at a higher level, node expanding is appropriate for more detailed analysis. Different from node collapsing/expanding, which affects a selected part of the visualization, filtering affects the whole of it. Filtering is used to display interesting information only, for example, statistics of some particular scanners, or nodes whose vulnerabilities exceed a threshold. Zooming provides a simple mechanism to view an interesting region more clearly. And the remaining interaction, details-on-demand, allows users to view more specific information about the vulnerabilities as well as brings them to particular tools' interfaces when required. Fig. 4 demonstrates some interactions mentioned above.

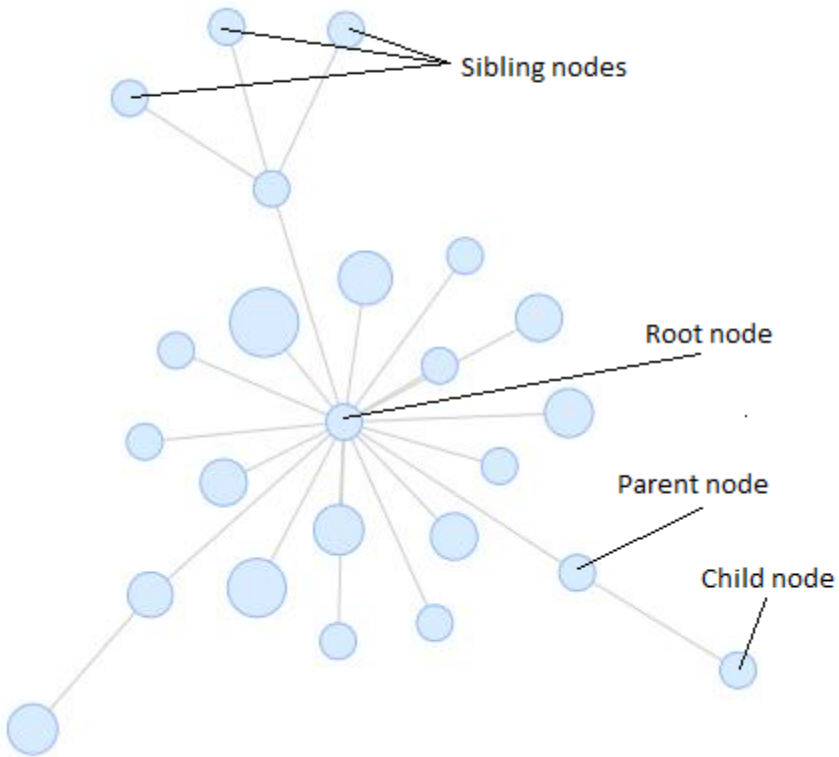


Fig. 2. Working of the Web application structure visualization component. The target web application is modeled as a rooted tree with the home page as the root node. Parent-child relationships are determined by URL values. Node sizes are decided by the number of vulnerabilities found on their respective pages.

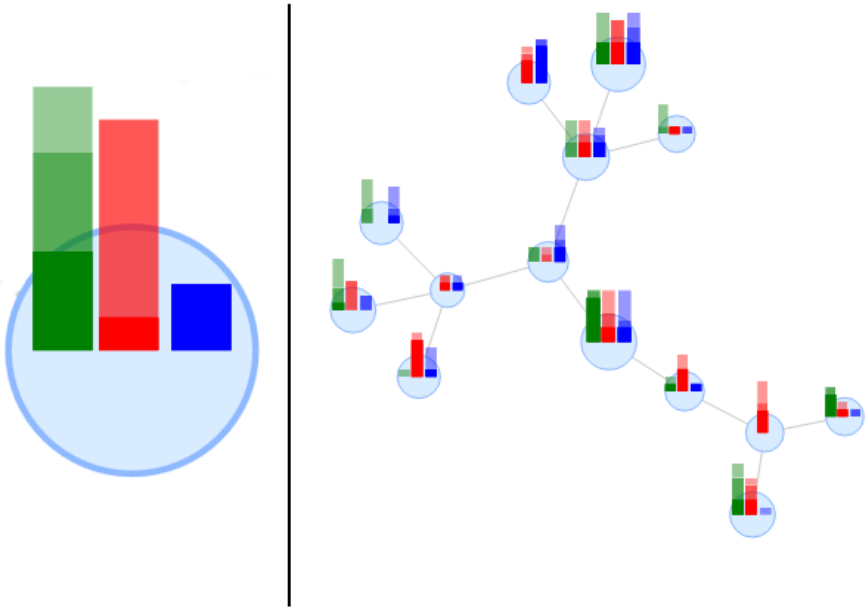


Fig. 3. The Statistics visualization component. The number of vulnerabilities are totaled for each scanner and grouped by their severity levels. On the left: stack charts are used to display the statistics; each scanner’s result has a unique color; and transparent levels are assigned to severity levels. On the right: the stack charts are overlaid on their respective nodes.



Fig. 4. Some interactions our framework supports. A: node collapsing/expanding. B: filtering user interface. C: details-on-demand in action.

5 Experiment

5.1 Implementation Details

We select three open source web application scanners for this framework prototype implementation: Arachni [26], w3af [27], and Wapiti [28]. There is no particular reason for the selection, except that they are free so we can start to work with them immediately. For each scanner, we create a related parser that can process that scanner's result into a common format, and then store them in one central database.

The visualization and user interaction are implemented on the web platform. Although web-based applications cannot offer as high performance as native applications do, we hope that the flexibility and popularity that the web offer can make our work more accessible. The library we use for the implementation is D3.js [29]. This library uses only W3C-compliant formats and languages like Scalable Vector Graphics (SVG), JavaScript, HTML5, and Cascading Style Sheets (CSS3), making it a good choice for our purpose.



Fig. 5. When the number of nodes is about 70, we cannot instantly grasp and understand the visualization result without some interactions. In this figure, it is not easy for us to decide which node is root or which node has the largest number of vulnerabilities by viewing only.

5.2 Testing the Framework

We create some virtual websites whose sizes increase from small to large to test the framework. Furthermore, we deliberately make some security holes on random pages in these sites. Our main goal in doing so is to create an artificial environment in which we can determine the framework's limit. As our experiment results point out, there are in fact two limits which we call technical-limit and user-limit.

Technical-limit is the limit of the visualization technique itself. It is the size of the target web application at which the performance of the framework decrease significantly. It is difficult for users to do analysis tasks at this limit, because of the lack of interactivity. Our prototype reaches this limit when the target website's number of pages is about 200. User-limit is the limit caused by users' cognitive ability. It is the size of the target web application at which the visualization is difficult for users to grasp and analyze without interaction. From our own experiments, this limit is much lower than the technical-limit, only about 70 pages. Fig. 5 demonstrates a case where the user-limit is reached for us.

6 Conclusion and Future Works

In this paper, we have described our proposal to solve two limitations of traditional web application vulnerability scanners: the first one is the lack of an effective interface to do dynamic analysis; and the second one is the difficulty in integrating scanning results of many tools together. To overcome the first problem, we created a novel visualization technique that can display scanning result statistics over the whole target web application. In addition, we apply basic interactions to provide more focused analysis when needed and to deal with medium to large websites. For the second problem, we design and implement a flexible framework that can integrate scanning results of many scanners together. We argue this work can enhance the effectiveness and efficiency of human users in security analysis task. At least, the proposed framework has three features support that belief: it allows users to compare performances among tools at various places; it helps them spot the similarities and differences between pages; and it shows them the overall vulnerability distribution in a novel way.

However, we need to do more real-world experiments to confirm the advantages of this work. As the experiment results point out, not only we need to test the framework's visualization technique, but more importantly, we also need to test it with regard to human cognitive perspective. Another essential, but not easy, task is getting feedbacks from actual web security penetration testers. These feedbacks are crucial in improving the practicality of the framework.

Acknowledgements. This research is funded by Ho Chi Minh City University of Technology (HCMUT) under grant number T-KHMT-2014-37. The authors would like to thank the D-STAR Lab members for their programming supports and constructive feedbacks. The authors would also like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper.

References

1. Symantec: 2014 Internet Security Threat Report, vol.19, http://www.symantec.com/security_response/publications/threatreport.jsp
2. Jovanovic, N., Kruegel, C., Kirda, E.: Precise Alias Analysis for Static Detection of Web Application Vulnerabilities. In: 2006 Workshop on Programming Languages and Analysis for Security, pp. 27–36. ACM, New York (2006)
3. Wassermann, G., Su, Z.: Sound and Precise Analysis of Web Applications for Injection Vulnerabilities. In: 2007 ACM SIGPLAN Conference on Programming Language Design and Implementation, pp. 32–41. ACM, New York (2007)
4. Wassermann, G., Su, Z.: Static Detection of Cross-Site Scripting Vulnerabilities. In: 30th International Conference on Software Engineering, pp. 171–180. ACM, New York (2008)
5. Rimsa, A., D'amorim, M., Pereira, F., Bigonha, R.: Efficient Static Checker for Tainted Variable Attacks. *Science of Computer Programming* 80, 91–105 (2014)
6. Huang, Y.-W., Huang, S.-K., Lin, T.-P., Tsai, C.-H.: Web Application Security Assessment by Fault Injection and Behavior Monitoring. In: 12th International Conference on World Wide Web, pp. 148–159. ACM, New York (2003)
7. Kals, S., Kirda, E., Kruegel, C., Jovanovic, N.: SecuBat: a Web Vulnerability Scanner. In: 15th International Conference on World Wide Web, pp. 247–256. ACM, New York (2006)
8. Balzarotti, D., Cova, M., Felmetsger, V., Jovanovic, N., Kirda, E., Kruegel, C., Vigna, G.: Saner: Composing Static and Dynamic Analysis to Validate Sanitization in Web Applications. In: 2008 IEEE Symposium on Security and Privacy, pp. 387–401. IEEE Computer Society, Washington (2008)
9. Doupé, A., Cavedon, L., Kruegel, C., Vigna, G.: Enemy of The State: a State-Aware Black-Box Web Vulnerability Scanner. In: 21st USENIX Conference on Security Symposium. USENIX Association, Berkeley (2012)
10. Visualization for Cyber Security, <http://www.vizsec.org/>
11. Paula, R., Ding, X., Dourish, P., Nies, K., Pillet, B., Redmiles, D., Ren, J., Rode, J., Filho, R.: In the Eye of the Beholder: a Visualization-Based Approach to Information System Security. *International Journal of Human-Computer Studies* 63, 5–24 (2005)
12. Leschke, T., Sherman, A.: Change-Link: a Digital Forensic Tool for Visualizing Changes to Directory Trees. In: 9th International Symposium on Visualization for Cyber Security, pp. 48–55. ACM, New York (2012)
13. Fischer, F., Mansmann, F., Keim, D.A., Pietzko, S., Waldvogel, M.: Large-Scale Network Monitoring for Visual Analysis of Attacks. In: Goodall, J.R., Conti, G., Ma, K.-L. (eds.) *VizSec 2008*. LNCS, vol. 5210, pp. 111–118. Springer, Heidelberg (2008)
14. Conti, G., Grizzard, J., Ahamad, M., Owen, H.: Visual Exploration of Malicious Network Objects Using Semantic Zoom, Interactive Encoding and Dynamic Queries. In: 2005 IEEE Workshops on Visualization for Computer Security, IEEE Computer Society, Washington (2005)
15. Abdullah, K., Lee, C., Conti, G., Copeland, J., Stasko, J.: IDS RainStorm: Visualizing IDS Alarms. In: 2005 IEEE Workshops on Visualization for Computer Security. IEEE Computer Society Press, Los Alamitos (2005)
16. Mansmann, F., Göbel, T., Cheswick, W.: Visual Analysis of Complex Firewall Configurations. In: 9th International Symposium on Visualization for Cyber Security, pp. 1–8. ACM, New York (2012)

17. Dang, TT., Dang, TK.: A Visual Model for Web Applications Security Monitoring. In: 2011 International Conference on Information Security and Intelligence Control, pp. 158-162. IEEE Computer Society, Washington (2011)
18. Dang, T.T., Dang, T.K.: Visualization of Web Form Submissions for Security Analysis. *International Journal of Web Information Systems* 9, 165–180 (2013)
19. Battista, G., Eades, P., Tamassia, R., Tollis, I.: *Graph Drawing: Algorithms for the Visualization of Graphs*. Prentice Hall PTR, Upper Saddle River (1998)
20. Kamada, T., Kawai, S.: An Algorithm for Drawing General Undirected Graphs. *Information Processing Letters* 31, 7–15 (1989)
21. Fruchterman, T., Reingold, E.: Graph Drawing by Force-Directed Placement. *Software: Practice and Experience* 21, 1129–1164 (1991)
22. Shneiderman, B.: Tree Visualization with Tree-Maps: 2-D Space-Filling Approach. *ACM Transactions on Graphics* 11, 92–99 (1992)
23. Munzner, T., Burchard, P.: Visualizing the Structure of the World Wide Web in 3D Hyperbolic Space. In: *First Symposium on Virtual Reality Modeling Language*, pp. 33–38. ACM, New York (1995)
24. Yee, K.-P., Fisher, D., Dhamija, R., Hearst, M.: Animated Exploration of Dynamic Graphs with Radial Layout. In: *2001 IEEE Symposium on Information Visualization*, pp. 43–50. IEEE Computer Society, Washington (2001)
25. Draper, G., Livnat, Y., Riesenfeld, R.: A Survey of Radial Methods for Information Visualization. *IEEE Transactions on Visualization and Computer Graphics* 15, 759–776 (2009)
26. Arachni, <http://www.arachni-scanner.com/>
27. w3af, <http://w3af.org/>
28. Wapiti, <http://wapiti.sourceforge.net/>
29. D3.js, <http://d3js.org/>

A Combination of Negative Selection Algorithm and Artificial Immune Network for Virus Detection

Vu Thanh Nguyen¹, Toan Tan Nguyen¹, Khang Trong Mai¹, Tuan Dinh Le²

¹ University of Information Technology, Vietnam National University,
HCM City, Vietnam

nguyenvt@uit.edu.vn, nguyentantoanuit@gmail.com

² Long An University of Economics and Industry, Long An province, Vietnam
le.tuan@daihoclongan.edu.vn

Abstract. This paper proposes a combined approach of Negative Selection Algorithm and Artificial Immune Network for virus detection. The approach contains the following stages: the first stage is data extraction and clustering. In the second stage, the negative selection algorithm is deployed to create the first generation of detectors. In the third stage, aiNet is used to improve detectors' coverage and enhance the ability of detecting unknown viruses. Finally, generated detectors are used to computing danger level of files and a classifier is used to train them and test performance of the system. The experimental results show that in the suitable conditions, the proposed approach can achieve reasonably high virus detection rate.

Keywords: Artificial Immune System (AIS), Negative Selection Algorithm (NSA), Artificial Immune Network (aiNet).

1 Introduction

Traditionally, two virus detection methods commonly used in Anti-Virus (AV) are data-based and behavior-based. As the development in virus programming techniques and the vast increase in the number of computer viruses, these methods are becoming less effective. Recently, many attempts employing knowledge from other field of studies are conducted. Researchers are looking to nature to solve human problems, especially computer viruses. Bio-inspired algorithms emerged as prospective models because of their ability to adapt naturally to the environment in which they applied.

Artificial Immune Systems (AIS) [1] are computational paradigms inspired by the biological immune system. This emerging area focuses on exploring and employing different immunological mechanisms such as negative selection theory and the idiotypic network theory to solve computational problems.

Negative Selection Algorithm (NSA) [2] was discussed which is based on the generation of T cells in the immune system. It has been shown to be efficient for anomaly detection problems in [3] [4]. In their approach, a set of detectors is generated by some randomized process that uses a collection of self (benign) as the input. Candidate detectors that match any of the self-samples are eliminated, whereas unmatched ones are kept for later detecting stage. Binary representation is traditional-

ly used in NSA, however, in some cases, this representation scheme need a large number of detectors to guarantee a good level of detection.

Artificial Immune Network (aiNet) [5] inspired by the idiotypic network theory [6] [7] suggested that the immune system, which is composed of a regulated network of molecules and cells, is not only capability of recognizing foreign particles, but also capability of recognizing and interacting with each other. aiNet consists of a set of nodes presenting antibodies and connection between them. These connections will be grown or pruned using a training algorithm. aiNet is a dynamic unsupervised learning method that has been used in clustering, data visualization, control, and optimization domains. The main drawbacks of aiNet are its high number of user-defined parameters and its high computational cost per iteration.

Aiming at building a light-weighted virus detection system, the approach proposed in the study is a combination of aiNet and NSA to get advantages of both models. NSA is a simple algorithm for generating good detectors and aiNet helps to improve the detection with a smaller detector needed.

2 Related Work

In 2009, Rui Chao and Ying Tan proposed a virus detection system (VDS) based on AIS [8]. In their model, a combination of NSA and Clonal Selection Algorithms [9] was proposed to generate detectors. They also presented a mechanism to determine the dangerous level of files based on these detectors. Their proposed VDS has a strong detection ability and good generalization performance.

In [3], they presented an improved negative selection algorithm by integrating a novel further training strategy into the training stage. The main purpose of further training is reducing self-samples to reduce computational cost in the testing stage as well as improving the self-region coverage. The experiment results showed that the proposed algorithm can achieve high detection rate and the low false alarm rate in most cases.

In [10], Artificial Immune Network Clustering approach is combined with Rough Set for Anomaly Intrusion Detection. While using Rough Set to get the most significant features of the dataset, aiNet is used to detect the novel attacks that have not been discovered in the training patterns.

In [11], a controllable and adaptable computer virus detection model based on immune system was proposed. Variable detector radiuses which can be set and adjusted automatically were used to improve the immune mechanism of the model.

3 The Proposed Approach

In our proposed approach, the first stage is data extraction and clustering. In the second stage, negative selection process is used to generate detectors. In the third stage aiNet is used to improve performance while reducing the number of detectors needed. After that, danger levels of files are computed and a classifier is used to training these danger levels.

3.1 Data Extraction and Clustering

In this stage, the virus gene set and the benign gene set are constructed by gene extraction from the virus files and benign files, respectively. Genes are bit-strings extracted from files by using a sliding window with a length L , L is chosen equal to 32. Two consecutive extracted strings will be overlapped $L/2$ with each other.

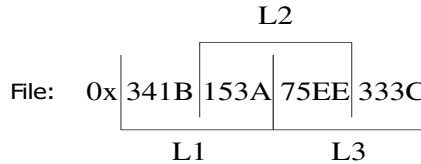


Fig. 1. Gene extraction mechanism

Generally, a virus requires a host program to infect. Note that similar genes may be existed in both virus gene set and benign gene set. In order to improve the accuracy of the detection, a gene in the virus gene set will be eliminated if the similarity between itself and a gene in the benign gene set is greater than or equal to the pre-defined matching threshold.

Note that the computational costs may be high because of the large number of genes extraction. To reduce the cost, in our approach the clustering algorithm Density-based spatial clustering of applications with noise (DBSCAN) [12] is deployed to reduce the size of the training data. After clustering, the distances from one to other elements are computed and get total distance in each cluster. In each cluster, the smallest total distance element will be kept as a delegate.

The main process of this stage is described as follow:

Notation

S_{VirusF} : training virus file set

$S_{BenignF}$: training benign file set

S_{virus} : virus gene set

S_{benign} : benign gene set

Extract(S): extract files of S to gene set

Cluster(G): cluster genes of G using DBSCAN and keep a delegate for each cluster

Begin

$S_{virus} := \text{Extract}(S_{VirusF}); S_{benign} := \text{Extract}(S_{BenignF})$

$S_{virus} := \text{Cluster}(S_{virus}); S_{benign} := \text{Cluster}(S_{benign})$

For each gene v in S_{virus} **do**

If v recognizes one or many genes in S_{benign} **then**
 v will be removed

End If

End For

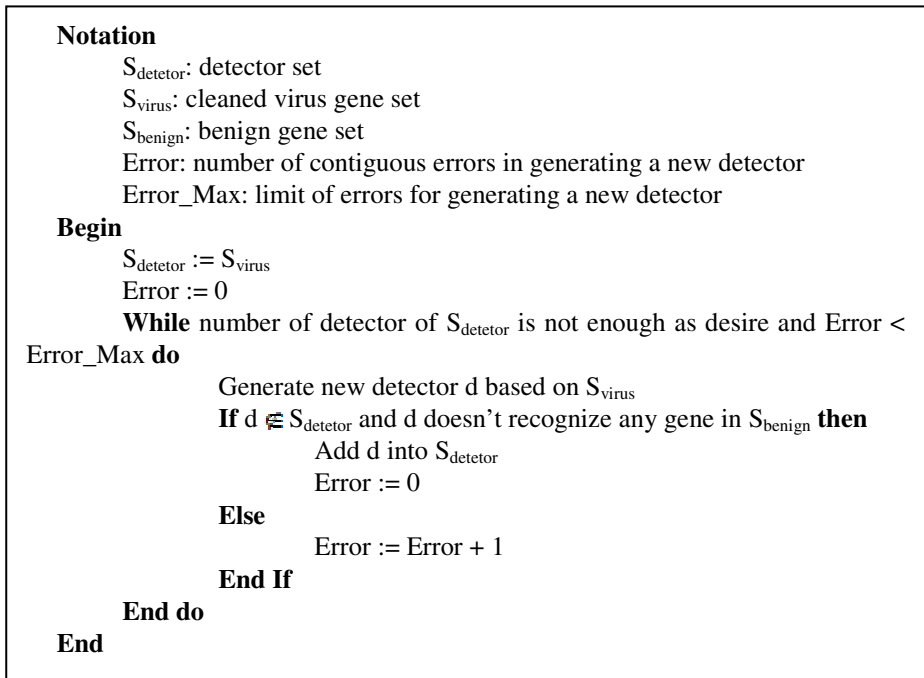
End

3.2 Negative Selection Process

After refining the training data, NSA is used to generate detectors that can discriminate between virus genes and benign genes. Besides using virus delegates as detectors, new detectors are generated from the delegates by randomly changing a bit string in the gene code bases on the pre-defined matching threshold. Some of these detectors can recognize not only virus genes but also benign genes.

Any of these detectors which can recognize a benign gene that is greater than or equal to the pre-defined matching threshold will be eliminated using NSA.

The process of negative selection is shown as follow:



3.3 AiNet Process

In the previous stage, a set of detectors generated can be used for virus detection. To increase the diversity of these detectors, a new generation of detectors is obtained during this stage with better coverage and ability to detect unknown viruses.

The aiNet algorithm is described as follow:

```

While not enough k iterations do
  S1.For each virus gene do
    S1.1.Choose m highest-affinity detectors with the current virus
    gene, clone these detectors proportion to their affinities and mutate the clones
    inversely proportional to their affinities.
    S1.2.Remove clonal detectors whose affinities with the current vi-
    rus gene are lower than the pre-defined matching threshold.
    S1.3.Suppress clonal detectors whose affinities with each other
    are greater than a pre-defined suppression threshold.
    S1.4.Select n desiring clonal detectors and save into a clonal de-
    tector set of the current virus gene.
  End For
  S2.Union clonal detector sets to remake the detector set.
  S3.Suppress detectors whose affinities with each other are greater
  than the pre-defined suppression threshold.
  S4.Remove detectors that can recognize some benign genes.
  S5.Add new random detectors to the detector set if this iteration is not
  the last iteration.
End While

```

In this algorithm, Step S1.1 and S1.2 are deployed to make the detectors become diversity and improve their qualities. Step S1.3, S1.4, S2, S3, and S4 is used to remove detectors that are similar to others and resizing the detector sets. Two suppression steps in this algorithm are S1.3 and S3. While S1.3 is used to suppress detectors in a group of detectors corresponding to the specific virus gene, S3 suppresses detectors between different groups of detectors corresponding to virus genes. Particularly, in suppression, if the affinity between two detectors is greater than the suppression threshold, in step S1.3, the one with smaller affinity to the current virus gene will be removed while in step S3, one of them will be removed randomly. Next, S4 works as an NSA to remove detectors recognizing benign genes. Last step S5 introduces new detectors to the detector set. The general diagram for generating detector set of the approach is shown in Fig. 2.

3.4 Affinity

Hamming distance h is used to identify the affinity between bit-strings. A threshold is used to determine if two strings are similar or not using the calculated affinity. Hamming distance is calculated as follow:

$$h(x, y) = \sum_{i=1}^N \overline{(X_i \oplus Y_i)} \quad (1)$$

where N is the length of strings, X_i and Y_i are the i^{th} bit of string x and the i^{th} bit of string y . $X_i \oplus Y_i$ denotes the XOR logic operation.

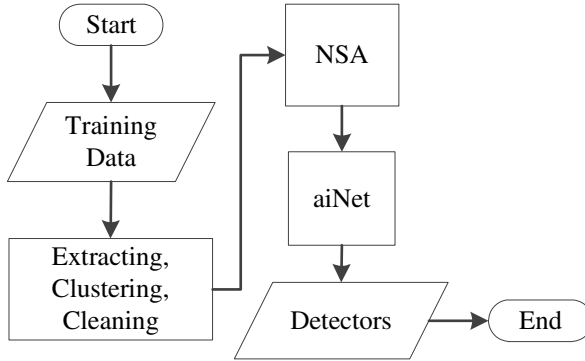


Fig. 2. The general model of generating detectors

Particularly, affinity *aff* will be identified based on hamming distance as follow:

$$aff(x, y) = N - h(x, y) \tag{2}$$

where *x* and *y* are two bit-strings with the same length and *N* is length of the bit-strings.

3.5 Danger Level

Mechanism to calculate strings or files danger level is similar to the one described in [8].

Danger level (*DL*) of a bit-string is used to determine how much dangerous the string is. And the danger level of a bit-string *x* of a file *l* is computed as follow:

$$DL(x) = \frac{\sum_{i=1}^{|DS|} \langle HA(x, DS), RCBA(x, DS, 12), RCBA(x, DS, 24) \rangle}{|DS|} \tag{3}$$

where *DS* is detector set, *x* denotes a bit-string extracted from file *L*, *HA*(*x*, *DS*) is average value of hamming distances between *x* and detectors in *DS*, *RCBA*(*x*, *DS*, *m*) is average value of results of R-Contiguous bits matching [2] between *x* and detectors in *DS* with the length of matching *m* bits. Each R-Contiguous bits matching will return one if two strings match or zero if they don't.

The danger level of file *L*, *DL*(*L*), is average value of all danger value of bit-strings extracted from file *L*.

$$DL(L) = \frac{\sum_{i=1}^{|L|} DL(x_i)}{|L|} \tag{4}$$

where *|L|* is the numbers of bit strings extracted from file *L* and *x_i* denotes a bit-string extracted from file *L*.

3.6 Training by Using Classifier

After the danger level of files in training data is calculated, a classifier, SVM [13] with a radial basis function (RBF) kernel [14] [15] is trained using these danger levels

to have ability for detecting virus files and then is used to detecting virus in the testing set.

The entire process in this study is shown in Fig. 3.

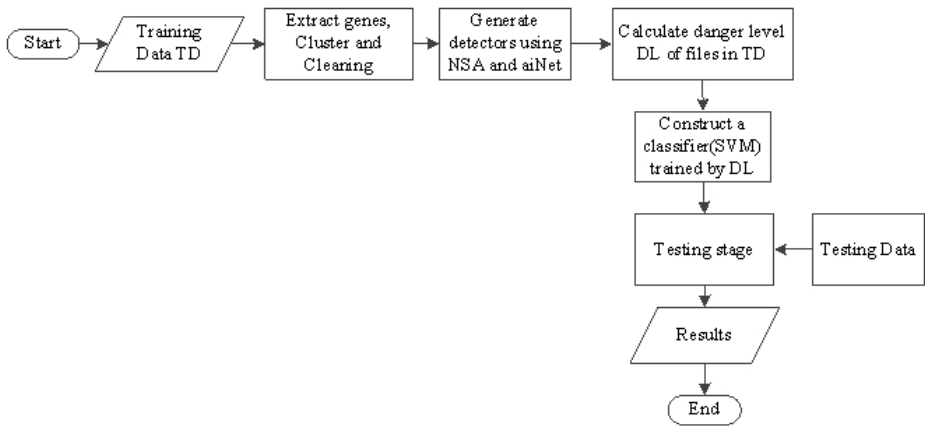


Fig. 3. The general model for virus detection system

4 Experimental Results

In order to evaluate our approach, a virus detection system was programmed in C# that allows user to define parameters, to observe the processes as well as to get experimental results.

4.1 Training and Testing Data

In this experiment, different datasets which include benign and virus files are generated. For each dataset, the ratio of 7 training set to 3 testing set is used, as shown in Table 1.

Table 1. The number of files in each dataset

Dataset	Training files		Testing files	
	Virus files	Benign files	Virus files	Benign files
Dataset 1	100	100	43	43
Dataset 2	200	150	86	64
Dataset 3	250	250	107	107
Dataset 4	400	250	171	107
Dataset 5	600	250	257	107
Dataset 6	800	300	343	129
Dataset 7	900	300	386	129

4.2 Experiments

Detection Rate. The first experiment was carried out to study the detection rate of our approach and the correlation between the performance of the model and the number of files in datasets. Note that the selection of matching thresholds is based on the condition that there were still remaining virus genes after benign genes had removed from virus gene set in cleaning virus gene set. The value of suppression threshold is manually selected so that it does not cause number of detectors to be significantly removed in the aiNet stage. This experiment has been done with the matching threshold is 84.375% and the suppression threshold is 90.625 %.

Table 2. The correlation between the number of files in data set and the performance of the model (the matching threshold is 84.375% and the suppression threshold is 90.625 %)

Dataset	Detection Rate (%)	
	Virus files	Benign files
Dataset 1	97.67	86.05
Dataset 2	98.84	73.44
Dataset 3	94.39	85.98
Dataset 4	98.25	91.59
Dataset 5	93.39	85.98
Dataset 6	93.29	91.47
Dataset 7	94.56	87.6
Average	95.77	86.02

As shown in Table 2, the average virus detection rate and benign detection rate in our experiment are 95.77% and 86.02%, respectively. The results showed that with suitable thresholds, our approach can achieve reasonably high virus detection rate.

The Correlation between Suppression Threshold and Number of Detectors. The main difference between aiNet and other AIS algorithms is at suppression stage, which mainly focuses on reducing the redundancy of similar detectors. The mechanism to determine the similarity of detectors is mainly based on the suppression threshold. The number of remaining detectors may affect the total coverage, which is consequent to performance of the system. In this experiment, we study how the suppression threshold affects to number of detectors and the performance.

Table 3. The correlation between the suppression threshold, the number of detectors generated by aiNet and detection rates (the dataset 3 with the matching threshold is 84.375%)

Suppression Threshold (% of length)	Number of detectors	Detection Rate (%)	
		Virus files	Benign files
90.625	2043	94.39	85.98
87.5	1466	97.2	82.24
84.375	659	95.33	78.5

As shown in Table 3, different suppression threshold's values result in changing in the number of detectors. If these detectors have approximately similar coverage, the performance will change insignificantly. This will help to reduce the cost of the system if we build a large system. Generally, the greater the suppression threshold is, the more overlap between detectors is. However, if the suppression threshold is small, the number of detectors may significantly decrease whereas the detectors are the important factors in the system which needs to be carefully taken into account. So choosing a reasonable suppression threshold is important to get good performance with a reasonable cost by decreasing the number of detectors needed while the total coverage is kept as the same approximately.

The Correlation between Matching Threshold and Number of Virus Genes. To determine the affinity between a gene and a gene or a gene and a detector, the hamming distance is introduced. In the cleaning virus gene set stage, if the affinity of a gene in unclean virus gene set and a gene in benign gene set are greater than or equal to the matching threshold, the two genes are considered as "similar" and then the virus gene is eliminated, otherwise. Because the number of remaining virus genes is critical to the our approach, the matching threshold is important factor that need to be considered in this experiment.

Table 4. The correlation between the matching threshold and the number of virus genes remained after removing benign genes from virus gene set (the dataset 3 is used)

Matching threshold (% of length)	Number of virus genes
84.375	2116
81.25	0

As shown in Table 4, the matching threshold affects the number of remaining virus genes after removing benign genes from virus gene set. The smaller the matching threshold is, the smaller the number of remaining virus genes is due to more of them were excluded when matching to genes in benign gene set. In contrast, if the matching threshold is large, it may be difficult to remove all genes that are similar to benign genes from virus gene set.

In overall, choosing a matching threshold should be considered as long as it doesn't meet a threshold that recognizes many virus genes as benign genes.

5 Conclusion

In this paper, an approach combining two bio-inspired models is proposed. While NSA is used for generating detectors, aiNet creates even better detectors for discriminate between virus genes and benign genes. Then, these detectors are used to calculate danger level of files and a classifier is trained to detect a virus file by its danger level. Our experiment results show that in suitable conditions, the proposed approach can achieve high virus detection rate. However, at the present the thresholds in our approach need to be determined manually. In the future, we will investigate how the system determines the best thresholds in order to achieve good performance and low cost.

References

1. Al-Enezi, J., Abbod, M., AI-Sharhan, S.: Artificial Immune Systems - Models, Algorithms and Applications. *International Journal of Research and Reviews in Applied Sciences* 3(2), 118–131 (2010)
2. Forrest, S., Perelson, A., Allen, L., Cherukuri, R.: Self-Nonself Discrimination in a Computer. In: *Research in Security and Privacy*, pp. 202–212. IEEE, Oakland (1994)
3. Gong, M., Zhang, J., Ma, J., Jiao, L.: An efficient negative selection algorithm with further training for anomaly detection. *Knowledge-Based Systems* 30, 185–191 (2012)
4. Sahu, A., Maharana, P.: Negative Selection Method for Virus Detection in a Cloud. *International Journal of Computer Science and Information Technologies* 4, 771–774 (2013)
5. de Castro, L., Von Zuben, F.: An evolutionary immune network for data clustering. In: *The IEEE SBRN (Brazilian Symposium on Artificial Neural Networks)*, pp. 84–89. Rio de Janeiro (2000)
6. Jerne, N., Cocteau, J.: Idiotypic Networks and Other Preconceived Ideas. *Immunological Reviews*, 5–24 (1984)
7. Jerne, N.: Towards a network theory of the immune system. *Ann Immunol (Paris)*, 373–389 (1974)
8. Chao, R., Tan, Y.: A Virus Detection System Based on Artificial Immune System. In: *International Conference on Computational Intelligence & Security*, vol. 1, pp. 6–10 (2009)
9. de Castro, L., Von Zuben, F.: Learning and optimization using the clonal selection principle. In: *Evolutionary Computation*, pp. 239–251. IEEE (2002)
10. Rassam, M., Maarof: Artificial Immune Network Clustering approach for Anomaly Intrusion Detection. *Journal of Advances in Information Technology* 3, 147–154 (2012)
11. Wang, X., Hua, J., Deng, Z.: A Controllable and Adaptable Computer Virus Detection Model. In: *Fifth International Joint Conference on INC, IMS and IDC*, Seoul, pp. 1977–1981 (2009)
12. Ester, M., Kriegel, H.-P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Simoudis, E., Han, J., Fayyad, U.M. (eds.) *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD 1996)*, pp. 226–231. AAAI Press (1996)
13. Cortes, C., Vapnik, V.: Support-Vector Networks. *Machine Learning* 20(3), 273–297 (1995)
14. Vert, J.-P., Tsuda, K., Schölkopf, B.: A primer on kernel methods. *Kernel Methods in Computational Biology*, 5–70 (2004)
15. Chang, Y.-W., Hsieh, C.-J., Chang, K.-W., Ringgaard, M., Lin, C.-J.: Training and testing low-degree polynomial data mappings via linear SVM. *J. Machine Learning Research* 11, 1471–1490 (2010)

De-anonymising Set-Generalised Transactions Based on Semantic Relationships

Hoang Ong and Jianhua Shao

School of Computer Science & Informatics
Cardiff University
{h.ong,j.shao}@cs.cardiff.ac.uk

Abstract. Transaction data are important to applications such as marketing analysis and medical studies. However, such data can contain personal information, thus must be sanitised before being used. One popular approach to protecting transaction data is set-based generalisation, where an item in a transaction is replaced by a set of items. In this paper, we study how well transaction data can be protected by this approach. More specifically, we propose de-anonymisation methods that aim to reconstruct original transaction data from its set-generalised version by analysing semantic relationship that exist among the items. Our experiments on both real and synthetic data show that set-based generalisation may not provide adequate protection for transaction data, and about 50% of the items added to the transactions during generalisation can be detected by our method with a precision greater than 80%.

1 Introduction

Transaction data are records that contain items about individuals. For example, Fig. 1 shows a set of 4 transactions, each recording a set of medical terms associated with a patient. TID is a transaction identifier which is included here for reference only; it will not be part of the released data.

Transaction data are important to applications such as marketing analysis and medical studies. However, such data can contain personal information, thus must be sanitised before being used. Unfortunately, simply removing identifying items such as names or telephone numbers is not sufficient to protect transactions, because combinations of other items in a transaction may still be used to identify an individual. For example, knowing that *ulcer* is associated with Mary will be enough to identify that Mary is the owner of T1 in Fig. 1, thereby revealing her identity within the dataset and disclosing other information about her.

One popular approach to protecting transaction data against this type of privacy disclosure is set-based generalisation, where an item in a transaction is replaced by a set of items. This is to ensure that combinations of certain items (which an adversary may use to attack the data) will not appear infrequently in the released dataset. For example, applying COAT (a set-based generalisation

TID	Items
1	gastric, ulcer, acid, bacteria
2	cancer, moles, bleeding, cough, bowels
3	diabetes, glucose, tiredness, itching, blurred vision
4	kidney disease, swelling, urination

Fig. 1. An Example of Transaction Data

TID	Items
1	gastric, (<i>ulcer, moles, glucose</i>), (<i>itching, acid, bleeding</i>), (<i>swelling, bacteria, tiredness</i>)
2	cancer, (<i>ulcer, moles, glucose</i>), (<i>itching, acid, bleeding</i>), cough, bowels
3	diabetes, (<i>ulcer, moles, glucose</i>), (<i>swelling, bacteria, tiredness</i>), (<i>itching, acid, bleeding</i>), blurred vision
4	kidney disease, (<i>swelling, bacteria, tiredness</i>), urination

Fig. 2. Set-Based Generalisation

method) [15] to Fig. 1 can result in Fig. 2¹, where items in brackets are *generalised* items. Now, knowing that Mary has ulcer will no longer be enough to determine if Mary is the owner of T1 in Fig. 2 with a probability greater than $1/3$.

Set-based generalisation is however a *syntactic* method. That is, it works on the assumption that items are meaningless literals, and how they form a set to replace (or generalise) an item is insignificant. The only requirement is that they should make some combinations of items appear frequently enough within the released dataset and they result in minimum distortion to the data. We argue that when semantic relationships among the items are considered, such protection may not be sufficient. For example, consider the generalised items in Fig. 2 again. Although (*ulcer, moles, glucose*) in T1 suggests that Mary could have *ulcer*, *moles*, *glucose* or any combination of them, the presence of *gastric* suggests that it is more likely to be *ulcer*, rather than *moles* or *glucose*. This type of semantic analysis will allow an adversary to reduce a generalised item to its original form.

In the paper, we study how well transaction data can be protected by set-based generalisation. More specifically,

- we propose de-anonymisation methods that aim to reconstruct original transaction data from its set-generalised version by analysing semantic relationships that exist among the items. This is in contrast to other studies on quantifying privacy risk involved in publishing transaction data [16,6,8] where attackers are assumed to have some auxiliary information about the individuals; we require no such information and rely on the released data only.

¹ In this example, we set $k = 3$, privacy constraints to be $\{(ulcer), (itching), (swelling)\}$, and utility constraints to be all the items. The reader is referred to [15] for details of how COAT works.

Thus, our de-anonymisation method is independent of the background knowledge that an attacker may have, and represents a realistic assessment of privacy risk associated with set-based generalisation.

- To determine semantic relationship among data items, we build our work on a measure called Normalised Google Distance [5]. This measure establishes semantic relationship between two terms by querying the Google repository of WWW pages: the more pages in which the two terms appear together, the more related they are considered to be. This eliminates the need to construct a single dictionary or corpus for testing term relationships and ensures that our approach is generic and realistic.

Our experiments on both real and synthetic data show that set-based generalisation may not provide adequate protection for transaction data, and about 50% of the items added to transactions during generalisation can be detected by our method with a precision greater than 80%. Note that our de-anonymisation approach uses information that is readily available from the released data and Google, thus the identified privacy risk is realistic.

The rest of the paper is organised as follows. Section 2 reviews the related work. Section 3 provides the necessary definitions. In Section 4, we introduce our de-anonymisation methods. We report our experimental results in Section 5. Finally, Section 6 concludes the paper.

2 Related Work

Protecting transaction data using set-based generalisation has been considered in [18,9,14]. Majority of these only aim to minimise the amount of data distortion during generalisation, that is, they seek to use a small set to generalise an item and to generalise as few items as possible. In so doing, they do not consider or constrain which items may appear in a generalising set, hence are vulnerable to the type of semantic attack we describe in this paper. COAT [15] on the other hand introduced utility constraints. These constraints limit what items can appear in a generalising set, thereby avoiding some incompatible items to be used together in generalisation. While this helps protect against semantic attack to a certain extent, utility constraints do not take other items in a transaction into account when generalising items. As a result, transactions sanitised by COAT are still vulnerable to semantic attack.

There is a body of work on the de-anonymisation of sanitised data [16,8,13]. The majority of these works focus on statistical data [12,10], where correlations and distributions are analysed to de-anonymise data that is perturbed through noise addition [1] or data swapping [4]. In contrast, we consider set-generalised data and rely on semantic relationships among the items, not statistic properties that may be observed from the released data. Narayanan and Shmatikov [16] proposed a method for identifying individuals from a set of published transactions. They assume that data is de-identified (i.e. having identifying items such as names removed), and that an attacker has some auxiliary information (i.e. knowing that several items are associated with an individual) when attacking

the data. We focus on set-generalised data, which is considered to provide better protection for privacy than simply de-identified data, and we do not require an attacker to possess any auxiliary information.

Anandan and Clifton studied how a sanitised term in a text may be re-identified based on its semantic relationships with others [2]. They assume that the term is generalised according to a specific taxonomy, and they measure its semantic relationships with others w.r.t. this taxonomy in order to re-identify it. In so doing, they rely on the existence of a taxonomy for semantic analysis, which is not entirely realistic in practice. Sánchez and Rovira [17], on the other hand, considered the possibility of uncovering a suppressed term from a sanitised text without using a taxonomy. Similar to our work, they use Normalised Google Distance to measure semantic relationships among the terms. While we share the view that semantic relationships among terms must be taken into account when anonymising data and the Normalised Google Distance is a practical way to analyse their relationships, our work has a different focus from theirs. Given a sensitive term, Sánchez and Rovira apply semantic analysis *before* anonymisation to assess whether it is sufficient to suppress this term only or other terms must also be suppressed in order to protect it. In contrast, we apply semantic analysis *after* anonymisation, i.e., attempting to reduce a set-generalised item to its original form.

3 Preliminaries

Let $\mathcal{I} = \{i_1, \dots, i_m\}$ be a finite set of literals called *items*. A *transaction* T over \mathcal{I} is a set of items $T = \langle a_1, a_2, \dots, a_k \rangle$, where each $a_j, 1 \leq j \leq k$ is a distinct item in \mathcal{I} . A transaction dataset $\mathcal{D} = \{T_1, \dots, T_n\}$ is a set of transactions over \mathcal{I} .

Definition 1 (Itemset and Support). *Any subset $I \subseteq \mathcal{I}$ is called an itemset. An itemset I is supported by transaction T if $I \subseteq T$. We use $\sigma(I, \mathcal{D})$ to represent the number of transactions in \mathcal{D} that support I , and we call these transactions supporting transactions of I in \mathcal{D} .*

For example, $\langle \text{gastric, ulcer, acid, bacteria} \rangle$ is a transaction in Fig. 1. $\langle \text{gastric, ulcer} \rangle$ is an itemset and is supported by T1, hence has the support of $\sigma(\langle \text{gastric, ulcer} \rangle, \mathcal{D}) = 1$. T1 is its supporting transaction.

When the support for an itemset is low, i.e. the itemset appeared infrequently within a transaction dataset, an attacker may use it to identify an individual with a high probability. A popular approach to ensuring that such itemsets would not compromise privacy is set-based generalisation [15], where some individual items are replaced by a set of items.

Definition 2 (Set-Based Generalization). *A set-based generalization is a partition $\tilde{\mathcal{I}}$ of \mathcal{I} in which each item $i \in \mathcal{I}$ is replaced by the partition to which it belongs. Each partition is called a generalized item, and each i is mapped to its generalised version \tilde{i} using a generalization function $\Phi : \mathcal{I} \rightarrow \tilde{\mathcal{I}}$. When an item is generalised to itself, we say that the item is trivially generalised.*

We denote a generalized item by listing its items in brackets, e.g. (*ulcer, moles, glucose*) in Fig. 2², and we interpret a generalised item as representing any non-empty subset of its member items, e.g. (*ulcer, moles, glucose*) may represent *ulcer* and *moles*. Generalization can help prevent identity disclosure as it increases the number of transactions in the dataset that may be linked to an individual through combination of items [15]. For example, consider the mapping of item *ulcer* in Fig. 1 to a generalized item (*ulcer, moles, glucose*) in Fig. 2. (*ulcer, moles, glucose*) is supported by 3 transactions in Fig. 2, whereas *ulcer* is supported by 1 transaction in Fig. 1.

Various privacy models have been proposed and they require different privacy constraints to be satisfied by the released data [18,15,19,7]. For the purpose of this paper, we use a simple, but commonly adopted privacy protection model based on support count.

Definition 3 (Protected Transactions). Let $\tilde{\mathcal{D}} = \{\tilde{T}_1, \tilde{T}_2, \dots, \tilde{T}_n, \}$ be a set of set-generalised transactions, and $c = (I, \sigma_{min})$ be a privacy constraint that requires an itemset I to have a minimum support of σ_{min} in \mathcal{D} . $\tilde{\mathcal{D}}$ is protected w.r.t. c if either $\sigma(I, \tilde{\mathcal{D}}) \geq \sigma_{min}$ or $\sigma(I, \tilde{\mathcal{D}}) = 0$.

Given a set of protected transactions w.r.t. a set of privacy constraints. we are interested to see if any constraint may be “violated” by performing some semantic analysis on the published (set-generalised) transaction data. That is, we are interested to know if some items in a generalised itemset could be removed based on their semantic relationships with other items in a transaction, thereby reducing the extent of generalisation and recovering some low frequency itemsets from the published transactions.

4 De-anonymisation Based on Semantic Relationships

In this section we describe our de-anonymisation methods. Given a generalised transaction $\tilde{T} = \langle \tilde{i}_1, \tilde{i}_2, \dots, \tilde{i}_k \rangle$, we examine if any item may be removed from $\tilde{i}_j, 1 \leq j \leq k$ based on semantic analysis. Our approach consists of two steps: *scoring* and *elimination*. We describe these two steps in detail in the next two sections.

4.1 Scoring

The scoring step is a process to establish semantic relationships among the items within a transaction. One approach to measuring how a given pair of items is related is to use an expert-specified ontology, such as the Wordnet [3]. Ontologies provide hierarchies of concepts and allow class inclusion or subsumption to be inferred, thus can help determine if the two items are related conceptually. However, such measures are not suitable for our purpose because they tend to

² For clarity, we will drop () when generalisation is trivial.

measure similarity rather than relatedness, and they do not take different contexts into account. For example, “string” and ”cord” may be deemed similar by an ontology when they are both taken to mean a thin rope, but it does not suggest if these two terms are likely to be used together, nor they actually represent similar concepts if we consider the use of the two terms in a programming context.

An alternative approach is to use a corpus of texts, and the relatedness of two items is judged by how they appear together within the corpus of texts [11]. This can help establish term relatedness based on how they are actually used in a context, rather than just if the terms are conceptually similar. However, this approach needs to construct a comprehensive and unbiased corpus for testing the usage of any terms, and this is not always feasible in practice.

In this paper, we follow the corpus based approach, but to avoid the need to construct a comprehensive corpus, we adopt the *Normalised Google Distance* (NGD)[5] measure which considers the entire world-wide-web as a corpus. Given two terms x and y , their semantic relatedness is established by

$$NGD(x, y) = \frac{\max(\log f(x), \log f(y)) - \log f(x, y)}{\log(N) - \min(\log f(x), \log f(y))} \quad (1)$$

where $f(x)$ denotes the number of Google pages containing x , $f(y)$ the number of pages containing y , $f(x, y)$ the number of pages containing both x and y , and N is the total number of pages Google has in its repository. The lower the NGD score is, the more closely the two terms are considered to be semantically related. For example, we have $NGD(\text{“paracetamol”}, \text{“HIV”}) > NGD(\text{“paracetamol”}, \text{“Cold”})$, which suggests that in general Paracetamol is more likely to be associated with Cold than with HIV.

For simplicity of discussion we assume that there is at least one trivially generalised item in a generalised transaction \tilde{T} .³ As trivially generalised items are original items in a transaction, we use them to identify any items in a generalised item that may not be in the original transaction by measuring their semantic relatedness. We call such items *context items*.

Given a generalised transaction, we may have a number of context items available, and any subset of these context items may be used to attack a given generalised item. Let C be a set of context items used in attacking an generalised item \tilde{i}_j and \hat{i} be an item in \tilde{i}_j . We measure semantic relatedness between \hat{i} and C by

$$d_{C, \hat{i}} = \frac{\sum_{j \in C} NGD(j, \hat{i})}{|C|}$$

where $|C|$ is the number of context items in C . That is, when multiple context items are used, an average score between \hat{i} and its context set C is used as a measure of how likely \hat{i} belongs to the transaction. For example, given

³ If no trivially generalised items available in a generalised transaction, then items in other generalised items may be used as context items. Detailed discussion on how this may be done is beyond the scope of this paper.

$\tilde{T} = \langle i_1, i_2, i_3, (i_7, i_8), i_4, i_5, i_6 \rangle$, the semantic relatedness between $i_7 \in (i_7, i_8)$ and its context $C = \{i_3, i_4\}$ is measured by $d_{C, i_7} = (NGD(i_3, i_7) + NGD(i_4, i_7))/2$.

One requirement of the set-based generalisation is that generalised items form k -equivalence groups. That is, each generalised item will appear at least k times within the released transactions. This is to ensure that the probability of using generalised items to link an individual to a transaction is no more than $1/k$. Therefore, when attacking a generalised item $\tilde{i} = (\hat{i}_1, \hat{i}_2, \dots, \hat{i}_s)$, we consider the whole equivalence group together by performing NGD scoring on each occurrence of \tilde{i} in different transactions and record the result in a distance table:

	\hat{i}_1	...	\hat{i}_s
C_1	d_{C_1, \hat{i}_1}	...	d_{C_1, \hat{i}_s}
...
C_k	d_{C_k, \hat{i}_1}	...	d_{C_k, \hat{i}_s}

Fig. 3. Distance Table

where columns are items in the generalised item under attack, and rows are context items selected from each transaction in the equivalence group to attack the generalised item. Note that while the generalised item \tilde{i} is identical in every transaction within the equivalence group, the context items that are selected and used to attack it need not to be the same. In fact, as each transaction is different and contexts are likely to be different, thereby allowing the membership of \hat{i} in \tilde{i} to be discriminated in a given transaction. For example, applying our scoring function to the generalised item (*ulcer, moles, glucose*) in Fig. 2, we obtain the following distance table:

This generalised item contains 3 items and forms a 3-equivalence group, therefore the distance tables has 3 columns and 3 rows. The largest distance is 1.98 between *moles* and *diabetes*, suggesting that they are not as related as others are, hence *moles* is likely to be an item introduced into T3 by the generalisation process, rather than an original item in T3. Note that in this example, we used a single context item to attack the generalised item. In general, any number of context items may be used if they are available.

4.2 Elimination

Once the semantic relatedness between the context items and the items in a generalised item is established, we employ some heuristics to eliminate those that are deemed to be not belonging to the original transactions from the generalised items. In the following sections, we give some heuristic methods to find such items.

Maximum Distance Attack. Given a distance table for an equivalence group of k transactions and an generalised item, it is easy to see from the definition of set-based generalisation that there exists at least one item that does not

belong to the original transactions. So a conservative method is to consider the one with the greatest distance in the distance table to be that item, and eliminate it from the generalised item. That is, we have

$$\mathcal{D}_e = \mathcal{D} \setminus MAX(\mathcal{D})$$

where $MAX(\mathcal{D})$ is the item with with greatest semantic distance in \mathcal{D} . For example, applying this method to Fig. 4, we eliminate *moles* from T3. However this method is very conservative, and does not attempt to eliminate all possible non-original items from a generalised item.

	ulcer	moles	glucose
$C_1 = \{gastric\}$	0.87	0.77	1.17
$C_2 = \{cancer\}$	1.02	1.45	0.85
$C_3 = \{diabetes\}$	1.11	1.98	0.73

Fig. 4. An Example Distance Table

Threshold Distance Attack. A more aggressive attack could consider all items with a distance above a certain threshold to be non-original, therefore eliminate them from the generalised item. That is, given a parameter δ and a distance table, we perform the following as long as d is not the last item left in a column or row in \mathcal{D} :

$$\mathcal{D}_e = \mathcal{D} \setminus \bigcup_{d \in \mathcal{D}, d > \delta} d$$

However, this attack requires the specification of a suitable δ , which may not be straightforward. One solution is to use the average distance in \mathcal{D} as δ :

$$\delta = \frac{\sum_{d \in \mathcal{D}} d}{|\mathcal{D}|}$$

where $|\mathcal{D}|$ is the number of items in \mathcal{D} . For example, the average threshold for Fig. 4 is 1.11, and eliminating items with a distance greater than this threshold from Fig. 4 we obtain Fig. 5. Note that while this method managed to eliminate more non-original items from generalised items, it has also eliminated a wrong one.

	ulcer	moles	glucose
$C_1 = \{gastric\}$	0.87	0.77	-
$C_2 = \{cancer\}$	1.02	-	0.85
$C_3 = \{diabetes\}$	1.11	-	0.73

Fig. 5. Result of Threshold Based Attack

Weighted Distance Attack. Intuitively, when an item is eliminated, it should have an affect on other items in a distance table. That is, removing one item should intuitively suggest that other items are more likely to be the original ones. Based on this intuition we propose a weighted distance attack which eliminates items from a distance table in iterations: one item is eliminated in each iteration, then the remaining distances in the table are updated w.r.t. the item eliminated. This continues until no more elimination can be performed.

Observe that each row or column in a distance table always contains at least one original item. So initially, if a row or column contains m items, then we assume that each item has a probability of $1/m$ to be an original one. As items are eliminated from the distance table, these probabilities will change and we use these probabilities as weights to revise the distances recorded in the distance table as follows:

Definition 4 (Weighted Distance). Let \mathcal{D} be a distance table and $\alpha_{ij} \in \mathcal{D}$ be the distance value at row i and column j in \mathcal{D} . The weighted distance α_{ij}^w for α_{ij} is calculated by

$$\alpha_{ij}^w = \alpha_{ij} \times \left(1 - \frac{1}{N_r - E_r^i}\right) \times \left(1 - \frac{1}{N_c - E_c^j}\right)$$

where N_r and N_c are the number of row and columns in \mathcal{D} , and E_r^i and E_c^j are the number of eliminated items in row i and column j , respectively.

That is, α_{ij} is first revised by the row probability $\left(\frac{1}{N_r - E_r^i}\right)$ and then by the column probability $\left(\frac{1}{N_c - E_c^j}\right)$. The more items are eliminated from a row (column), the more likely the remaining items in the row (column) will be original, and revision given in Definition 4 reflects that. In the following, we illustrate how this method works through an example.

Consider Fig. 4 again. To start, we assume that each item is equally likely to be an original item in each column and row, hence the two weights tables as shown in Fig. 6(a) and Fig. 6(b). The entries in Fig. 4 are then revised using these two weights tables according to Definition 4 to produce Fig. 6(c):

	ulcer	moles	glucose
C_1	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
C_2	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
C_3	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

(a) Row Weights

	ulcer	moles	glucose
C_1	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
C_2	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
C_3	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

(b) Column Weights

	ulcer	moles	glucose
C_1	0.39	0.34	0.52
C_2	0.45	0.65	0.38
C_3	0.49	0.88	0.32

(c) Weighted Table

Fig. 6. First Iteration

The elimination of an item from Fig. 6(c) is then carried out, based on the following conditions: a) the item has the greatest distance in the table, b) the item is not the last one in a row or column, and c) its distance is greater than the average distance in the table. Note that in this case, the average threshold is

calculated from the revised table, i.e. $\delta = 0.49$. *moles* in T3 satisfies these three conditions, hence is eliminated. After *moles* is removed, the two weights tables are updated and the results are shown in Fig. 7(a) and Fig. 7(b). These two tables are then used to revise Fig. 4 to give Fig. 7:

	ulcer	moles	glucose
C_1	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
C_2	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
C_3	$\frac{1}{2}$	-	$\frac{1}{2}$

(a) Row Weights

	ulcer	moles	glucose
C_1	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{1}{3}$
C_2	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{1}{3}$
C_3	$\frac{1}{3}$	-	$\frac{1}{3}$

(b) Column Weights

	ulcer	moles	glucose
C_1	0.39	0.26	0.52
C_2	0.45	0.48	0.38
C_3	0.37	-	0.24

(c) Weighted Table

Fig. 7. Second Iteration

Following the same process, *glucose* in T1 is eliminated, and weights updates and distance revision produce the following:

	ulcer	moles	glucose
C_1	$\frac{1}{2}$	$\frac{1}{2}$	-
C_2	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
C_3	$\frac{1}{2}$	-	$\frac{1}{2}$

(a) Row Weights

	ulcer	moles	glucose
C_1	$\frac{1}{2}$	$\frac{1}{2}$	-
C_2	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{1}{2}$
C_3	$\frac{1}{3}$	-	$\frac{1}{2}$

(b) Column Weights

	ulcer	moles	glucose
C_1	0.29	0.19	-
C_2	0.45	0.48	0.28
C_3	0.37	-	0.18

(c) Weighted Table

Fig. 8. Third Iteration

Now, as no more distances are above the threshold, the elimination process terminates. It is interesting to compare the result to that obtained from the Threshold Distance Attack: it did not wrongly eliminate *moles* from T2.

5 Experiments

In this section, we experimentally examine and compare the effectiveness of our proposed methods. We first describe the datasets used in our experiments, and then compare our methods to a baseline method which randomly determines if an item in a generalised item is original or not. We measure the effectiveness of our methods in terms recall and precision.

5.1 Datasets and Preparation

We used 3 datasets with different characteristics in the experiments. These transactions are extracted from i2b2 documents⁴, articles from GoArticles.com⁵)

⁴ i2b2 (www.i2b2.org) is sets of fully de-identified notes from the Research Patient Data Repository at Partners Health Care for a series of NLP Challenges organized by Dr. Ozlem Uzuner.

⁵ The data is collected in various topics from article directories

<http://www.goarticles.com>

Table 1. Datasets used in the experiments

Properties	AOL	i2b2	GoArticle
Size (transactions)	758	643	263
Length (items in a transaction)	1 to 5	150 to 200	150 to 200
Items require protection	127	112	45
Original/Generalised	1/6	1/6	1/3
Quality	Few typos	Many typos and abbreviations	Content cleansed
Context	Multiple	Single (Medical)	Multiple

and AOL search query logs⁶, and their characteristics are summarised in Table 1.

The AOL dataset is already in the form of transaction: each user’s search session is a transaction and each searching keyword is an item in the transaction. The i2b2 transactions are extracted from documents, using the *Stanford POS Tagger* package⁷. We extract nouns and noun-phases only from the documents. Each document results in one transaction and the nouns and noun phases extracted from the document become items in the transaction. For the GoArticle dataset, we selected a set of articles that share contexts and then manually extracted terms from the articles to form transactions. All our transactions do not contain duplicated items and the order in which the terms appear in a document is preserved in the extracted transaction.

After the transactions are prepared, we anonymise the data using COAT [15], a set-base generalisation method for anonymising transactions. Items to be protected (i.e. privacy constraints) are randomly selected from the transactions, and the number of items selected for each data set is given in Table 1. The utility constraint is the most general one, it allows any items to be used in any generalisation set.

5.2 Random Attack

We compare our proposed heuristics to a baseline method which performs a random attack on generalised items. The baseline method essentially assumes that an adversary has no other information than the released dataset, and he or she can only randomly guess whether an item in a generalised item is an original one or not. There are ways that an adversary may perform a random attack:

- On each item, the adversary decides whether it should be eliminated or not. Assuming that each item is equally like to be original or introduced by the generalisation, each item has a 50% chance to be eliminated.

⁶ Data source is found in <http://gregsadetsky.com/aol-data/>

⁷ <http://nlp.stanford.edu/software/tagger.shtml>

- The adversary randomly picks up a random number of items to eliminate from generalised items.

We use the second method in the experiment as it allows more variation in outcome and offers better comparison for our heuristics.

5.3 Results and Discussion

We use *precision* (p) and *recall* (r) to measure how well our methods can detect non-original items correctly:

$$r = \frac{\text{correct eliminations}}{\text{all non-original items}} \quad p = \frac{\text{correct eliminations}}{\text{all eliminations}}$$

We also use the standard F-score to measure their overall quality.

Fig. 9 shows the results of our experiments. We have not included the Maximum Distance Attack as it does not attempt to remove all non-original items, hence it is not meaningful to measure its recall and compare it to other methods. In attacking a generalised item \tilde{i} , we use two closest context items on either side of \tilde{i} . We measured recall, precision and F-score against k , the minimum size that a equivalence group must have in the released dataset. The higher the k is, the

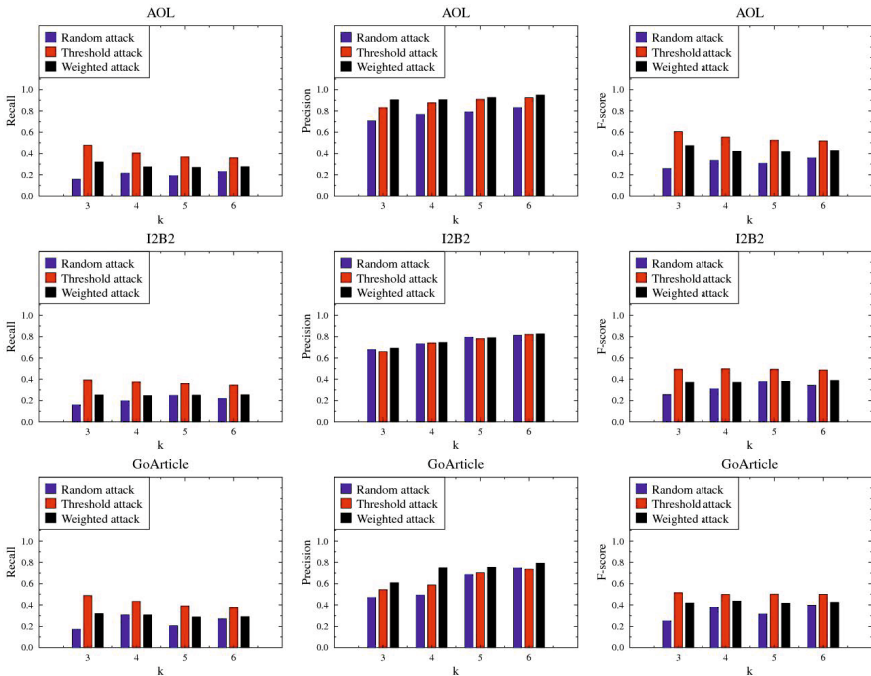


Fig. 9. Comparison of Random, Threshold and Weighted Attacks

more items are likely to be added into generalised items, hence more eliminations are to be expected.

It is easy to see that the Threshold and Weighted methods both outperformed the Random Attack. This suggests that semantic relationship among the items of transactions can be used to de-anonymise set-generalised transactions. This is especially so when we deal with transactions that are extracted from text, as such extracted items (nouns and noun phrases) are often related in some context, as we have observed in our experiments. For example, about 50% of the introduced items were removed from generalised items and precision of doing so was as high as 80%. So transaction data sanitisation without considering semantic relationships among their items may not provide sufficient protection for individual privacy.

It is worth noting that as k increased, recall actually decreased as can be observed in Fig. 9. We attribute this to the use of thresholds in both Threshold and Weighted Attacks. As k increases, more non-original items are likely or needed to be added into a generalised item in order to form required equivalence groups. As a result, the average distance between the items in a generalised item and its context items is likely to increase, as added items are likely to be less semantically related to the context items. This will result in a higher threshold and a lower recall. How to set a suitable threshold needs to be investigated further.

The Weighted Attack did not perform as well as the Threshold method in terms of the overall F-score in our experiments. This is a surprise, but we believe that this is mainly due to the characteristics of the datasets used in the experiments. For all three datasets, we observed that relatively largely number of items were added into generalised items, because the data were highly dimensional and sparse. This resulted in the NGD scores for the original items to be mostly below the average threshold. With the Threshold method, this gives a very good recall (and F-score), as all items above the threshold were removed. The Weighted method, on the other hand, is more conservative. Anytime an item is removed, it makes the rest more likely to be original. Consequently, it eliminates less, and has a lower recall and a higher precision. This is evident from Fig. 9. To verify this, we undertook further experiments to vary the thresholds used in elimination. The result is shown in Fig. 10.

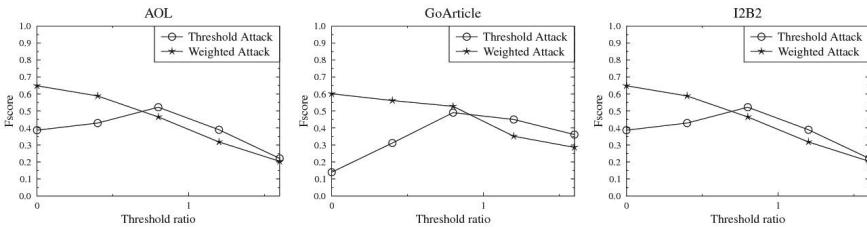


Fig. 10. Comparison of Effect of Threshold

As can be seen, when thresholds set very low (i.e. anticipating most of the items non-original), the Weighted method performed better. This is because as the thresholds lowered, more original items will have NGD scores that are above the threshold. They will therefore be wrongly removed by the Threshold method, significantly reducing precision and F-score. The Weighted method on the other hand is able to use the “enlarged” range to remove more non-original items while maintain relatively good precision due to its iterative process of elimination. This results in a better overall F-score. When the thresholds have increased to a point where it places most of the original scores below it, the threshold method works better. Again, how to find an appropriate threshold needs to be investigated further.

6 Conclusions

In this paper, we examined if set-based generalisation can provide sufficient protection for transactions. We proposed methods which use semantic relatedness among the items to detect if certain items are unlikely to be in a generalised transaction. We have shown that about 50% of the non-original items can be eliminated from generalised items with a precision greater than 80% in our experiments. This suggests that without considering semantic relationships, anonymising transactions using set-based generalisation may not provide protection for individual privacy. Furthermore, unlike other works, we do not assume any adversary background knowledge in attacking the data. The only information that an adversary needs in order to attack the data is the released data and Google repository. Thus, the risk we identified here is a real one.

References

1. Agrawal, R., Srikant, R.: Privacy-preserving data mining. In: SIGMOD 2000 Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, pp. 439–450 (2000)
2. Anandan, B., Clifton, C.: Significance of Term Relationships on Anonymization. In: 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, pp. 253–256. IEEE (2011)
3. Budanitsky, A., Hirst, G.: Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. *Evaluation* (1998)
4. Carlson, M.: A data-swapping technique for generating synthetic samples; A method for disclosure control (2000)
5. Cilibrasi, R.L., Vitányi, P.M.B.: The google similarity distance. In: *Knowledge and Data Engineering*, pp. 370–383 (2007)
6. Datta, A., Sharma, D., Sinha, A.: Provable de-anonymization of large datasets with sparse dimensions. In: Degano, P., Guttman, J.D. (eds.) *POST 2012. LNCS*, vol. 7215, pp. 229–248. Springer, Heidelberg (2012)
7. Ghinita, G., Tao, Y., Kalnis, P.: On the anonymization of sparse high-dimensional data. In: 2008 IEEE 24th International Conference on Data Engineering, pp. 715–724 (2008)

8. Giannella, C.R., Liu, K., Kargupta, H.: Breaching Euclidean distance-preserving data perturbation using few known inputs. *Data & Knowledge Engineering* (301) (2012)
9. He, Y., Naughton, J.F.: Anonymization of set-valued data via top-down, local generalization. *Proceedings of the VLDB Endowment* (2009)
10. Huang, Z., Du, W., Chen, B.: Deriving private information from randomized data. In: *SIGMOD 2005* (2005)
11. Islam, A., Inkpen, D.: Semantic text similarity using corpus-based word similarity and string similarity. *ACM Transactions on Knowledge Discovery from Data* 2(2), 1–25 (2008)
12. Kargupta, H., Datta, S., Wang, Q., Sivakumar, K.: On the privacy preserving properties of random data perturbation techniques. In: *Data Mining*, pp. 99–106 (2003)
13. Kifer, D.: Attacks on privacy and deFinetti’s theorem. In: *Proceedings of the 35th SIGMOD International Conference on Management of Data, SIGMOD 2009*, p. 127 (2009)
14. Liu, J., Wang, K.: Anonymizing transaction data by integrating suppression and generalization. In: *Advances in Knowledge Discovery and Data Mining*, vol. 1, pp. 1–10 (2010)
15. Loukides, G., Gkoulalas-Divanis, A., Malin, B.: COAT: COntstraint-based anonymization of transactions. *Knowledge and Information Systems* (2010)
16. Narayanan, A., Shmatikov, V.: Robust De-anonymization of Large Sparse Datasets. In: *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pp. 111–125 (May 2008)
17. Sánchez, D., Batet, M., Viejo, A.: Detecting Term Relationships to Improve Textual Document Sanitization. In: *PACIS 2013* (2013)
18. Terrovitis, M., Mamoulis, N., Kalnis, P.: Anonymity in unstructured data. In: *Very Large Data Bases (VLDB) Conference*, pp. 1–21 (2008)
19. Xu, Y., Fung, B.C.M., Wang, K.: Publishing sensitive transactions for itemset utility. In: *Eighth IEEE International Conference on Data Mining, ICDM 2008*, pp. 1109 – 1114 (2008)

An Implementation of a Unified Security, Trust and Privacy (STP) Framework for Future Integrated RFID System

Mohd Faizal Mubarak^{1,2}, Jamalul-Lail Ab Manan³,
and Saadiah Yahya¹

¹ Faculty of Computer and Mathematical Sciences,
Universiti Teknologi MARA Malaysia, 40450 Shah Alam, Selangor, Malaysia

² Information Communication Technologies (ICT) Division, MIMOS Bhd,
57000 Technology Park Malaysia, Kuala Lumpur, Malaysia

³ Strategic Advanced Research Cluster (STAR), MIMOS Bhd,
57000 Technology Park Malaysia, Kuala Lumpur, Malaysia
{faizal.mubarak, jamalul.lail}@mimos.my,
saadiah@tmsk.uitm.edu.my

Abstract. Previous and existing RFID systems with security protection have discussed the solution in term of security, trust and privacy in silo. Moreover, security and privacy solutions protect secret data and provide anonymity services for RFID system. Unfortunately, both solutions are not used to verify integrity of the integrated and interconnected platforms. RFID platform without any system integrity verification could be hijacked and adversary could use the infected platform to attack other integrated platforms. This research aims at designing RFID system with a combination approach of security, trust and privacy. The unified combination of security and privacy with trust in this solution compliments each other because integrity report is encrypted by using lightweight-based encryption, encryption key is sealed by using a sealing key of the trusted platform module, and identity of every platform is anonymous and protected by using an anonymizer. This paper presents an implementation of a framework which emulates the real hardware prototype system for RFID with security, trust and privacy.

1 Introduction

The positive and encouraging impact of RFID system is that it saves lots of times, which especially for tracking and managing massive items in large areas. For instance, it simplifies tasks such as inventory checking for the inventory worker to keep track of consumer products in the warehouse, librarian to manage books in the library and farmer to supervise livestock on the farm. However, this prevalence of RFID technology introduces emerging security and privacy risks because unprotected and unverified RFID system includes its components could easily be tracked or attacked by an adversary [1].

The normal conventional setup of RFID is a closed-based system which is not connected to the internet. However, the future technology of RFID system would eventually be integrated and interconnected via the network to other systems. For example, the current NFC system is a mobile phone that is capable of communicating with RFID reader [2, 3]. Another example is the implementation of UWB radio in RFID system that is the combination of RFID and wireless network [4, 5]. In the future, there are lots more possibilities for RFID system to be integrated with other systems such as sensors, electrical and electronic components, home appliances and which could also be connected to the internet. This would definitely expose RFID systems and components to more threats and attacks.

Things could be bad for the integrated RFID system because dangerous threats could come from the other side of the integrated components of the system. For instance, an adversary can invade the integrated system before it launches further attacks to RFID system by using malware [6, 7, 8]. A possible dangerous scenario could be, if the adversary is able to maliciously use the installed privacy-preserving or security components which are supposed to protect the entire system but ironically for it to become villain to the host system. This attack could be done by adversary through advanced hacking techniques and attacking tools which can be downloaded easily from the internet [9, 10]. History has proven that even the isolated and restricted SCADA system could be attacked by a novice hacker [11, 12]. This leads to a very important fact that system integrity verification is very crucial to be included in the RFID system. This is by virtue that it verifies every component inside the platform and provides integrity report for the external party inclusive of the privacy-preserving solution in the RFID system.

Security and privacy solution provides protection for RFID system against adversary attacks. Usually, RFID tag would be the target due to its mobility and transferability. Since RFID tag and reader communicates through radio frequency channel wirelessly, messages can also be captured by unknown reader. Thus, an important security objective of RFID is to ensure it is resistance against such attacks and also the secret data is protected from being intercepted by the adversary. Usually, encryption is one of the best approaches to protect the secret information in RFID system and especially in the communication channel. Normal type of encryption is not suitable to be used in RFID system because of resource limitation in the tag. The practical way is to encrypt the secret data in RFID system by using lightweight-based encryption [13, 14, 15]. The RFID system with encryption and other type of security techniques [16, 17] can only protect secret data from being exposed to the adversary, but it still can be traceable and tracked by unauthorized device.

Traceability and illegal tracking creates privacy issues in RFID system which could hijack user-related data including user identity and location. Usually, previous RFID protocols discussed about privacy and security, but not as one complete solution and everything is in silo. Almost all of currently used RFID systems do not provide privacy solution at all [18, 19]. Any compromised RFID system could expose location of the person who carried any item with affected embedded RFID tag. This scenario gives bad impressions about RFID to consumer because nobody wants their whereabouts to be tracked especially without their knowledge. The compromised RFID system also could expose secret data of any consumer product to adversary. This situation could give advantages to rival business entities because it can be further exploited

and exposed other details such as marketing and sales overview of that product. Supposed, every company does not want its trade secret information such as numbers of selling products to be exposed to anybody especially to competitors. Thus, the main objectives of privacy-preserving solution for RFID system is to protect the system against illegal tracking (location privacy) and it has to be untraceable (anonymous) to any unauthorized system. Similar to encryption, the privacy-preserving solution for RFID system has to fulfill the limited resources criteria in the tag. One good and suitable privacy preserving solution which can be used in low-cost RFID tag is anonymizer [20].

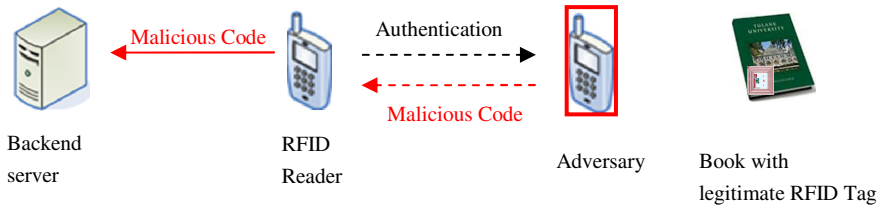


Fig. 1. Malicious code attacked on RFID system

Several previous RFID security research [20, 21, 22, 23, 24, 25] focused more towards security and privacy compared to trust or trusted computing. Security and privacy solutions provide secret data protection and privacy-preserving services for RFID system, but it would not verify the integrity of every component in the system. There is not so much discussion about trust and trusted computing in the research community of RFID security. An example of RFID protocol that is related to trust and integrity verification is provided by us, Mubarak et al. [26]. Other solutions [27, 28] are related to RFID reader hardware design and they are not discussed about protocol and system communication at all. RFID platform and protocol without any integrity verification cannot be trusted because it can be exposed to virus or malware attacks [7]. Moreover, it also can be hijacked by adversary to use the compromised platform as infected base or zombie to attack other platforms. Fig. 1 shows the authentication message from RFID reader could be intercepted by adversary system which injected malicious code to RFID reader and infected other system such as the backend server. RFID system with trust and integrity verification is very hard for malicious code or zombie attacks because integrity verification of every element and applications are being measured and verified by remote verifier. Any RFID system component with incorrect integrity measurement is detected to be rogue and communication with other systems have to be disconnected from further damage. Trusted based system could also detect any unexpected modifications at boot and runtime levels because compromised and unauthorized platforms are not having the same integrity measurements as the legitimate platform.

This research work provides a design of an integrated framework for security and privacy based RFID system with trust. The security part used for our proposed framework is a lightweight based encryption that protects secret information inside RFID system and in the communication channel. The privacy-preserving technique for

our proposed solution is by using anonymizer. Trust is another part of our proposed framework which provides integrity verification for RFID platforms by using the capabilities of the TPM.

The sequel of the paper is structured as follows:

- We describe several previous security and privacy based RFID solutions (Section 2).
- We briefly discuss about the system architecture and concept of RFID with STP (Section 3). Surprisingly, this has not been done before.
- We explain precisely our system implementation of RFID with STP (Section 4).
- Finally, we conclude the paper (Section 5).

2 Related Work

Back in 2003, research on security and privacy of RFID started after researchers from MIT [1] have reported that the unprotected information in the RFID wireless communication channel can easily be leaked to unauthorized reader and adversary. Since then, numerous previous related studies on security and privacy-based RFID solutions have been introduced by researchers to protect the RFID system. There are many security and privacy threats which always targeted RFID tags as victims [29], [30, 31] because the unprotected wireless communication in the contactless radio frequency electromagnetic field between RFID reader and tag is vulnerable to attacks.

The complexity of integrated and interconnected RFID system would create more security issues. For instance, unprotected wireless communication and mobility of RFID tag opens up many possibilities of its being tracked by adversary which violates user data and location privacy. Advanced technology such as NFC (Near Field Communication) type of mobile phone which is capable to communicate with RFID reader is also vulnerable to be tracked and traced by adversary system. RFID middleware that is connected to the network or internet makes RFID reader and backend server also vulnerable to attacks. Several RFID systems and protocols have been proposed by previous researchers in order to solve lots of security and privacy issues. Unfortunately, these RFID solutions could only solve specific issues such as security or privacy, but not a unified solution of security, privacy and trust [32].

An example of privacy-preserving solution for NFC type of mobile phone by using Direct Anonymous Attestation (DAA) technique has been proposed by Dietrich [2]. This solution is very good because it provides anonymity services for mobile host with NFC module which protects the privacy of mobile phone and its owner. Unfortunately, this DAA based solution is quite complex and heavy loads for anonymous authentication between mobile phone and NFC terminal. It took about eighteen steps to complete the authentication protocol between mobile phone and NFC terminal. The author is planning to implement a more efficient RFID protocol in the near future, so it could solve this performance issue. This solution only provides privacy-preserving solution and could be improved by including security protection for it.

Another example of privacy-preserving solution for RFID is the privacy-friendly RFID protocol that has been proposed by Asadpour & Dashti [33]. This solution is used to mitigate traceability issues in a substantial number of existing RFID protocols.

Authors of this solution have claimed that this protocol is anonymous and scalable because it used a pool of reusable anonymous tickets to protect identity of RFID tag. However, the communication process in this RFID protocol is done without proper authentication and integrity verification in place. Therefore, the reader would not know the integrity status of the tag and backend server due to this issue.

Another previous RFID protocol is the mutual authentication protocol for mobile RFID which has been proposed by Zhou *et al.* [34]. This solution is used to address the normal problem that occurs in RFID systems with authentication that needs RFID reader or backend server to search for secret keys exhaustively just to authenticate a single RFID tag. From our analysis and observation, we found that the mutual authentication process of this mobile RFID reader and tag is less efficient because it needs at least four steps just to be authenticated. This is quite longer process than usual that normally it takes for about one to two steps only. Another issue which is related to the protocol is regarding the connection between RFID reader and backend server. While it has been considered as unsecured-link between reader and backend server but there is not even a simple authentication process between the reader and backend server is provided by this protocol in order to protect both platforms.

RFID system with passive tag has resource limitation and it cannot do heavy processing such as public-key cryptography. Only the lightweight type of encryption is suitable to be used for security protection in this environment. This resource constraint is also applicable to other applications in the system such as privacy-preserving module. This situation is suitable for anonymizer because it provides anonymity services without using lots of resources from tag [20]. Sadeghi *et al.* [20] proposed the anonymizer-based RFID protocol that is very efficient and secure against impersonation attack. Unfortunately, this protocol has to be carefully considered because it requires an additional protocol between tags and anonymizers that is vulnerable to attack. This protocol also really needs lots of anonymizers to anonymize tag. Moreover, it's just assumed RFID system to always trust anonymizers. This protocol would create more problems if an anonymizer already has been infected by adversary and it could definitely make most RFID tags vulnerable to be attacked too and becoming traceable. The worse problem could be occurred as related to the adversary attack could also jeopardise the whole RFID system.

Several privacy-preserving protocols are really dependent to use lots of anonymizers because they have to make sure RFID tags remain anonymous mode all the time. This situation would definitely increase overall operation costs for the system implementation. Armnecht *et al.* [22] has proposed anonymous RFID protocol by using the enhanced version of the DAA protocol which has already been implemented by Chen *et al.* [35] in their DAA prototype. This protocol is very efficient because it used Elliptic Curve Cryptography (ECC) [13] type of DAA protocol rather than common types of DAA. Unfortunately, this protocol has the same problem with previous anonymous RFID protocol proposed by Sadeghi *et al.* [20] which is related to system availability because it always needs multiple anonymizers to anonymize each tag frequently.

From all of the above discussions on solving security and privacy issues in RFID protocols, there is no RFID protocol that provides system integrity verification in their scheme. Existing RFID protocols are able to resolve a number of security and privacy issues, but still unable to overcome other security and privacy related issues as reported by Henrici and Müller [36]. This is exasperated further by the non-existence of

integrity verifications in previous RFID protocols, which is important to create trust within the systems.

3 RFID System with Unified STP

The unified STP solution is proposed in order to provide holistic protection for RFID system. As far as our knowledge goes, almost all of previous and existing RFID protocols have discussed the solution in term of security and privacy in silo. Normally, security and privacy would always be the main focused for every RFID system and protocol, but not for trust or integrity which usually it is only being discussed as a part of security.

Our proposed RFID system with unified STP covers every system element from being attacked or traced by an adversary. Security solution such as lightweight-based encryption protects data in the communication channel between every platform in an RFID system. Trust or integrity verification process protects RFID platforms from being infected by malware or from being suffered through masquerade attacks. Privacy-preserving solution for RFID system is used to protect user identity and confidential data from being exposed or traceable by illegal entities. Location tracking by illegal party is another privacy issue of RFID system that is also can be protected by using our proposed solution. Our proposed RFID model is as shown in Fig. 2.

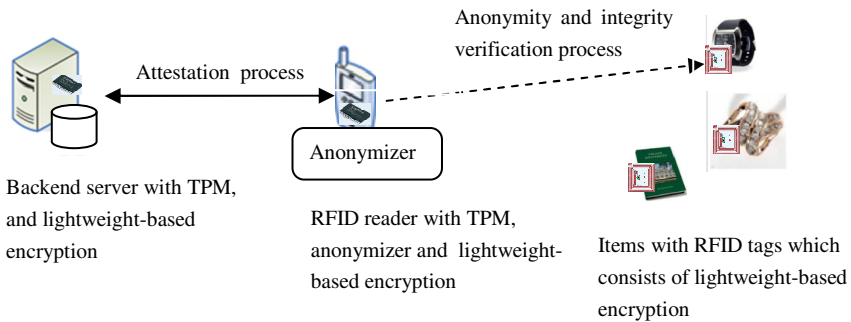


Fig. 2. A model of RFID system with unified STP

Our proposed RFID system with unified STP solution provides a complete security protection for data at the network level, storage level and also data at runtime. This security protection for our proposed RFID system is provided by using lightweight-based encryption technique, privacy-preserving protection by using an anonymizer to provide anonymity, and integrity verification process or trust by using the TPM and MJS-trust application. The TPM is the trusted computing component proposed by the Trusted Computing Group (TCG) [37] and MJS-trust is our trust-based of the in-house developed application.

This RFID with STP model consists of RFID reader embedded with TPM chip to perform attestation and integrity verification process by the backend server and

RFID tag. The challenge-response or attestation process occurs between the RFID reader and backend server is a mutual attestation process. Normally, a challenge-response process between RFID platforms is done through authentication. As we have already known, authentication and authorization mechanisms are very good security protection, but it's just for an element of access control only. Moreover, authentication and authorization processes could not provide protection against impersonation and malicious code attacks. For example, if the secret identity or password of authentication process fell off to the hand of unauthorized party may be through eavesdropping technique or man in the middle attack, it could easily be exposed and compromised to more dangerous threats and attacks such as malicious code and impersonation attacks. Therefore, attestation and integrity verification by using functionalities from TPM and MJS-trust provide protections against malicious code and impersonation attacks. Intruders that try to insert any alien code or malicious code to RFID platforms can be detected by MJS-trust. Attackers that try to impersonate any RFID platform by changing the secret information or system configuration of the platform can also be detected through the attestation process by using the TPM.

Our proposed framework also consists of anonymizer which placed in the RFID reader that is used to provide privacy-preserving solution such as creating anonymous data for RFID tags, backend server and also to the RFID reader. The communication channel in this proposed solution between RFID reader, backend server, and the tag is well secured by using the lightweight-based encryption of ECC [13], [38].

3.1 Anonymizer

In order to protect the RFID system from any kind of attacks which especially privacy-based attacks, RFID system must be equipped with the privacy-preserving solution such as an anonymizer. The privacy-preserving solution with an anonymizer is suitable to be used in RFID system because it provides protection for data and location privacy for RFID system and it is easily can be inserted into RFID tag because it does not need high power consumption [20]. It provides confidentiality, anonymity and unlinkability protections [20] for confidential data and location privacy [39, 40] of RFID reader, tags and backend server from being tracked or traced by an adversary. The low cost RFID tag has very limited resources, so it cannot be provided with algorithm or tools which needs high power consumptions.

Several previous anonymizer-based RFID protocols [20], [22] have almost similar issues, especially on providing an honest type of anonymizer. Most of them have other difficulties such as relying on multiple anonymizers to always refresh the tag. It must be emphasized that anonymizer without any integrity verification cannot be trusted because it could be hijacked by adversary and its component can be infected by malicious code and could be vulnerable to malware attacks [7]. Once the RFID system is hijacked by adversary, the system owner cannot confirm that it is operating as it is being expected by the system owner and to other parties [27]. This is more dangerous compared to the man in the middle attacks because the infected component could spread the virus to other components and could work together with the adversary to track and trace RFID system and its components.

System availability is another issue for previous anonymizer-based RFID protocols that can interrupt the process of RFID system to create anonymous tags. This issue

occurs because the anonymizer cannot perform its tasks normally, it could be due to it has been infected by malicious code or invaded by an unauthorized entity. It could be more dangerous because tags which are not anonymous can directly be exposed and tracked by adversary. These two most prominent previous research on anonymizer-based RFID solutions [20], [22] have tried to solve this issue by using multiple anonymizers. Unfortunately, it could produce others unexpected problems such as collision, logistics and system management issues. Moreover, multiple anonymizers are not cost effective as well and it is definitely would increase the maintenance costs.

System collision could occur between two or more anonymizers that are trying to provide anonymity services for the same tag. Another issue on system collision is about finding the right location for every anonymizer so that they would not compete with each other in providing anonymity services for tags. This kind of setup really depends on anonymizers to always providing anonymity services for tags regularly. The term “*regularly*” is related to unclear explanation because it is not really sure which time frequency is the best solution, i.e. whether the time is for every second, minutes or hours. If the anonymity services which provided by anonymizers for tags occurs too frequently, it will be good for the privacy of tags but it will consume lots of processing resources and it is very costly. On the other hand, the less frequent anonymity services for tags would give advantage for adversary to track tags.

System trust is really suitable to solve lots of issues which are related to previous anonymizer-based RFID protocols. System integrity verification for every component includes the anonymizer in RFID system guarantees that all of the components are operating as they have been expected by the verifier. Any intruder which tries to insert alien codes inside the trusted RFID system can easily be detected because it is not having the right integrity report. Moreover, system trust provides integrity verification for anonymizer in RFID system to be a trusted anonymizer-based RFID system which we proposed in our research work. More than that, this situation creates a complete trust and privacy-based RFID system.

3.2 Attestation

The message processing scenario of our proposed RFID with STP model would start by backend server verifying integrity of anonymizer and RFID reader through the attestation process. The integrity of anonymizer and RFID reader is represented by the integrity report that is created based on measurements of anonymizer and other components and applications inside the RFID reader platform. This integrity report is encrypted by using lightweight-based encryption of ECC. The backend server needs to decrypt the encrypted integrity report by using ECC encryption techniques based on pre-set key between the RFID reader and the backend server at the initialization stage. If the backend server found and verified integrity report from RFID reader to be an invalid report, it terminates the transaction with the RFID reader. Otherwise, if it found the integrity report from RFID reader to be a valid report, backend server accepted RFID reader as a legitimate platform and continue the transaction with the reader.

Then, the backend server responded to this challenge-response process by sending it's encrypted integrity report to be verified by the RFID reader. This integrity report has been encrypted by using ECC encryption based on pre-set key earlier. Both of this

challenge-response between RFID reader to backend server and from backend server to RFID reader is called mutual attestation process. These two platforms would accept each others as legitimate platforms after their decrypted and verified those integrity reports of each of the platforms as valid reports. Then, the backend server and the RFID reader are said to be trusted platforms which provide trusted communication services between both of the platforms.

After the mutual attestation process between the RFID reader and the backend server is completed, the RFID reader is open to communicate with any RFID tag because it has already been verified as a trusted platform. The communication process between the RFID reader and the RFID tag started by the reader interprets the radio waves to obtain an identity and information of any item which is embedded with the tag. At the same time, RFID reader also sends its encrypted integrity report to be verified by the RFID tag. RFID tag needs to decrypt the encrypted integrity report by using ECC encryption techniques based on pre-set key between the RFID reader and the tag. If the tag decrypted and verified the integrity report of the reader to be an invalid report, it stops from communicating with the illegitimate or compromised reader. Otherwise, if the integrity report of the reader is found to be a valid report, the tag continues to communicate with the reader by sending its anonymous ID (identity) to the reader. The identity of RFID tag has already been set and pre-installed to be anonymous by using the anonymizer in RFID reader.

The anonymous ID of RFID tag remains as anonymous, even if unauthorized RFID reader could get the anonymous ID from RFID tag, maybe from impersonation attack. The unauthorized RFID reader still could not decipher the message because it does not have anonymizer and the right secret keys to open the message. Although the authorized RFID reader could decipher the anonymous ID from RFID tag by using anonymizer, it still remains anonymous with RFID reader because RFID reader did not keep the real information of the tag. However, the authorized RFID reader knows and recognizes the tag as a legitimate platform. The identity of RFID tag is set to be anonymous so that only the trusted and authorized reader knows the real information about the tag and also to item that is related which later can be retrieved from database in the backend server.

4 System Implementation

This section provides the implementation of RFID with unified STP by using Rifidi emulator which is a tool in Rifidi Toolkit [41]. Rifidi emulator could easily emulate real hardware implementation of RFID readers and tags which taken from commonly be used the industrial-based RFID system. Normally, before RFID system developer embarks into real hardware-based implementation, they will provide a proof of concept (POC) or RFID design by using Rifidi development toolkit. In this research, our focus is to implement the complete STP framework without testing the system performance. We know that performance is very important but the complete STP framework is more important to be implemented and tested in the scope of security, trust and privacy. Therefore, this section provides the integration part of security element (ECC), integrity element (TPM and MJS-trust), and privacy-preserving element (anonymizer) in RFID with unified STP by using Rifidi emulator. First of all, Fig. 3 shows the attestation

process from the backend server to RFID reader. Fig. 4 shows the returned attestation process from RFID reader to the backend server. Both Fig. 3 and Fig. 4 show mutual attestation processes between RFID reader and backend server.

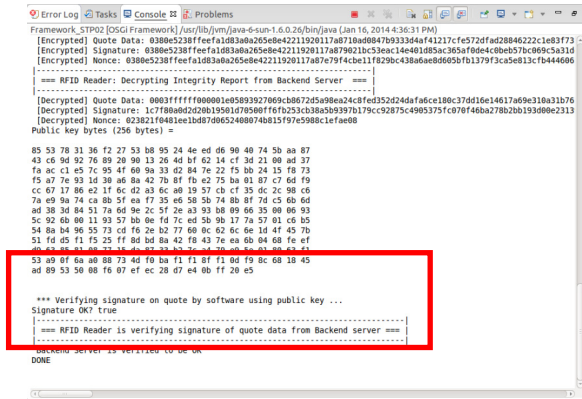


Fig. 3. RFID reader verifies integrity of the backend server

Fig. 3 shows backend server has encrypted its integrity report before the backend server sends it to be verified by the verifier (RFID reader). After RFID reader receives the integrity report of the backend server, RFID reader has to decrypt and verify the integrity report of the backend server. Fig. 3 shows the integrity report consists of quote data, signature and nonce. After that, the AIK from the backend server would be used by RFID reader to verify the signature and if it is found as a valid signature, the backend server would be recognized as the legitimate platform by the reader. The integrity verification process of the integrity report of the backend server by RFID reader is shown in the provided rectangles in Fig. 3. However, if the signature from the backend server found to be mismatched with the AIK, so RFID reader would recognize the backend server as illegitimate or unauthorized platform. The same process applies to RFID reader when it has to be verified by the backend server as shown in Fig. 4.

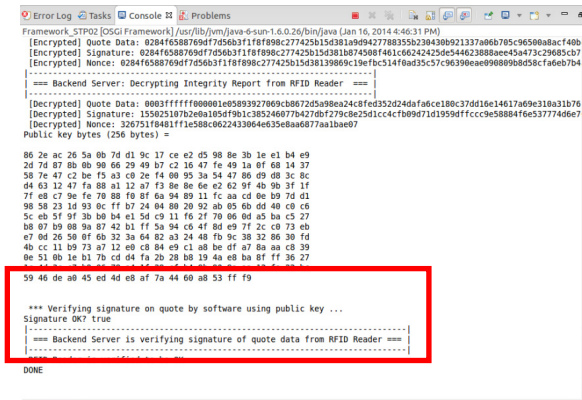
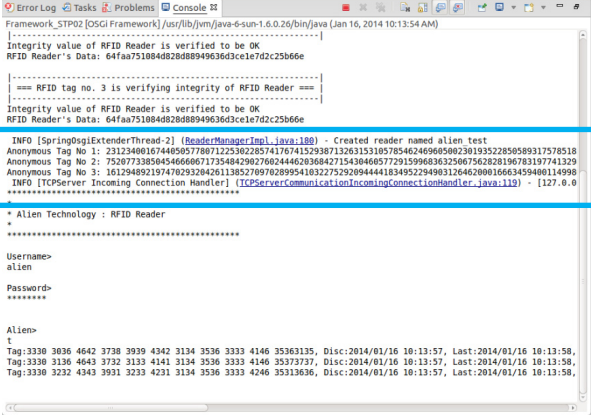


Fig. 4. Backend server verifies integrity of the reader

Moreover, the provided arrow shows the same values for the real tag IDs of the three RFID tags in Fig. 6 after it is listed by using “r” command in the reader.

After we have completed in building and constructing the RFID with STP it means that we have included security, trust and privacy-preserving modules in the RFID system.



```

Error Log Tasks Problems Console
Framework_STP02 [JSGI Framework] /usr/lib/jvm/java-6-sun-1.6.0.26/bin/java (Jan 16, 2014 10:13:54 AM)
[.....]
Integrity value of RFID Reader is verified to be OK
RFID Reader's Data: 64faa751084d828d8949636d3ce1e7d2c25b66e
[.....]
=== RFID tag no. 3 is verifying integrity of RFID Reader ===
[.....]
Integrity value of RFID Reader is verified to be OK
RFID Reader's Data: 64faa751084d828d8949636d3ce1e7d2c25b66e

INFO [SpringOsgiExtenderThread-2] (ReaderManagerImpl.java:188) - Created reader named alien test
Anonymous Tag No 1: 2312346016744695877897122530220574137674152938713283153105705424696650823019322858689317570510
Anonymous Tag No 2: 7529773385645466966717548429827692446283586427154384466577291599686325667562828196783197741329
Anonymous Tag No 3: 16129489219747829328426113852709782899541032275292094444183495229498312646280016663459400114998
INFO [TCPServer Incoming Connection Handler] (TCPServerCommunicationIncomingConnectionHandler.java:119) - [127.0.0.0
.....]
* Alien Technology : RFID Reader
* .....
Username>
alien
Password>
*****
Alien>
Tag:3330 3036 4642 3738 3939 4342 3134 3536 3333 4146 35363135, Disc:2014/01/16 10:13:57, Last:2014/01/16 10:13:58,
Tag:3330 3136 4643 3732 3133 4141 3134 3536 3333 4146 35373737, Disc:2014/01/16 10:13:57, Last:2014/01/16 10:13:58,
Tag:3330 3232 4343 3933 3233 4233 3134 3536 3333 4246 35333936, Disc:2014/01/16 10:13:57, Last:2014/01/16 10:13:58,

```

Fig. 6. The real tag IDs of RFID tag viewed by the legitimate RFID reader

5 Conclusion and Future Works

In this paper, we presented our proposed unified model for RFID with security, trust and privacy (STP). Our proposed model uses TPM as trusted computing principles and components to solve issues highlighted by previous works in RFID protocols. We combine the strengths of lightweight-based encryption, mutual attestation and privacy enhancement to form a unified model for STP of RFID system. This proposed model also provides a holistic protection for RFID system. We recommend future works that could be done for this research are to provide performance test and an integration with other components such as sensors in IoT (Internet of Things).

References

1. Sarma, S.E., Weis, S.A., Engels, D.W.: RFID systems and security and privacy implications. In: Kaliski Jr., B.S., Koç, Ç.K., Paar, C. (eds.) CHES 2002. LNCS, vol. 2523, pp. 454–469. Springer, Heidelberg (2003)
2. Dietrich, K.: Anonymous RFID authentication using trusted computing technologies. In: Ors Yalcin, S.B. (ed.) RFIDSec 2010. LNCS, vol. 6370, pp. 91–102. Springer, Heidelberg (2010)
3. Kumar, A., Arora, A., Islam, C.J.: Near Field Communication(NFC): An expertise primer. *Discovery* 2(4), 20–25 (2012)
4. Dardari, D., D’Errico, R., Roblin, C., Sibille, A., Win, M.Z.: Ultrawide bandwidth RFID: The next generation? *Proceedings of the IEEE* 98(9), 1570–1582 (2010)

5. Zou, Z., Baghaei-Nejad, M., Tenhunen, H., Zheng, L.R.: An efficient passive RFID system for ubiquitous identification and sensing using impulse UWB radio. *e & I Elektrotechnik und Informationstechnik* 124(11), 397–403 (2007)
6. Grunwald, L.: New attacks against RFID-systems. GmbH Germany (2006)
7. Rieback, M.R., Simpson, P.N., Crispo, B., Tanenbaum, A.S.: RFID Malware: design principles and examples. *Pervasive and Mobile Computing* 2(4), 405–426 (2006)
8. Shankarapani, M.K., Sulaiman, A., Mukkamala, S.: Fragmented malware through RFID and its defenses. *Journal in Computing Virology* 5(3), 187–198 (2009)
9. Armstrong, H.L., Forde, P.J.: Internet anonymity practice in computer crime. *Information Management and Computer Security* 11(5), 209–215 (2003)
10. Barber, R.: Hacking techniques: The tools that hackers use, and how they are evolving to become more sophisticated. *Computer Fraud & Security* 2001(3), 9–12 (2001)
11. Byres, E., Franz, M., Miller, D.: The use of attack trees in assessing vulnerabilities in SCADA systems. In: *Proceedings of the International Infrastructure Survivability Workshop* (2004)
12. Hentea, M.: A perspective on security risk management of SCADA control systems. In: *Proceedings of ISCA 23rd International Conference on Computers and their Applications, CATA 2008*, pp. 222–227 (2008)
13. Hein, D., Wolkerstorfer, J., Felber, N.: ECC is Ready for RFID – A Proof in Silicon. In: *Conference on RFID Security, Budapest, Hungary* (2008)
14. Hermans, J.: *Lightweight Public Key Cryptography* (Dortoral dissertation, PhD thesis, Katholieke Universiteit Leuven, Leuven, Belgium) (2012)
15. Feldhofer, M., Dominikus, S., Wolkerstorfer, J.: Strong Authentication for RFID Systems Using the AES Algorithm. In: Joye, M., Quisquater, J.-J. (eds.) *CHES 2004*. LNCS, vol. 3156, pp. 357–370. Springer, Heidelberg (2004)
16. Deng, G., Li, H., Zhang, Y., Wang, J.: Tree-LSHB+: An LPN-Based Lightweight Mutual Authentication RFID Protocol. *Wireless Personal Communications*, 1–16 (2013)
17. El Moustaine, E., Laurent, M.: GPS+: A back-end coupons identification for low-cost RFID. In: *Proceedings of the sixth ACM Conference on Security and Privacy in wireless and Mobile Networks*, pp. 73–78 (2013)
18. Atmel Corporation: Innovative IDIC solutions (2007), http://www.atmel.com/dyn/resources/prod_documents/doc4602.pdf
19. Calypso Networks Association: Web site of Calypso Networks Association (2007), <http://www.calypsonet-asso.org/>
20. Sadeghi, A.R., Visconti, I., Wachsmann, C.: Anonymizer-enabled Security and Privacy for RFID. In: Garay, J.A., Miyajiri, A., Otsuka, A. (eds.) *CANS 2009*. LNCS, vol. 5888, pp. 134–153. Springer, Heidelberg (2009)
21. Juels, A., Rivest, R., Szydlo, M.: The Blocker Tag: Selective Blocking of RFID Tags for Consumer Privacy. In: *Proceedings of the 10th ACM Conference on Computer and Communication Security*, pp. 103–111. ACM (2003)
22. Armknecht, F., Chen, L., Sadeghi, A.-R., Wachsmann, C.: Anonymous Authentication for RFID Systems. In: Ors Yalcin, S.B. (ed.) *RFIDSec 2010*. LNCS, vol. 6370, pp. 158–175. Springer, Heidelberg (2010)
23. Toirul, B., Lee, K.O.: An Advanced Mutual-Authentication Algorithm Using AES for RFID Systems. *International Journal of Computer Science and Network Security* 6(9B) (2006)
24. Fishkin, K.P., Roy, S., Jiang, B.: Some Methods for Privacy in RFID Communication. In: Castelluccia, C., Hartenstein, H., Paar, C., Westhoff, D. (eds.) *ESAS 2004*. LNCS, vol. 3313, pp. 42–53. Springer, Heidelberg (2005)

25. Peris-Lopez, P., Hernandez-Castro, J.C., Estevez-Tapiador, J.M., Ribagorda, A.: An Efficient Authentication Protocol for RFID Systems Resistant to Active Attacks. In: Denko, M.K., Shih, C.-s., Li, K.-C., Tsao, S.-L., Zeng, Q.-A., Park, S.H., Ko, Y.-B., Hung, S.-H., Park, J.-H. (eds.) EUC-WS 2007. LNCS, vol. 4809, pp. 781–794. Springer, Heidelberg (2007)
26. Mubarak, M.F., Manan, J.A., Yahya, S.: Mutual Attestation Using TPM for Trusted RFID Protocol. In: 2nd International Conference on Network Applications, Protocols and Services, NETAPPS 2010, pp. 153–158 (2010)
27. Soperra, A., Burbidge, T., Boekhuizen, V.: Trusted RFID Readers for Secure Multi-Party Services. In: EU RFID Forum (2007)
28. Molnar, D., Soppera, A., Wagner, D.: Privacy for RFID through Trusted Computing. In: Proceedings of the 2005 ACM Workshop on Privacy in The Electronic Society, pp. 31–34. ACM (2005)
29. Dimitriou, T.: A lightweight RFID protocol to protect against traceability and cloning attacks. In: Security and Privacy for Emerging Areas in Communications Networks, SecureComm 2005, pp. 59–66. IEEE (2005)
30. Juels, A.: RFID Security and Privacy: A Research Survey. *IEEE Journal on Selected Areas in Communications* 24(2), 381–394 (2006)
31. Piramuthu, S.: Protocols for RFID tag/reader authentication. *Decision Support Systems* 43(4), 897–914 (2007)
32. Mubarak, M.F., Manan, J.L.A., Yahya, S.: A Unified Model for Security, Trust and Privacy (STP) of RFID System. *Journal of Information Assurance and Security* 7(3), 119–126 (2012)
33. Asadpour, M., Dashti, M.T.: A Privacy-friendly RFID Protocol using Reusable Anonymous Tickets. In: 2011 IEEE 10th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), pp. 206–213 (2011)
34. Zhou, J., Zhou, Y., Xiao, F., Niu, X.: Mutual Authentication Protocol for Mobile RFID Systems. *Journal of Computational Information Systems* 8(8), 3261–3268 (2012)
35. Chen, L., Page, D., Smart, N.P.: On the Design and Implementation of an Efficient DAA Scheme. In: Gollmann, D., Lanet, J.-L., Iguchi-Cartigny, J. (eds.) CARDIS 2010. LNCS, vol. 6035, pp. 223–237. Springer, Heidelberg (2010)
36. Henrici, D., Müller, P.: Tackling Security and Privacy Issue in Radio Frequency Identification Devices. In: 2nd International Conference on Pervasive Computing, pp. 149–153. IEEE (2004)
37. Sumrall, N., Novoa, M.: Trusted Computing Group (TCG) and the TPM 1.2 Specification. *Intel Developer Forum* 32 (2003)
38. Batina, L., Lee, Y.K., Seys, S., Singelée, D., Verbauwhede, I.: Extending ECC-based RFID Authentication Protocols to Privacy-Preserving Multi-Party Grouping Proofs. *Personal and Ubiquitous Computing* 16(3), 323–335 (2012)
39. Kim, H.W., Lim, S.Y., Lee, H.J.: Symmetric encryption in RFID Authentication Protocol for Strong Location Privacy and Forward-Security. In: International Conference on Hybrid Information Technology, vol. 2, pp. 718–723 (2006)
40. Roberts, C.M.: Radio Frequency Identification (RFID). *Computer Security* 25(1), 18–26 (2006)
41. Heubner, A., Facchi, C., Janicke, H.: Rifidi Toolkit: Virtually for Testing RFID. In: The 7th International Conference on Systems and Networks Communications, pp. 1–6 (2012)

Toward a Nexus Model Supporting the Establishment of Business Process Crowdsourcing

Nguyen Hoang Thuan, Pedro Antunes, and David Johnstone

School of Information Management, Victoria University of Wellington,
PO Box 600, Wellington, New Zealand

{Thuan.Nguyen, Pedro.Antunes, David.Johnstone}@vuw.ac.nz

Abstract. Crowdsourcing is an emerging strategy that has attracted attention from organizations for harvesting information, labour, expertise and innovation. However, there is still a lack of a way to establish crowdsourcing as an organizational business process. Adopting a design science paradigm, the current study fills the gap by building a model supporting the establishment of business process crowdsourcing. In particular, we combined a structured literature review method, identifying individual components of business process crowdsourcing, and the design theory nexus, connecting these identified components. Our results identify twelve components that were widely proposed by the literature. These components are structured into a preliminary model concerning three stages of business process crowdsourcing: the decision to crowdsourcing, design, and configuration. Discussions on each component of the model and related implications are provided.

Keywords: Business process crowdsourcing, crowdsourcing, design science, nexus model, structured literature review.

1 Introduction

With the development of information technology that enables an online global workforce [1], many organizations have begun to shift from a strategy of inner-sourcing and outsourcing to a strategy of crowdsourcing. Crowdsourcing, which utilizes mass individuals in the crowd to perform specific tasks [2, 3], has attracted attention from the organizations for gaining information, skills, and labour, and reducing cost [4, 5]. Consequently, the list of organizations adopting a crowdsourcing strategy has become longer, including Threadless, iStockPhoto, Amazon, Boeing, Procter and Gamble, Colgate-Palmolive, Unilever, L’Oreal, Eli Lilly, Dell, Netflix, and Lexus [2, 5].

While early literature has demonstrated the success of several crowdsourcing initiatives, recent literature has emphasized that organizations need to build dedicated business processes to effectively utilize the crowdsourcing business model [6]. In crowdsourcing, although tasks are performed outside organizations, several other activities, such as task definition and quality control, remain inside [7]. Thus, it is necessary to establish crowdsourcing as an organizational business process, namely business process crowdsourcing (BPC) [8], which tightens and streamlines the external and internal activities.

This establishment has become more significant recently as crowdsourcing was used for complex organizational processes, such as product development [6].

Yet, in terms of establishing an approach to BPC, crowdsourcing has not been transferred from an emerging strategy to common practice. The current lack of a way to establish BPC has been identified by several researchers [9-11]. In particular, Vukovic et al. [11] state that one major challenge in the crowdsourcing domain is “how does crowdsourcing become an extension of the existing business process” (p. 7). Similarly, Khazankin et al. [9] recently noted the lack of a way to execute BPC, i.e. as repeated organisational practice. Consequently, the following research question needs to be further investigated.

Research Question: How to support the analysis, design, and configuration of business process crowdsourcing?

To address the research question, the current study aims to develop a model supporting the establishment of BPC. According to Aalst and Hee [12], a business process is defined as a number of tasks and a set of conditions determining the order to perform these tasks. Adopting this definition, the current study examines BPC as the overall coordination of internal tasks and crowdsourcing tasks. Effective coordination involves 1) classification of tasks across entities (e.g. between internal and external entities), where tasks corresponding to an entity comprise a sub-process, referring to a component in the ‘to-be-built’ model; and 2) integration of these sub-processes to execute the entire business process.

Although there are currently no frameworks for supporting the establishment of BPC, the literature has investigated these two aspects of BPC separately. In the first aspect, a large number of crowdsourcing studies have examined diverse topics within a particular crowdsourcing sub-process, including crowd management [13] and quality control [14]. However, these studies mostly focus on isolated aspects [15] and examine a crowdsourcing sub-process in an ad-hoc manner [16]. Addressing the second aspect, a few studies chose an integrated view and proposed several linked sub-processes or components of BPC [17, 18]. However, different studies suggest different lists of components, making it difficult to establish a common framework supporting the planning, analysis, designing and configuration of BPC.

Fulfilling this gap, the objective of this study is twofold. First, we want to identify and analyse what components constitute BPC. Second, we aim to integrate these components into a model supporting the establishment of BPC. To design this model, the current study followed a design science paradigm [19]. In particular, our research method combines a structured literature review method (SLR) [20] with the design theory nexus (DTN) [21]. While a SLR enables the formation of a knowledge base for developing a design science artifact [22, 23], the DTN can “connect numerous design theories with alternative solutions” (p. 1) [21] that result from the SLR. As a result, this combination helps to systematically identify and synthesize individual findings from the related literature into components comprising a BPC model.

By doing so, this study contributes to knowledge by consolidating our understanding on how to establish crowdsourcing as an organizational business process, addressing the current lack of a way to organize business processes based on crowdsourcing [9]. Another contribution of this study is to develop a model supporting the establishment of BPC. As this model is developed by incorporating the most significant findings highlighted in the BPC literature, it overcomes the ad-hoc manner emphasized by

the crowdsourcing literature [15, 16]. From a practical point of view, our research provides practical implication on how to analysis, design and deploy BPC, which moves forward the application of crowdsourcing in practice.

2 Background

2.1 Concept of Crowdsourcing

Since the term ‘crowdsourcing’ was first coined by Howe [2], referring to a strategy utilizing mass individuals to perform specific tasks in form of an open call, this concept was conceptualized by several researchers. Many of them conceptualized crowdsourcing by comparing this notion with similar concepts, including open innovation, outsourcing, and open source [4, 24, 25]. Within these concepts, crowdsourcing has often been classified to the open innovation paradigm, where organizations harvest knowledge and expertise from the outside, as opposite to closed innovation. However, Schenk and Guittard [24] stress two important differences between crowdsourcing and open innovation. The first one is that open innovation only focuses on innovative processes, while crowdsourcing can be used for varied types of tasks. Second, organizations explicitly interact with other firms and their customers in open innovation, but rely on members of the crowd in crowdsourcing activities [7].

Although organizational demands to use external agents are similar in crowdsourcing and outsourcing [2, 25], the differences between them can still be clearly identified. A major difference lies in the manner of who performs the activities. Actors performing tasks in crowdsourcing are members in the crowd, while they are supplier firms in outsourcing [24]. This leads to the second difference of managing these actors. Compared to official contracts with some preselected suppliers in outsourcing, crowdsourcing uses an open call to popularise the tasks [2, 7]. Finally, motivation for task performers in crowdsourcing is based on not only financial incentives as in outsourcing but diversity, including both intrinsic (e.g. love of community) and extrinsic motivation (e.g. financial incentives) [26].

It is also necessary to distinguish crowdsourcing from open source. Although both concepts rely on the power of the community to accomplish tasks, Brabham [4] suggests distinguishing these two concepts in terms of how the activities can be managed and performed. In crowdsourcing, organizations manage their workflows, whereas in open source, these activities are driven by the community. Examining how activities are performed, Zhao and Zhu [7] note that crowdsourcing outcomes can be achieved either independently or collaboratively, but open source’s outcomes can only be achieved through collaboration. Furthermore, unlike open source, crowdsourcing has clearer ownership and does not restrict to software [24]. Given the above discussion, it can be stated that crowdsourcing is a distinctive notion, and thus the current study investigates crowdsourcing as a concept *per se*.

2.2 Business Process Crowdsourcing

The term Business Process Crowdsourcing (BPC) was first introduced by Vecchia and Cisternino [8] as an alternative to business process outsourcing. Etymologically, BPC combines the word *crowdsourcing*, utilizing the crowd to perform particular jobs

(Section 2.1), with the phrase *business process*, referring to a number of tasks and the coordination of these tasks [12]. Thus, BPC should be examined as both a number of individual tasks across crowdsourcing entities, and the coordination of these tasks forming an entire business process.

The literature has highlighted several roles of BPC in crowdsourcing activities. First, BPC can help streamline internal and external tasks in the crowdsourcing process. In other words, the lack of an integrated workflow to link these tasks is an obstacle for crowdsourcing applications [9]. Second, BPC can preserve the knowledge necessary to accomplish several crowdsourcing tasks, like problem solving. Lopez et al. [10] state that “organizations require integration of the crowdsourced tasks with the rest of the business process. [...] the solutions are never reintegrated to the enterprise causing knowledge to be lost” (p. 539). Finally, an establishment of BPC enables crowdsourcing to become a common organizational practice, as opposed to one-off projects.

In spite of its promise, how to establish BPC has not been fully examined in the literature. Khazankin et al. [9] identify “the lack of an integrated way to execute business processes based on a crowdsourcing [platform]” (p. 1). Yet, these authors investigated only a part of the problem, which optimized task properties for supporting business process execution. Similarly, Satzger et al. [13] seek to help organizations “fully automate[d] deployment of their tasks to a crowd, just as in common business process models” (p. 67), but focus only on choosing suitable workers to perform tasks. As a result, the establishment of BPC still needs to be further investigated.

As previously mentioned, an investigation on how to establish BPC needs to consider both individual tasks of a crowdsourcing process and the integration of these tasks. Each of these two aspects has been explored separately in the crowdsourcing literature, but not in concert. In the first aspect, a large number of studies examine diverse topics of crowdsourcing tasks within a particular sub-process [13, 14, 27]. Though these studies provide several implications for establishing BPC, the overall picture is still unveiled due to their ad-hoc foci [15, 16]. As a result, various and disparate, sometimes conflicted, findings and sub-processes related to BPC exist in the literature, confusing organizations in their BPC establishment.

In the other aspect, although the integration of crowdsourcing process has featured in several studies, a comprehensive approach is still missing. For instance, Geiger et al. [17] propose crowdsourcing processes as a sequence of four components: preselection of contributors, accessibility of peer contributions, aggregation of contributions, and remuneration for contributions. As the names imply, these components, however are mainly related to the contributors or external processes, and thus do not clarify internal organisational sub-processes. Examining both internal and external processes, Hetmank [18] suggests other components of a crowdsourcing system, including user management, task management, contribution management, and workflow management. Although this study considers both internal and external components, it has quite a narrow view due to its chosen technical perspective [18].

In summary, while recent literature has emphasised the importance of standardising BPC, the diversity of perspectives around BPC have made this difficult. Additionally, these multiple perspectives across different disciplines have led to inconsistent findings and propositions, making it more difficult for the establishment of crowdsourcing practices. Addressing this gap, the current study aims to build a model supporting the establishment of BPC.

3 Method

To develop a model supporting the establishment of BPC, the current study followed a design science paradigm proposed by Hevner et al. [19]. In the design science paradigm, studies usually require a design method that guides the development of the artifact [19, 21] and a knowledge base that forms a background for the development [19, 28]. Although several design methods were proposed [19, 21, 28], the choice of using a particular method appears disparate in the existing literature, and seems to depend on the particular problems. In this study, as the establishment of BPC forms a wicked problem, in which a variety of views, sub-processes, issues and alternative solutions exist, we adapted a DTN proposed by Pries-Heje and Baskerville [21]. The DTN enables numerous design theories and different views to be connected [21], and therefore seems well-suited to consolidate the various views and individual foci existing in the crowdsourcing literature.

In addition, a design science study requires a suitable knowledge base, which can be populated with the research problem [19, 28]. However, crowdsourcing is a new research field [7], leading to the difficulty of finding a corresponding knowledge base for the establishment of BPC. This problem is not rare in design science [19]. Addressing the problem of non-existent knowledge base, several researchers suggest utilizing the best research evidence from the literature [22, 23]. It is worth noting that Pries-Heje and Baskerville [21], when proposing their DTN, also recommend “a survey of existing literature and findings” (p. 737-738) to identify the existing theories and solutions related to the targeted problem. Given the above discussion, the current study adapted and combined the DTN [21] and the SLR method [20].

Table 1 compares the stages of the current study with the equivalent stages of the DTN and SLR. As seen via Table 1, our method includes five stages: selecting articles, filtering articles, data extraction and classifying articles, data synthesis, and model building. These stages are based on and thus comparable to the SLR [20]. While based on the SLR, each stage in our method has a similar purpose compared to the steps of DTN that were summarized in column 3 of Table 1. In particular, our first three stages aim at identifying the literature related to BPC and extracting findings, approaches and applied conditions, consistent to the first two steps of DTN [21]. In the next stage, the extracted findings and conditions are synthesized and formalised into components. The final stage in our study follows the DTN (the last row of Table 1) by first designing a decision making process and then structuring the identified components into the decision process in order to develop a model supporting BPC. Detailed stages of our research method are presented in the following sections.

Selecting Articles. This stage involved the search for relevant articles addressing crowdsourcing subjects. Following a concept-centric approach that was not restricted but open to multiple sources of literature [29], we conducted keyword searches across eight popular online bibliographic databases, including ACM, EcoHost, IEEE, Emerald, Sage, ScienceDirect, Springer Link, and Wiley, between September and November 2013. The searched keywords included ‘crowdsource’, ‘crowdsourcing’, ‘crowdsourced’, ‘crowdsourcer’, and ‘crowdsources’. Additional criteria for selection were that articles have been written in English and available in full text. As a result, the selecting stage identified a total of 877 articles (Table 2).

Table 1. Stages of our research method, in comparison to the SLR [20] and the DTN [21]

<i>The current study</i>	<i>Structured literature review</i>	<i>Design theory nexus</i>
Selecting articles	Searching for the literature	Identify different approaches in a given area
Filtering articles	Practical screen	
Data extraction & classifying articles	Data extraction Quality assessment	Analyse approaches to identify their applied conditions
Data synthesis	Data synthesis	Formulate the identified conditions into assertions
A model of BPC (Results section)		Design a decision making process Develop an artifact

Table 2. Results of crowdsourcing searches on the eight chosen bibliographic databases

Document types	ACM	Eco Host	Emerald	IEEE	Sage	Science Direct	Springer Link	Wiley	Total
Conference	408	-	-	170	-	-	89	-	667
Journal	3	6	11	47	20	53	58	12	210
Total	411	6	11	217	20	53	147	12	877

Filtering Articles. Using a screening technique [20], this stage filtered out articles that were clearly irrelevant to the focus of the current study by two following steps. We first excluded duplicates, editorial letters, posters, tutorials, work in progress (e.g. abstracts and in-brief papers). This step also eliminated conference articles that were extended and published as journal articles, in order to prevent repeated analysis. The second step restricted the pool of articles by the research question. The elimination was based on the articles’ titles and keywords. Focusing on BPC, this step thus excluded articles applying crowdsourcing to medical and behaviour research, citizen science, learning, and games with a purpose. Adopting a tolerant view suggested by [30], decision to include rather than exclude was made for studies that broadly refer to BPC. As a result, a total of 536 articles remained in the initial pool.

Data Extraction and Classifying Articles. The current study, aligned with Okoli [20], developed a coding form for data extraction and used extracted data in order to classify articles. In detail, the form codified the following four dimensions for extracting data and two questions for classifying articles. The first recorded dimension was general information about the article (article reference, year of publication, data of coding, and additional notes), which are typically extracted in structured literature reviews [30]. Next, our attention turned to the article’s topics. Focusing on articles addressing BPC, we believed that analysing the topics of these articles helped identify the main components of BPC. In particular, we codified this dimension based on themes suggested from previous works, such as ‘task design’, ‘task decomposition’, ‘workflow design’ and ‘incentive mechanism’ proposed by [1] and [7], but still open for emerging categories as an inductive approach.

Another concerned dimension was the research findings, which reflect the different approaches and alternative solutions, necessary for developing a nexus model [21]. In addition, we considered how knowledge was generated from the findings, i.e.

whether these findings can be generalised to other situations or limited to a similar context [31]. The last considered dimension codified practical implications of the articles, including recommendations, to whom these recommendations were targeted, and applied contexts.

To classify articles, the coding form consists of two questions for deciding to include articles: ‘are the topics relevant to BPC?’; and ‘does the article present findings supporting the establishment of BPC?’. Only articles that are both relevant and helpful for the establishment of BPC were fully codified and remained in the reviewed pool. Following Kitchenham et al. [32], the data extraction and classification were undertaken by one researcher, while the other authors randomly checked the procedure. As a result, a total of 238 articles related to the focus of the current study were reviewed in the final pool.

Data Synthesis. This stage synthesized the extracted data to build a model supporting BPC. We reviewed the data extracted by the coding forms, focusing on the articles’ topics, to identify components of the model. This is a four-step process. First, extracted topics were compared and aggregated to several components. We then merged together duplicate components, such as ‘quality control’ and ‘quality estimation’. Third, we mapped sub-components into more generic components. For instance, the sub-component ‘detection of gaming the system’ was mapped to ‘quality control’. Finally, the findings and implications of the reviewed articles were also synthesized, supporting our discussions on the model and its components.

4 Results

As a result of the previous stages, we identified 238 articles related to BPC, of which 71% are conference articles and 29% are journal articles. Regarding the years of publications, Fig. 1 shows the review articles distributed per year from 2008 to 2013. Through this figure, we note an increase on the number of studies published every year, reflecting the mature of the crowdsourcing field. It also indicates that more recently studies have provided more findings that can be generalised to other situations (the top part of the columns in Fig. 1). This leads to the plethora of recently tested and validated findings, solutions, and approaches, which can be seen as promising materials for developing a nexus model supporting BPC.

A closer look at the pool of reviewed articles reveals two groups of studies related to BPC: studies with an integrated view (29 articles) and studies addressing individual aspects (209 articles). In the first category, Table 3 summarises topics and number of articles that adopted an integrated view. From Table 3, the results are that ‘deployment of crowdsourcing’ is the most common topic in this category with 23 articles that focused on designing and deploying several integrated components of a crowdsourcing application. As the articles in this group [e.g. 33] described several specific components of BPC, we further analysed them for their components, and the results were combined with the analysis of the second group that address individual aspects of BPC, as presented in the next sections.

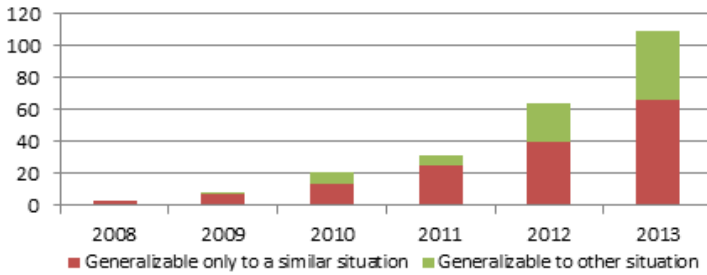


Fig. 1. Reviewed articles per year and how knowledge can be generalised from the findings

Table 3. Topics related to business process crowdsourcing with an integrated view

<i>Main topics</i>	<i>No. of supporting articles</i>
Deployment of crowdsourcing	23
Crowdsourcing framework	4
Design principles for crowdsourcing	2

4.1 Components of Business Process Crowdsourcing

In this section, more detailed results are reported. Focusing on the components of BPC, our analysis on both integrated-view and individual-aspect articles reveals a diverse of components related to BPC. In particular, more than 20 components and sub-components were suggested by the reviewed articles. However, the number of articles supporting these components is largely different. For instance, ‘guide crowdsourcing with Artificial Intelligent’ was supported by only one article, whereas ‘task design’ was discussed by 29 articles. Following a basic assumption of crowdsourcing that groups of researchers are smarter than the smartest individual experts [34], we focused on components proposed by multiple articles.

Table 4 highlights 12 components of BPC that were supported by at least 10 reviewed articles. Within these components, quality control and incentive mechanism are the two most popular components studied in the BPC literature. As crowdsourcing performers are voluntary members in the crowd [7, 14], it is hard for organizations to control the performance of these members. Thus, quality control mechanisms are necessary to make sure that “outcome fulfils the requirements of the requester [organization]” [35]. Also because of the voluntary nature of crowd members, incentive mechanisms are necessary to attract and motivate these members to perform the tasks [36]. To a lesser extent, these results further indicate other components of BPC, including crowd management, task design, results aggregation, workflow design, capability and characteristic of crowdsourcing, task assignment, output, platform, technical configuration, and circumstance to crowdsource and decision factors.

Table 4. Main components of business process crowdsourcing

<i>Components of BPC</i>	<i>No. of supporting articles (n>10)</i>
Quality control	42
Incentive mechanism	37
Crowd management	32
Task design	29
Results aggregation	26
Workflow design	25
Capability & Characteristic of crowdsourcing	23
Task assignment	21
Output	17
Platform	16
Technical configuration	16
Circumstance to crowdsource & Decision factors	16

5 A Nexus Model Supporting the Establishment of BPC

Based on the components identified in the previous section, this section aims at building a model supporting the establishment of BPC. Following the DTN method [21] that starts by designing a decision making process, we first identified the main stages related to BPC. Our analysis on the targeted audiences of the reviewed articles suggests three most important roles related to BPC, including manager (66 articles), designer (186 articles), and programmer (35 articles). Based the traditional system development life cycle [37], we transferred these roles into three stages of BPC, namely decision to crowdsource, design, and configuration. We then used these three stages to structure the identified components, which results a preliminary nexus model supporting the establishment of BPC (Fig. 2). We note that some components in Table 4 were combined together in the model. For instance, both ‘capability and characteristic of crowdsourcing’, and ‘circumstance to crowdsource and decision factors’ help organizations evaluate whether crowdsourcing is a suitable approach, and thus were combined into the ‘decision to crowdsource’. ‘Technical configuration’ and ‘platform’ were also merged because crowdsourcing configuration should be examined on a particular platform. Besides, ‘task decompositions’ was integrated to ‘workflow design’, while ‘task assignment’ was combined with ‘crowd management’. The detailed model is discussed below.

Decision to Crowdsource. According to the reviewed articles [38, 39], *the decision to crowdsource* is positioned in the first phase of the crowdsourcing activity. Therefore, it is presented as the initial component in our model (component 1). Using *the input*, this component initially conceptualizes the crowdsourcing application in order to “decide whether the crowdsourcing approach is appropriate to solve their internal problem/problems [tasks]” (p. 322) [38]. Examining this component, our previous study has already identified and analysed several factors influencing the decision to crowdsource [40]. That study classified and structured the identified

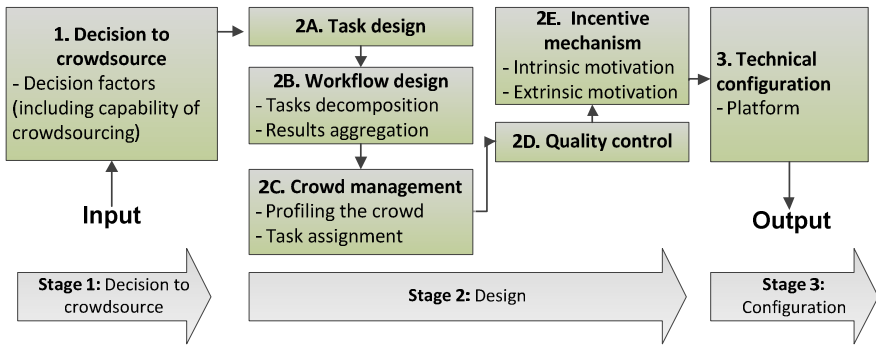


Fig. 2. A preliminary nexus model supporting the establishment of BPC

factors into a decision framework, considering task, people, management, and environmental factors. Based on the framework, the study [40] proposed a series of decision tables with actionable guidelines for making a crowdsourcing decision.

Design. After an organization decides to crowdsource, the design stage transfers the conceptual information determined by the decision factors into concrete design. In this stage, *task design* is important in the crowdsourcing activity, and thus was proposed as the second component in the model (component 2A). Both Malone et al. [41] and Rosen [42] suggest clearly defining what tasks are crowdsourced. Similarly, most studies in our review that deployed a crowdsourcing application have focused on designing tasks as a crucial part of their deployment [33, 43]. To design crowdsourcing tasks, the task properties suggested by [44] can be used as a starting point.

The next component, *designing workflow*, “facilitate[s] decomposing tasks into subtasks, managing the dependencies between subtasks, and assembling the results” [1]. Adopting this definition, we integrated ‘task decomposition’ and ‘results aggregation’ as two sub-components of ‘workflow design’ (component 2B). The role of this component has been highlighted by several researchers, who do not examine individual tasks but the whole crowdsourcing workflow [1, 27]. Organizations can choose different actors to design workflow, including the organization [25], the crowd [45], or a combination between the crowd and the organization [27].

Crowd management is the component that refers to how organizations manage members in the crowd to achieve defined tasks (component 2C). Addressing this component, the literature suggests two sub-components: profiling the crowd [46, 47] and assigning tasks according to profiles [48]. In profiling the crowd, organizations need to evaluate the completeness and effectiveness of crowd members when performing tasks [1, 47], and use this evaluation to build the member profiles. Based on these profiles, different mechanisms can be devised to assign tasks to suitable members, such as the auction-based mechanism [13] and scheduled mechanism [48].

According to Table 4, *quality control* (component 2D) is the most popular component addressed by the reviewed articles, which implies its important role in BPC. The fact that crowdsourcing workers have diverse background and knowledge [14] and work voluntarily may lead to poor results. Thus, quality control is necessary. Agreeing on the necessity of this component, Naroditskiy et al. [49] extend this

component by including functions for preventing malicious behaviours from the crowd members. In the reviewed literature, several quality control mechanisms were proposed, which can be generally grouped into two approaches: design-time and run-time [35]. At design time, organizations can design tasks in a robust way for reducing malicious behaviours, like several mechanisms proposed by [50]. At run-time, organizations can choose three mechanisms for controlling crowdsourcing quality, using the crowd, using experts, and relying on third-party organizations [7].

Organizations, which aim to successfully design a crowdsourcing application, need to attract and engage the crowd members. This attraction can be done through *incentive mechanisms* (component 2E). Borrowing from psychology that two main types of motivation are intrinsic and extrinsic ones [51], incentive mechanisms in BPC should influence different factors of the intrinsic and (or) extrinsic motivation. For extrinsic motivation, most of the reviewed articles examine the usage of financial incentives [26, 36]. For intrinsic motivation, several other factors were suggested, such as love of the community [26] and help other people through meaningful tasks [52].

Configuration. The final component focuses on how to *configure* crowdsourcing in a certain platform (component 3). In general, organizations can choose to develop or use an existing platform. However, with the availability of several crowdsourcing platforms, where a large number of members exist, the choice of utilising available platforms seems to be more attractive, and was supported by several studies [53, 54]. Given that, we suggest configuring crowdsourcing applications on a chosen platform, rather than developing a new platform. Another reason for choosing an existing platform is that the current literature has proposed several tools supporting the configuration, such as Turkit [55] and Crowdforge [45]. As a result, this component returns an *output* of the process, which includes an installation of the crowdsourcing application and the accomplished tasks that were crowdsourced.

6 Conclusion and Future Work

Addressing the lack of a way to establish crowdsourcing as an organizational business process [9-11, 56], this study proposed a preliminary nexus model supporting the establishment of BPC. We identified and synthesized several important components of BPC. We then chose 12 components that were suggested by at least 10 reviewed articles and integrated them into a model supporting the establishment of BPC. From the ‘wisdom of the researchers’ where the collective researchers are smarter than the few [34], it can be stated that our model and its components capture the main business processes of crowdsourcing as they were supported by multiple articles. As a result, the current study has provided important implications for both academics and practitioners.

From the academics’ perspective, our study adopted a broad view of what the literature has reported on BPC, overcoming the ad-hoc issues in the crowdsourcing literature [15, 16]. As a result, the study provides a good starting point for academics from both the crowdsourcing field and other disciplines that aim to follow up the components or model discussed in this work. For instance, researchers from computer

security, who may use crowdsourcing for collecting and processing malware datasets, can use our model for building the corresponding business process.

Methodologically, the current study validates the design science method proposed by Pries-Heje and Baskerville [21] when applying it to the context of crowdsourcing. Additionally, we extend this method by combining it with a SLR [20] that systematically identifies existing approaches and components in the crowdsourcing literature, which is a key requirement for this design science method [21]. From another methodological aspect of IS literature review, our study is one of the most comprehensive reviews in the crowdsourcing field, in terms of number of reviewed articles. We analysed 238 articles, compared to 55 articles in a review by Zhao and Zhu [7]. Consequently, our review contributes to establish background for the emerging of crowdsourcing field [29].

From the practical view, our study provides insights for organisations to employ business processes based on crowdsourcing. In particular, our model has seven sequent components that were structured corresponding to three stages: the decision to crowdsource, design, and configuration, which can be used to guide how to plan, analyse, design, and configure BPC. Based on this model, we also provided discussions and implications about approaches and solutions in each component, contributing to organise case evidences that are currently unarranged in the crowdsourcing practices [57].

As future work, an interesting direction is research on transferring the model into a tool supporting BPC. This requires detailed rules or assertions that can be directly applied to the decision making process [21]. Thus, we plan to extend our preliminary model by further analysing the reviewed articles. In fact, a part of this analysis was conducted by our previous work [40], where we analysed the decision to crowdsource and proposed a series of decision tables for making crowdsourcing decision. Another future direction includes explicit formalizing concepts related to BPC and exploring relationship between these concepts, which can be based on the components of our proposed model. This direction can lead to an ontology enriching the understanding on BPC and providing a mean for sharing knowledge in the domain.

References

1. Kittur, A., et al.: The Future of Crowd Work. In: Proceedings of the 2013 Conference on Computer Supported Cooperative Work (2013)
2. Howe, J.: The rise of crowdsourcing. *Wired Magazine*, 1–4 (2006)
3. Estellés-Arolas, E., González-Ladrón-de-Guevara, F.: Towards an integrated crowdsourcing definition. *Journal of Information Science* 38(2), 189–200 (2012)
4. Brabham, D.C.: *Crowdsourcing*. The MIT Press, Cambridge (2013)
5. Gassenheimer, J.B., Sigauw, J.A., Hunter, G.L.: Exploring motivations and the capacity for business crowdsourcing. *AMS Review* 3(4), 205–216 (2013)
6. Djelassi, S., Decoopman, I.: Customers' participation in product development through crowdsourcing: Issues and implications. *Industrial Marketing Management* 42(5), 683–692 (2013)
7. Zhao, Y., Zhu, Q.: Evaluation on crowdsourcing research: Current status and future direction. *Information Systems Frontiers*, 1–18 (2012)

8. La Vecchia, G., Cisternino, A.: Collaborative Workforce, Business Process Crowdsourcing as an Alternative of BPO. In: Daniel, F., Facca, F.M. (eds.) ICWE 2010. LNCS, vol. 6385, pp. 425–430. Springer, Heidelberg (2010)
9. Khazankin, R., Satzger, B., Dustdar, S.: Optimized execution of business processes on crowdsourcing platforms. In: IEEE 8th International Conference on Collaborative Computing: Networking, Applications and Worksharing, Pittsburgh, PA (2012)
10. Lopez, M., Vukovic, M., Laredo, J.: PeopleCloud Service for Enterprise Crowdsourcing. In: 2010 IEEE International Conference on Services Computing (SCC), Miami, FL (2010)
11. Vukovic, M., Laredo, J., Rajagopal, S.: Challenges and experiences in deploying enterprise crowdsourcing service. In: Benatallah, B., Casati, F., Kappel, G., Rossi, G. (eds.) ICWE 2010. LNCS, vol. 6189, pp. 460–467. Springer, Heidelberg (2010)
12. van der Aalst, W., Hee, K.M.: Workflow management: models, methods, and systems. The MIT Press, Cambridge (2004)
13. Satzger, B., Psailer, H., Schall, D., Dustdar, S.: Stimulating skill evolution in market-based crowdsourcing. In: Rinderle-Ma, S., Toumani, F., Wolf, K., et al. (eds.) BPM 2011. LNCS, vol. 6896, pp. 66–82. Springer, Heidelberg (2011)
14. Hirth, M., Hoßfeld, T., Tran-Gia, P.: Analyzing costs and accuracy of validation mechanisms for crowdsourcing platforms. *Mathematical and Computer Modelling* 57(11-12), 2918–2932 (2012)
15. Geiger, D., Schader, M.: Personalized task recommendation in crowdsourcing information systems—Current state of the art. *Decision Support Systems* (in press, 2014)
16. Man-Ching, Y., King, I., Kwong-Sak, L.: A Survey of Crowdsourcing Systems. In: Privacy, security, risk and trust (passat), 2011 IEEE third International Conference on Social Computing (socialcom), Boston, MA (2011)
17. Geiger, D., et al.: Managing the crowd: towards a taxonomy of crowdsourcing processes. In: Proceedings of the Seventeenth Americas Conference on Information Systems (2011)
18. Hetmank, L.: Components and Functions of Crowdsourcing Systems—A Systematic Literature Review. In: 11th International Conference on Wirtschaftsinformatik, Leipzig, Germany (2013)
19. Hevner, A., et al.: Design science in information systems research. *MIS Quarterly* 28(1), 75–105 (2004)
20. Okoli, C.: A critical realist guide to developing theory with systematic literature reviews. Available at SSRN 2115818 (2012)
21. Pries-Heje, J., Baskerville, R.: The design theory nexus. *MIS Quarterly* 32(4), 731–755 (2008)
22. Carlsson, S.A., et al.: Socio-technical IS design science research: developing design theory for IS integration management. *Information Systems and e-Business Management* 9(1), 109–131 (2011)
23. Gregor, S., Jones, D.: The anatomy of a design theory. *Journal of the Association for Information Systems* 8(5), 312–335 (2007)
24. Schenk, E., Guittard, C.: Towards a characterization of crowdsourcing practices. *Journal of Innovation Economics* 7(1), 93–107 (2011)
25. Rouse, A.C.: A preliminary taxonomy of crowdsourcing. In: Proceedings of the 21st Australasian Conference on Information Systems, pp. 1–10 (2010)
26. Kaufmann, N., Schulze, T., Veit, D.: More than fun and money. worker motivation in crowdsourcing—a study on mechanical turk. In: Proceedings of the Seventeenth Americas Conference on Information Systems, Detroit, MI (2011)

27. Kulkarni, A., Can, M., Hartmann, B.: Collaboratively crowdsourcing workflows with turkomatic. In: Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, pp. 1003–1012. ACM, Seattle (2012)
28. Peffers, K., et al.: A design science research methodology for information systems research. *Journal of Management Information Systems* 24(3), 45–77 (2007)
29. Webster, J., Watson, R.T.: Analyzing the past to prepare for the future: writing a literature review. *MIS Quarterly* 26(2), xiii–xxiii (2002)
30. Okoli, C., Schabram, K.: A guide to conducting a systematic literature review of information systems research. *Sprouts: Working Papers on Information Systems* 10(26) (2010)
31. Mingers, J.: The paucity of multimethod research: a review of the information systems literature. *Information Systems Journal* 13(3), 233–249 (2003)
32. Kitchenham, B.: Guidelines for performing systematic literature reviews in software engineering Version 2.3, in EBSE Technical Report, Keele University and University of Durham (2007)
33. Bojin, N., Shaw, C.D., Toner, M.: Designing and deploying a ‘compact’ crowdsourcing infrastructure: A case study. *Business Information Review* 28(1), 41–48 (2011)
34. Surowiecki, J.: *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business*. Economies, Societies and Nations. Doubleday, New York (2004)
35. Allahbakhsh, M., et al.: Quality control in crowdsourcing systems: Issues and directions. *IEEE Internet Computing* 17(2), 76–81 (2013)
36. Mason, W., Watts, D.J.: Financial incentives and the “performance of crowds”. In: Proceedings of the ACM SIGKDD Workshop on Human Computation, pp. 77–85 (2009)
37. Lucas, H.C.: *Information technology: Strategic decision making for managers*. John Wiley & Sons, Hoboken (2005)
38. Muhdi, L., et al.: The crowdsourcing process: an intermediary mediated idea generation approach in the early phase of innovation. *International Journal of Entrepreneurship and Innovation Management* 14(4), 315–332 (2011)
39. Wexler, M.N.: Reconfiguring the sociology of the crowd: exploring crowdsourcing. *International Journal of Sociology and Social Policy* 31(1/2), 6–20 (2011)
40. Thuan, N.H., Antunes, P., Johnstone, D.: Factors influencing the decision to crowdsource. In: Antunes, P., Gerosa, M.A., Sylvester, A., Vassileva, J., de Vreede, G.-J. (eds.) *CRIWG 2013*. LNCS, vol. 8224, pp. 110–125. Springer, Heidelberg (2013)
41. Malone, T.W., Laubacher, R., Dellarocas, C.: The collective intelligence genome. *IEEE Engineering Management Review* 38(3), 38–52 (2010)
42. Rosen, P.A.: Crowdsourcing Lessons for Organizations. *Journal of Decision Systems* 20(3), 309–324 (2011)
43. Corney, J., et al.: Putting the crowd to work in a knowledge-based factory. *Advanced Engineering Informatics* 24(3), 243–250 (2010)
44. Zheng, H., Li, D., Hou, W.: Task Design, Motivation, and Participation in Crowdsourcing Contests. *International Journal of Electronic Commerce* 15(4), 57–88 (2011)
45. Kittur, A., et al.: Crowdforge: Crowdsourcing complex work. In: Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology, pp. 43–52 (2011)
46. Celis, L.E., Dasgupta, K., Rajan, V.: Adaptive crowdsourcing for temporal crowds. In: Proceedings of the 22nd International Conference on World Wide Web Companion. International World Wide Web Conferences Steering Committee, pp. 1093–1100. Rio de Janeiro, Brazil (2013)

47. Allahbakhsh, M., et al.: Reputation management in crowdsourcing systems. In: 8th International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom) (2012)
48. Khazankin, R., Schall, D., Dustdar, S.: Predicting qoS in scheduled crowdsourcing. In: Ralyté, J., Franch, X., Brinkkemper, S., Wrycza, S. (eds.) CAiSE 2012. LNCS, vol. 7328, pp. 460–472. Springer, Heidelberg (2012)
49. Naroditskiy, V., et al.: Crowdsourcing dilemma. arXiv preprint arXiv:1304.3548 (2013)
50. Eickhoff, C., de Vries, A.: Increasing cheat robustness of crowdsourcing tasks. *Information Retrieval* 16(2), 121–137 (2013)
51. Ryan, R.M., Deci, E.L.: Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions. *Contemporary Educational Psychology* 25(1), 54–67 (2000)
52. Chandler, D., Kapelner, A.: Breaking monotony with meaning: Motivation in crowdsourcing markets. *Journal of Economic Behavior & Organization* 90, 123–133 (2013)
53. Feller, J., et al.: ‘Orchestrating’ sustainable crowdsourcing: A characterisation of solver brokerages. *The Journal of Strategic Information Systems* 21(3), 216–232 (2012)
54. Chanal, V., Caron-Fasan, M.L.: The difficulties involved in developing business models open to innovation communities: the case of a crowdsourcing platform. *M@n@gement* 13(4), 318–340 (2010)
55. Little, G., et al.: TurkIt: human computation algorithms on mechanical turk. In: Proceedings of the 23rd Annual ACM Symposium on User Interface Software and Technology, pp. 57–66 (2010)
56. Satzger, B., et al.: Auction-based crowdsourcing supporting skill management. *Information Systems* 38(4), 547–560 (2012)
57. Kärkkäinen, H., Jussila, J., Multasuo, J.: Can crowdsourcing really be used in B2B innovation? In: Proceeding of the 16th International Academic MindTrek Conference, pp. 134–141 (2012)

Link Prediction in Social Networks Based on Local Weighted Paths

Danh Bui Thi¹, Ryutaro Ichise², and Bac Le¹

¹Computer Science Department, VNUHCM-University of Science, Vietnam
{btdanh, lhbac}@fit.hcmus.edu.vn

²National Institute of Informatics, Tokyo, Japan
ichise@nii.ac.jp

Abstract. A graph path, a sequence of continuous edges in a graph, is one of the most important objects used in many studies of link prediction in social networks. It is integrated in measures, which can be used to quantify the relationship between two nodes. Due to the small-world hypothesis, using short paths with bounded lengths, called local paths, nearly preserves information, but reduces computational complexity compared to the overall paths in social networks. In this paper, we exploit local paths, particularly paths with weight, for the link-prediction problem. We use PropFlow [16], which computes information flow between nodes based on local paths, to evaluate a relationship between two nodes. The higher the PropFlow, the higher the probability that the nodes will connect in the future. In this measure, link strength has a strong link to the measure's performance as it directs information flow. Therefore, we investigate ways of building a model that can efficiently combine more than one useful property into link strength so that it can improve the performance of PropFlow.

Keywords: Link prediction, information flow, link strength.

1 Introduction

Link prediction is a basic computational problem underlying the evolution of social networks. Given a snapshot of a social network at time t , this problem attempts to accurately predict the edges that will be added to the network during the interval from t to a given future time t' . Due to the recent exponential growth of online social networks, link prediction holds a great attraction for researchers. However, it is a hard problem because social networks are highly sparse and dynamic and have a collective structure, and therefore the outcome is difficult to foresee. Solving this problem would be helpful in understanding human relationships in society, and would be applicable to other issues in a wide variety of domains, such as molecular biology, criminal investigations, and recommendation systems. Link prediction can be used for suggesting friends or collaborations, seeking missing links in criminal systems, predicting genetic or protein-protein interactions, or recommending article references.

Approaches dealing with link prediction vary. Liben-Nowell and Kleinberg [13] examined unsupervised measures for making predictions, such as common neighbors, *Adamic/Adar*, *Jaccard*, *Katz*, *Rooted Pagerank*, *Simrank*, and *Hitting time*. All of these measures are based on network structure, such as node degree, common nodes, or a global topology which detects overall paths. *Katz* directly sums over a collection of paths, exponentially damped by length to count short paths more heavily, while *Rooted PageRank* for (u, v) is the stationary probability of reaching v in a random walk starting from u , measured by the propagation of the node's transition probability. These experiments show that a number of the proposed measures significantly outperform a random predictor. This suggests that there is useful information contained in network topology. One drawback of the above methods is that global topology is expensive in computational cost. Extending Liben-Nowell and Kleinberg's study, other authors proposed new measures which only use local topology. The small-world phenomenon in social networks, the principle that people are all linked by short chains of acquaintances, allows these other researchers to implement local topology. This can reduce not only the time and space complexity, but also reduce noise data. Some noteworthy measures are *Friend Link* [15], *PropFlow* [16], and *T-Flow* [14]. *Friend Link* [15] is a customization of the *Katz* measure that only considers a graph path with bounded length l . *PropFlow* is somewhat similar to *Rooted PageRank*, but it is a more localized measure of propagation and is insensitive to topological noise far from the source node [16]. *T-Flow* [14] is an enhancement of *PropFlow* (which will be explained in the next section) that includes link activeness, and this activeness is reflected through the time during which interaction occurs between two nodes.

Supervised machine learning is another major approach in which link prediction is considered as a binary classification. In this approach, most methods encounter two primary challenges. First, the nodes that have connections account for a very small fraction compared to all the nodes in the network. This causes class imbalance, leading to misclassification of the minority class. The second challenge is feature selection, which refers to how to obtain good features from the network structure and the attributes of nodes/edges. To overcome the imbalance issue, we can apply resampling methods to balance a dataset or use classifier ensembles [7]. In [5], the authors built link predictors using classifier ensembles such as *AdaBoost*, *Bagging*, *Rotation Forest*, and *Random Forest*, and obtained better results than with *C4.5*, *k-NN*, and *SVM*. With regard to the second challenge, many studies have searched for useful features, with common neighbors and node degrees being popular features. These features can be used independently or integrated with metrics such as *Adamic/Adar*, *Jaccard*, *Preferential Attachment*, and *Transitive Friends*. Graph topology also has been used, as seen in *Katz*, *Rooted PageRank*, and *Simrank*[13]. Some studies characterize a network in context of triads, subgraphs consisting of three nodes. Triad distribution is used to examine social network evolution [10] or as a feature in supervised learning [8], [12]. In addition to graph topology, the attributes of nodes or edges are also of interest as, for example, keywords [17], places [18],

email [3], and transaction events [9]. Due to security and privacy concerns, this type of information is generally limited in datasets.

In this paper, we investigate the effects of weighted local paths on link prediction in social networks. We examine different measures based on local paths and choose PropFlow (a measure proposed by Lichtenwalter et al. [16]) to evaluate the relationship of a pair of nodes in our model. This procedure uses link strength to observe information flow in a social network. Information flow starts at a source node and runs to other nodes via edges; its value corresponds to the probability that the flow reaches one node from a particular source node. The higher the probability is, the more likely a link will be created in future. One drawback of their research is that it uses only the quantity of interactions between nodes for link strength. If the quantity of the interactions is not the unique thing depicting how information flows in a network, such as kind of interaction and content of interaction, link strength computation should be considered. Furthermore, link formation in a social network often depends on more than one standard, and using one standard cannot provide enough information to make the flow efficient enough for link prediction. Therefore, we propose a model which computes link strength as a combination of multiple standards so that we can improve the accuracy of link prediction. Our model is a customization of a genetic algorithm. It searches for the optimal combination so that the link strength, which is computed as a function of the combination, will guide the information flow starting at a node, and move to potential nodes that are more likely than other nodes.

The idea of modeling link strength has been investigated in previous studies, including unsupervised and supervised learning. The method proposed in [19] is unsupervised. The authors formulated a latent variable model integrating interaction activities and user profiles by considering link strength to be a hidden effect of them. Then, they developed a coordinate ascent optimization procedure to infer strength from the model. Gilbert and Karahalios [4] approached the problem as a supervised issue. They modeled link strength as a linear combination of predictive variables plus terms for dimension interactions and network structure. These predictive variables consist of 74 variables, separated into seven categories: intensity, intimacy, duration, reciprocal services, structure, emotional support, and social distance. They used an iterative variant of the OLS regression to fit their model. They needed some participants to manually notate labels for training samples. As a result, the training set comprises about 2000 links, which not large enough to cover the probable cases. Unlike [4], [9] took advantage of a “top friends” tag to indicate whether a relationship is strong. However, the tag is not available in many social network datasets. *Supervised Random Walk* [11] is a training algorithm for learning the link strength estimation function. It attempts to direct parameters in the function so that link strengths in the network guide a random walker more likely to visit the nodes to which new links will be created in the future. Supervised Random Walk works with Rooted PageRank, which is a global topology-based measure and therefore has a high computational complexity.

The rest of this paper is organized as follows. Section II describes the PropFlow measure, i.e., how to determine it for each node pair. Link strength and the search framework are presented in Section III. Experiment results are given in Section IV. Finally, Section V is the Conclusion.

2 PropFlow Measure

PropFlow [16] is an unsupervised measure which corresponds to the probability that restricts a random walk starting at v_s to end at v_d in l steps or fewer using link strength as transition probabilities. PropFlow is somewhat similar to Rooted PageRank, but examines local paths instead of overall paths. An algorithm computing PropFlow is a simple modification of a breadth-first search limited to height l . The detailed procedure for PropFlow estimation is given in Algorithm 1.

Algorithm 1. PropFlow Estimation

Require: the network $G(V,E)$, node v_s , max length l

Ensure: PropFlow values for all $n \leq l$ -degree neighbors v_d of v_s .

begin

 insert v_s into Found

 push v_s into NewSearch

 insert $(v_s, 1)$ into S

For CurrentDegree = 0 to l

 OldSearch \leftarrow NewSearch

 empty NewSearch

 While OldSearch is not empty

 pop v_i from OldSearch

 find NodeInput using v_i in S

 SumOutput \leftarrow 0

 For each v_j in neighbors of v_i

 add strength of e_{ij} to SumOutput

 End For

 Flow \leftarrow 0

 For each v_j in neighbors of v_i

$w_{ij} \leftarrow$ strength of e_{ij}

 Flow \leftarrow NodeInput $\ast \frac{w_{ij}}{\text{SumOutput}}$

 insert or sum v_j, Flow into S

 If v_j is not in Found

 insert v_j into Found

 push v_j into NewSearch

 End If

 End For

 End While

End For

end.

3 Link Strength Model

According to [6], link strength should be satisfied by the following definition: the strength of a tie is a combination of the amount of time, the emotional intensity, the intimacy (mutual confiding), and the reciprocal services which characterize the tie. Link strength indicates relationships more faithfully than a binary representation. According to the theory of cognitive balance, as formulated by Heider and especially by Newcomb [6], if strong links A-B and A-C exist, and if B and C are aware of one another, a positive link would be introduced between B and C, since C will want his feelings to be congruent with his good friend, A, and similarly for B. Where the links are weak, such consistency is psychologically less crucial [6]. Therefore, mining link strength provides more useful information to predict links. This paper models link strength as a linear combination of the involved factors. A framework is built by using a genetic algorithm to search for the optimal combination.

3.1 Model Specification

As mentioned above, we model link strength as a linear combination of multiple information factors as equation (1). The factors are a reflection of link strength into observable things, which we call strength features.

$$a_{uv} = \frac{1}{1 + \exp(-wx)} \tag{1}$$

where x is the feature vector of link (u, v) , and each feature corresponds to an information factor impacting link strength; w is the parameter vector. Our task now is to learn the parameters w of link strength function a_{uv} . As manual labeling for training samples is an expensive task in supervised learning, especially in link strength classification, our aim is to build a framework which can use available link information to learn w .

G_t and G_{t+1} are snapshots of the social network at time t and $t+1$. Consider a sub-graph of G_t which consists of the four nodes a, b, c, d , shown on the left-hand side of Figure 1. The graph on the right-hand side of Figure 1 is the subgraph at future time $t + 1$. We denote this subgraph as G_Δ . As shown in the figure, a connection appears from c to d during the interval from t to time $t + 1$, but does not occur between c and a even if they have the same common neighbor, b . The hypothesis is posed so that at time t , c is aware of d more than of a , and this affects c 's decisions to make friends. From this hypothesis, we define the optimization problem to find the optimal parameters w as follows:

$$\operatorname{argmax}_w f = |\{G_\Delta(\{a, b, c, d\}, E_\Delta), \operatorname{PropFlow}_{c,a}(w) < \operatorname{PropFlow}_{c,d}(w)\}| \tag{2}$$

$$G_\Delta \subseteq G_t, E_\Delta = \{(b, a), (b, c), (b, d)\} \\ (a, c) \notin E_t, (c, d) \notin E_t, (a, c) \notin E_{t+1}, (c, d) \in E_{t+1}$$

This means finding the parameter w that can maximize the number of subgraphs $G_{\Delta}(a, b, c, d)$ satisfying that the PropFlow of the pair node (c, a) is less than pair node (c, d) . To solve the problem, we apply a genetic algorithm with the fitness function $f(w)$ in equation (2). Details of the optimization process are given in Algorithm 2.

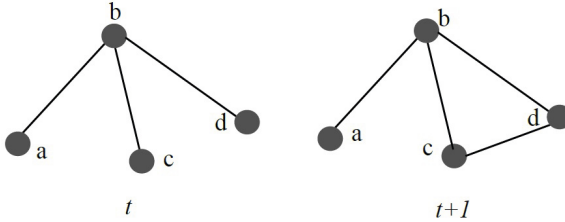


Fig. 1. Subgraph pattern G_{Δ}

Algorithm 2. Search for optimal parameters w

Require: k , network G_t, G_{t+1}

Ensure: parameters w

begin

 Initialize $S = \emptyset$

 Edge e appears during the interval from t to $t+1$

 Find all subgraphs G_{Δ} which involve to e , put into S

 End For

 Randomly initialize k sets of the parameters w

 Each set of parameters is one chromosome.

 Evaluate the fitness value using equation 2 on S

 Repeat

 Create a new population by repeating the following steps:

Selection

Crossover

Mutation

 Evaluate the fitness values using equation 2 on S

 Publish new population with the k best chromosomes

 Until objective function Converged

end.

To find the subgraphs G_{Δ} which involve to a new directed edges $e = (c, d)$ adding to the social network during the interval from t to $t + 1$, we use the following procedure: At first, it lists all common neighbors b of c and d , then explores other nodes that are neighbors of a but not create link with d . Each collection consisting of four above nodes (a, b, c, d) forms a subgraph G_{Δ} . In our experiments, we use $k = 40$, a crossover rate of 0.85, a mutation rate of 0.15, and

a maximum loop number of 50. The fitness value slowly increases for each loop. The chromosomes are represented as an array of bits (0/1). Due to the features being normalized to $[0, 1]$, we can limit each parameter to the $[-1, 1]$ range. A stochastic universal sampling technique[1] is applied to selection step, one-point crossover for crossover step, and flip bit for mutation.

3.2 Features for Link Strength

Importance Level: The importance level implies how important the link is to the involved nodes. We define three levels corresponding to the number of common neighbors at the time the link appears. If no common neighbors exist, there is a strong possibility that these nodes know each other outside of the social network (in real world), and so we set the highest level, as 3. If the number of common neighbors is less than 3, we assign the level as 2; otherwise, we assign 1.

Activeness: Interactions refer to communication activities over time between two nodes such as wall posting, picture tagging, and collaboration. They depict the activeness of links, if node pairs interact more recently, then the corresponding links become more active. Link activeness is estimated as follow:

$$activeness_{u,v} = \sum_{t=1}^T \frac{|I_{uv}(t)|}{T-t+1}, \quad (3)$$

where t is the time stamp of an associated network snapshot when the interaction takes place. It can be hours, days, months, or years; T is the time stamp of the latest snapshot; and $I_{uv}(t)$ is a collection of the interactions occurring during the interval from time $t-1$ to time t .

Common Neighbors: This feature refers to the size of the neighbor set that two nodes share but it only enumerates neighbors appearing after the time u, v make a connection. These neighbors show the impact of the (u, v) link on the attitude of other nodes. In other words, they show the strength of the link, while common neighbors existing before the (u, v) link creation make u and v aware of each other. It is expected that the larger the number of common neighbors, the higher the chances that both nodes are close.

Similarity: Each node has its own attributes such as gender, hobby, working or home address, number of friends, and activeness. Similarity implies a similarity in the attributes of the nodes. We expect that the more similarity nodes have, the more stable their relationship. This is due to their common attributes making them communicate easily and frequently.

4 Experimental Evaluation

To evaluate our proposed link strength model, we use PropFlow and T-Flow measures as features in a supervised learning method. We choose the Bagging algorithm for the learning method due to the imbalance class issue in the datasets.

Table 1. Statistics of Facebook datasets

Dataset	Nodes	Edges	Snapshot quantity
D1	25, 334	233, 523	10
D2	34, 268	388, 392	15
D3	44, 159	630, 325	20
D4	57, 924	1, 175, 701	25

Table 2. Statistics of training and testing data

Dataset	Training (node pair)	Testing (node pair)
D1	8, 328	10, 908
D2	15, 118	14, 910
D3	52, 899	53, 021
D4	112, 143	166, 080

We estimate T-Flow and PropFlow by using the number of interactions between nodes as link strength, and PropFlow based on the link strength which comes from our model. The following subsections describe how to prepare data for training/testing and the experimental results for each dataset.

4.1 Data Preparation

We work with a Facebook dataset collected from the regional Facebook network of New Orleans from September 2006 to January 2009 [2]. This dataset consists of wall postings exchanged by 60,290 users who are connected by 1,545,686 links. Wall postings are considered as interactions between users. We take snapshot for the network each month and arrange them in order of time. The first 25 snapshots are used to construct four datasets for our experiments; each dataset includes T snapshots from G_1 to G_T . The details of these datasets are shown in Table 1.

To ascertain parameters w of the link weight function, we apply Algorithm 2 to snapshots G_{T-2} and G_{T-1} . For training the Bagging algorithm, we collect positive samples that are the links appearing during the interval from time $T-2$ to $T-1$. A classifier is used to predict the links appearing from time $T-1$ to T . Due to the presence of overwhelming negative samples or node pairs without connections in the datasets, we conduct the resampling method as follows: for each positive sample, we randomly select 5 negative samples. Table 2 shows the statistics of training and testing data for each dataset.

The features of training samples are extracted from snapshots G_1, \dots, G_{T-2} , while the features of the testing samples come from snapshots G_1, \dots, G_{T-1} . Below are the features used in our experiments.

Common Neighbors: The number of neighbors two nodes have in common.

$$score(u, v) = |\Gamma(u) \cap \Gamma(v)|$$

where $\Gamma(u)$ indicates the neighbor set of node u .

Adamic/Adar: This measure considers two nodes, related or not, based on the friends of their common neighbors. If their common neighbors have more friends with whom to communicate, then they are less related.

$$score(u, v) = \sum_{w \in \Gamma(u) \cap \Gamma(v)} \frac{1}{\log |\Gamma(w)|}$$

Jaccard's Coefficient: The Jaccard coefficient is a commonly used similarity measure in information retrieval. In social network analysis, Jaccard's coefficient of two nodes can be computed as the proportion of common neighbors to the number of their friends. This indicates that two nodes are more similar if most of their friends are common.

$$score(u, v) = \frac{|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u) \cup \Gamma(v)|}$$

Preferential Attachment: This is considered as a model of the growth of a network [13]. The basic premise is that if a node has many neighbors, then the probability of the other nodes attaching to it is high because it is a popular or well-known node. According to this premise, the probability of creating a link between node u and v is correlated with the product of the number of neighbors of u and v .

$$score(u, v) = |\Gamma(u)| \cdot |\Gamma(v)|$$

PropFlow: We mine using two types of PropFlow. One is PropFlow using the number of interactions as link strength, and the other is PropFlow using the link strength which comes from our proposed model.

T-Flow [14]: T-Flow is an enhancement of PropFlow, based on the major idea that the information transition between two nodes is depicted via not only link strength but also the time for which this link is active. A link being active means that there is activity between the nodes at that time. Compared to PropFlow, T-Flow produces a coefficient which is aware of the effect of link activeness on transition probabilities, as in the following equation:

$$p(u, v) = \frac{a_{uv}}{\sum_{w \in \Gamma(u)} a_{uw}} \cdot (1 - \alpha)^{|t_x - t_y|}, \quad (4)$$

where a_{uv} is the strength of link (u, v) , which is determined by the number of interactions between u and v ; $\alpha (0 \leq \alpha \leq 1)$ is a decaying factor; t_x is the time stamp of the link when a random walker comes into node u ; and t_y is the time stamp of the link when the random walker is going to node w .

4.2 Experimental Results

The details of the experiments and the features used are described in Table 3, where the “ $\sqrt{\quad}$ ” sign means the corresponding feature is available, and the “-”

Table 3. List of Features

Feature	Baseline	Ex-01T	Ex-01P	Ex-02P	Ex-03P
Common Neighbors	✓	✓	✓	✓	✓
Adamic/Adar	✓	✓	✓	✓	✓
Jaccard's Coefficient	✓	✓	✓	✓	✓
Preferential Attachment	✓	✓	✓	✓	✓
T-Flow	-	✓	-	-	-
PropFlow	-	-	✓	-	-
PropFlow with link strength	-	-	-	✓	-
PropFlow+	-	-	-	-	✓

Table 4. Performance of Bagging classifier for positive class

Method	Precision	Recall	F-measure	Method	Precision	Recall	F-measure
D1 Dataset				D2 Dataset			
Baseline	0.272	0.036	0.064	Baseline	0.267	0.022	0.04
Ex-01T	0.351	0.054	0.093	Ex-01T	0.38	0.038	0.07
Ex-01P	0.462	0.057	0.101	Ex-01P	0.387	0.042	0.076
Ex-02P	0.394	0.066	0.113	Ex-02P	0.356	0.069	0.116
Ex-03P	0.477	0.114	0.184	Ex-03P	0.497	0.11	0.18
D3 Dataset				D4 Dataset			
Baseline	0.236	0.013	0.025	Baseline	0.301	0.018	0.034
Ex-01T	0.381	0.037	0.067	Ex-01T	0.38	0.029	0.054
Ex-01P	0.415	0.045	0.081	Ex-01P	0.407	0.037	0.068
Ex-02P	0.43	0.064	0.112	Ex-02P	0.443	0.049	0.088
Ex-03P	0.486	0.099	0.165	Ex-03P	0.521	0.094	0.159

sign means the corresponding feature is absent. Using the same classification algorithm, Bagging, we change the feature vector such that it includes the link strength from our model in some cases, but not in other cases. The baseline is the feature vector consisting only of basic measures: *common neighbors*, *Adamic/Adar*, *Jaccard's coefficient*, and *Preferential Attachment*. Other case studies are combinations of the base line with Prop Flow, T-Flow in succession. The depth of local paths for three measures is set to 3, which means we excluded nodes that are more than three links away from a node. In the experimental process, we recognize that a person u makes friends with some people whose information flow starting at u is not high, but it is higher than the others. Therefore, we introduced one more experiment by adding one feature into the learning method. This feature is the ratio of $\text{PropFlow}(u, v)$ to the average $\text{PropFlow}(u)$ starting from u , called PropFlow+. The experimental results show PropFlow+ remarkably improves the predictor in both precision and recall.

Table 4.2 shows the performance of feature combinations for a positive class on Facebook data. The result is low due to the imbalance class issue. The result shows that PropFlow improves the baseline combination and achieves better per-

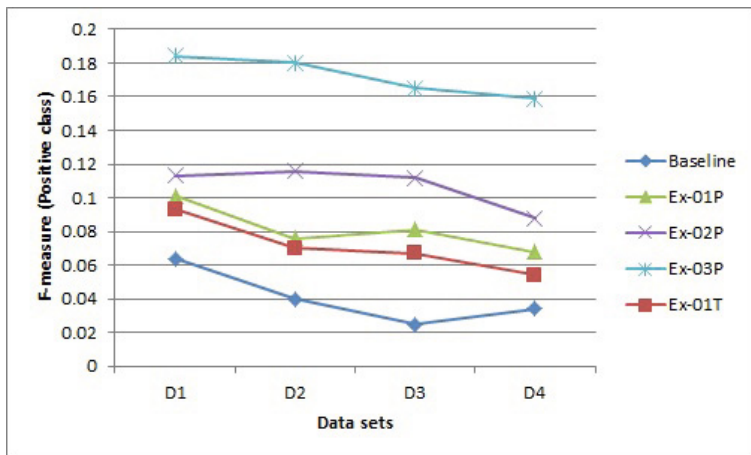


Fig. 2. F-measure (for positive class) of Bagging classifier with different features in each dataset

Table 5. Average F-measure of Bagging classifier with different features in each dataset

Feature	D1	D2	D3	D4
Baseline	0.705	0.737	0.763	0.789
Ex-01T	0.713	0.744	0.77	0.792
Ex-01P	0.718	0.745	0.773	0.794
Ex-02P	0.718	0.749	0.777	0.797
Ex-03P	0.734	0.765	0.787	0.809

formance than that of T-Flow. We also can see that the precision of PropFlow using interactions is better than that of PropFlow using the link strength model in the first two datasets, but when the social network becomes larger, it is no longer better. This could be caused by our link strength mining neighbor information: the network is larger, which means more neighbors appear and this explains more efficiently the strength between two nodes. The recall of the link strength model gives better results than interaction usage alone, and the same thing happens with the F-measure. Although the difference is not dramatic, it shows that the information integration model works. Another thing which we can see from the experiment is the improvement caused by introducing average PropFlow into the system. This increases the performance not only in recall but also in precision. One thing that can be observed in Figure 2 is that the F-measure tends to go down when the network becomes larger for all feature combinations.

Table 5 enumerates the average F-measure of the Bagging classifier of both classes for each feature combination. In general, PropFlow with the link strength model and PropFlow+ obtain performances better than those of the other methods.

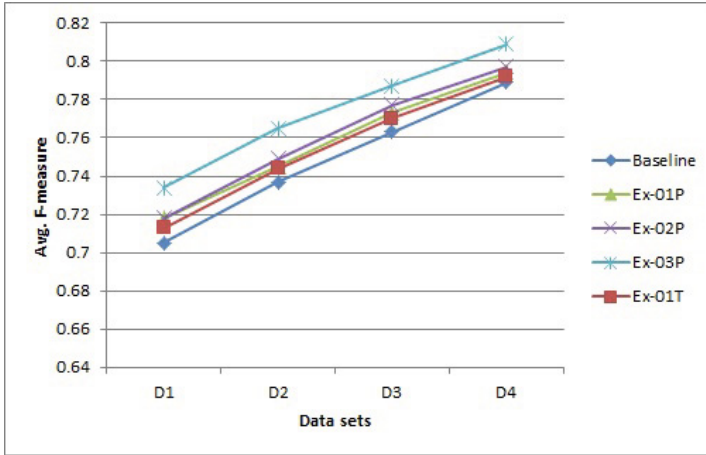


Fig. 3. Average F-measure of Bagging classifier with different features in each dataset

5 Conclusion

In this paper, we conducted an investigation of local paths with weight for link prediction in a social network. PropFlow and T-flow measures were chosen for the examinations conducted. We proposed a link strength model and integrated it into PropFlow to improve the performance of the prediction system. The experiments we conducted showed results that motivate us to continue our work in the future.

Our link strength model currently searches for the optimal combination of strength features by using two snapshots. This eliminate global optima. Therefore, we need improve the model to be able to cumulatively learn from multiple network snapshots. The class imbalance issue is another drawback which we encountered when we applied a supervised learning method for prediction. It dramatically impacted predictor performance, especially for the positive class. Therefore, solving this issue is a very important task for future work.

References

1. Baker, J.E.: Reducing Bias and Inefficiency in The Selection Algorithm. In: Proceedings of the Second International Conference on Genetic Algorithms (1987)
2. Viswanath, B., Mislove, A., Cha, M., Gummadi, K.P.: On the Evolution of User Interaction in Facebook. In: Proceedings of the 2nd ACM Workshop on Online Social Networks, pp. 37–42 (2009)
3. Thi, D.B., Hoang, T.-A.N.: Features Extraction for Link Prediction in Social Networks. In: 13th International Conference on IEEE Computational Science and Its Applications (ICCSA) (2013)
4. Eric, G., Karahalios, K.: Predicting Tie Strength with Social Media. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (2009)

5. Fire, M., Tenenboim, L., Lesser, O., Puzis, R., Rokach, L., Elovici, Y.: Link Prediction in Social Networks Using Computationally Efficient Topological Features. In: Privacy, Security, Risk and Trust (PASSAT), 2011 IEEE Third International Conference on Social Computing (SOCIALCOM), pp. 73–80 (2011)
6. Granovetter, M.: The Strength of Weak Ties. *American Journal of Sociology* 78(6), 1360–1380 (1973)
7. Japkowicz, N., Stephen, S.: The Class Imbalance Problem: A Systematic Study. *Journal Intelligent Data Analysis* 6(5), 429–449 (2002)
8. Ye, J., Cheng, H., Zhu, Z., Chen, M.: Predicting Positive and Negative Links in Signed Social Networks by Transfer Learning. In: Proceedings of the 22nd International Conference on World Wide Web, pp. 1477–1488 (2011)
9. Indika, K., Neville, J.: Using Transactional Information to Predict Link Strength in Online Social Networks. In: ICWSM (2009)
10. Juszczyszyn, K., Musial, K., Budka, M.: Link Prediction Based on Subgraph Evolution in Dynamic Social Networks. In: Privacy, Security, Risk and Trust IEEE 3rd International Conference (2011)
11. Backstrom, L., Leskovec, J.: Supervised Random Walks: Predicting and Recommending Links in Social Networks. In: Proceeding of the 4th ACM International Conference on Web Search and Data Mining, pp. 635–644 (2011)
12. Jure, L., Huttenlocher, D., Kleinberg, J.: Predicting Positive and Negative Links in Online Social Networks. In: Proceedings of The 19th International Conference on World Wide Web. ACM (2010)
13. Liben-Nowell, D., Kleinberg, J.: The Link Prediction Problem for Social Networks. *Journal of the American Society for Information Science and Technology* 58(7), 1019–1031 (2007)
14. Lankeshwara, M., Ichise, R.: Link Prediction in Social Networks using Information Flow via Active Links. *IEICE Transactions on Information and Systems* 96(7), 1495–1502 (2013)
15. Papadimitriou, A., Symeonidis, P., Manolopoulos, Y.: Scalable Link Prediction in Social Networks based on Local Graph Characteristics. In: 9th International Conference on Information Technology: New Generations (ITNG) (2012)
16. Lichtenwalter, R.N., Lussier, J.T., Chawla, N.V.: New Perspectives and Methods in Link Prediction. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2010)
17. Mrinmaya, S., Ichise, R.: Using Semantic Information to Improve Link Prediction Results in Network Datasets. *International Journal of Computer Theory and Engineering* 3, 71–76 (2011)
18. Scellato, S., Noulas, A., Mascolo, C.: Exploiting Place Features in Link Prediction on Location-based Social Networks. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1046–1054 (2011)
19. Rongjing, X., Neville, J., Rogati, M.: Modeling relationship strength in online social networks. In: Proceedings of the 19th International Conference on World Wide Web. ACM (2010)

An Architecture Utilizing the Crowd for Building an Anti-virus Knowledge Base

Nguyen Hoang Thuan^{1,2}, Pedro Antunes¹, David Johnstone¹,
and Minh Nhat Quang Truong²

¹ School of Information Management, Victoria University of Wellington,
PO Box 600, Wellington, New Zealand
{Thuan.Nguyen, Pedro.Antunes, David.Johnstone}@vuw.ac.nz

²Can Tho University of Technology,
256 Nguyen Van Cu Street, Can Tho city, Vietnam
{nhthuan, tmnquang}@ctu.edu.vn

Abstract. Recently, the behaviour-based technique was received attentions for its ability to detect unknown viruses. However, the literature suggests that this technique still needs to be improved due to high false-positive rates. Addressing the issue, the current work-in-progress proposed an architecture utilizing the crowd for building an anti-virus knowledge base, which considers not only virus behaviour but also behaviour from the new applications. This architecture also utilized anti-virus experts in the crowd for classified objects that are unclassified by machines. Using the classified objects, it used a machine learning algorithm to analyse application behaviour from the crowd for updating the knowledge base, and thus the corresponding anti-virus system can correctly diagnose and classify objects, reducing the false-positive rates.

Keywords: Anti-virus, Behaviour-based detection technique, Business process crowdsourcing, Crowdsourcing, Knowledge Base, Machine learning.

1 Introduction

Most computer users rely on anti-virus software to protect their computers from computer viruses. Although anti-virus software is offered by different vendors, it seems that the existing anti-virus software does not effectively protect the users. A recent study on six anti-virus products shows that “the anti-virus software doesn’t always block malware from performing code injection” (p. 69) [1]. This requires an improvement of anti-virus strategies, especially an improvement on detection techniques. The call for improving existing detection techniques was recommended by several researchers [1, 2], who believe that the current detection methods are not enough powerful against the evolutionary of viruses.

By and large, two detection techniques have been deployed in anti-virus systems: signature-based detection and behaviour-based detection [1, 3]. The signature-based techniques, which are mainly based on the known threats’ signature, can only recognize viruses from their known datasets and may have problems when viruses use

obfuscation or polymorphism techniques [4]. Because of the fundamental constraints, the behaviour-based techniques have received the focus of recent studies [3-5]. Some advantages of the behaviour-based approach have been highlighted in the literature. For instance, it can detect unknown computer viruses and new variants of existing viruses [5-7]. It can also avoid processing a huge number of virus samples [4], and adapt over time by using machine learning algorithms [8, 9].

Although the benefits of behaviour-based techniques are well recognized, this approach still has problems for distinguishing malicious from regular behaviours [8]. In particular, the behaviour-based approach may lead to two types of errors. The first type currently considered too high [1] is false-positive detection, where benign applications are classified as malicious. For instance, Unikey, a Vietnamese-keyboard application, may be seen as a virus because of its hook functions, which are functionally similar to key logger threats. On the other hand, the second type of error is the false-negative. It refers to a failure identifying files containing harmful code but acting (or pretending) to be normal. Examples include malicious adware programs often integrated into free downloaded software [10]. In these cases, end-users, rather than anti-virus systems, may identify abnormal activities on their computers, e.g. too high memory use and large network traffic.

Addressing these errors, many studies [e.g. 11, 12, 13] propose techniques for analysing virus samples and deriving significant virus behaviour. These techniques allow building crucial knowledge bases over time [3]. Although these studies improved the overall effectiveness of the behaviour-based approach, they may not be comprehensive in terms of its diversity and timeliness of analysed datasets. The main reason is that knowledge bases have to be shaped and therefore may take time to be updated. Besides, by only analysing a pre-gathered virus datasets, the deriving knowledge base is not thorough, because it does not consider behaviour of new used applications, which “are unknown and therefore have no expected normal behaviour” (p. 64) [1]. As a result, the two aforementioned problems still remain, especially the problem of detecting a new benign application as a virus.

When deploying the D32 (also known as D2) anti-virus project [14], we identified the problems caused by having incomplete knowledge base. Therefore, we suggest that end-users may play important roles in solving the problem. Users that develop knowledge about malicious behaviour by using applications can suggest whether applications are benign or malicious. Besides, collecting users’ data about application behaviour may help building confidence in detecting and classifying viruses. However, the current literature does not propose any framework or technique considering the role of end-users in building anti-virus knowledge bases.

Since crowdsourcing was first introduced by Howe [15] as a strategy that relies on the crowd to achieve specific tasks, the crowdsourcing model has been suggested for doing tasks that require large human resources, like building knowledge bases [16]. Indeed, crowdsourcing has been utilized for building knowledge bases in different application domains. For instance, Wikipedia is a typical example of organizations successfully applying a crowdsourcing strategy when using a significant number of its anonymous users to perform writing and editing activities [17, 18]. Recently, Vukovic et al. [19] deployed a crowdsourcing application to capture IT Inventory knowledge.

Other examples of building knowledge bases by the crowd have been reported by [20-22].

The current work in progress proposes a new architecture utilizing the crowd for building an anti-virus knowledge base. This architecture extends the research by Truong and Hoang [8] that introduced a machine learning algorithm for analysing virus datasets. In particular, we introduce mechanisms to collect users' feedback on application behaviour, and then use both internal and the crowd anti-virus experts for analysing application behaviour and classifying received feedbacks. A machine learning algorithm is applied to analyse these data, and results are used to update the anti-virus knowledge base. We call this architecture CrowdMAV (Applying **Crowd**-sourcing to **Machine Learning Anti-Virus System**).

The main contribution of this work-in-progress is the CrowdMAV architecture that, for the first time, utilizes a crowdsourcing strategy for building a knowledge base of an anti-virus malicious and benign behaviour. This database can reduce the detection of false-positive. Another expected contribution of the current work is its mechanism to handle a huge amount of data receiving from the crowd by also utilizing anti-virus experts from the crowd, which is known as one mechanism of result aggregation in the crowdsourcing field [18]. From a practical perspective, this study helps improving the overall performance of the behaviour-based technique.

2 Related Work

2.1 Anti-virus Detection Techniques

A computer virus is defined as “a program that can ‘infect’ other programs by modifying them to include a possibly evolved copy of itself” (p. 23) [23]. Computer viruses can degrade the performance of a computer by disabling, damaging and destroying computer resources [8], gathering private data, and using resources in unintended ways. With the widespread of computer viruses and malwares, anti-virus software from different vendors has become very popular, in trying to prevent problems by identifying and stopping viruses immediately when they have entered a computer [1]. The existing anti-virus software generally deploys one of the two following detection techniques: a signature-based technique and the behaviour-based technique.

In the signature-based technique, the anti-virus software scans every received (or copied) file with code and compares them with known threat signatures [1]. In particular, this technique requires having an up-to-date signature database extracted from known threats [13], which is often updated daily by vendors. Within the database, each virus has a unique tag that is used to classify suspect files. When a computer receives a new file, the signature-based anti-virus software analyses the content of the received file to determine if it has known malicious tags. Although this technique was widely used in the past, it has two major problems. First, this technique cannot identify viruses that are not recorded in the knowledge base, including polymorphic viruses [1]. Second, this technique needs to analyse file code, which is difficult or sometimes impossible because of obfuscated or packed viruses [4].

Up to now, behaviour-based technique has been studied for more than a decade. This technique dynamically examines unknown files and monitors the file code execution in a controlled environment to detect its malicious behaviour [1]. Several types of behaviour are analysed. For instance, memory usage is examined in dynamic taint analysis [24]. Bayer et al. [4] suggest analysing execution traces. Other behaviour that is typically analysed is Windows API or system calls [3, 25], information flow [11] and network messaging [12]. Classifying the existing malicious behaviour, Hu [26] finds six classes of behaviour, including file-related, process-related, window-related, network-related, register-related, and windows-service behaviour.

Depending on the types of behaviour that are analysed, anti-virus vendors build their knowledge bases, “representing the execution behaviour of a family of malware instances” (p. 1) [3] and consisting of rules for detecting viruses [8]. These knowledge bases largely influence the effectiveness of anti-virus systems. Therefore, several efforts have been made to develop and improve the quality and completeness of knowledge bases [8, 27]. However, despite these efforts, the false-positive rates remain high [1]. We believe that one reason for this failure is that existing anti-virus knowledge bases do not consider new applications used by end-users [1], and thus the corresponding anti-virus system may classify the benign behaviour from these applications as malicious because the knowledge bases were not updated. Another problem with the current knowledge bases is that virus developers are well aware of the behavioural attitudes detected and developed counter measures, as stated that “we have to accept that virus authors are one step more ahead because they decide how to attack first” (p. 7) [2]. To address this problem, anti-virus software needs knowledge bases built from application behaviour reported by the world wide end-users. Addressing this need, we propose an architecture utilizing the crowd for building and extending the knowledge base.

2.2 Crowdsourcing for Building Anti-virus Knowledge Bases

‘Crowdsourcing’ is a concept introduced in 2006 by Howe [15], who referred to crowdsourcing as a model utilizing the crowd for achieving organizational tasks. Since its introduction, crowdsourcing has been widely studied and conceptualized by several researchers, leading to the existence of different definitions. For instance, some authors compared crowdsourcing to the concept of outsourcing [15, 28]. Others considered crowdsourcing as a model for micro-tasks, where users provide their free time to accomplish particular tasks [18, 29, 30]. Recently, Estellés-Arolas and González-Ladrón-de-Guevara [31] analysed the existing definitions of crowdsourcing, and proposed an integrated definition. However, the definition is wordy and complex [32], thus we already adapted and simplified it into the following definition in our previous work [33], which is also used in the current study.

Crowdsourcing is defined as an online strategy, in which an organisation proposes defined task(s) to the members of the crowd via a flexible open call. By undertaking the task(s), the members contribute their work, knowledge, skills and/or experience, and receive reward. The organisation will obtain these contributions and utilize the results for the defined goals.

Literature has also showed that crowdsourcing can be utilized for different applications. Howe [15] discusses the concept of crowdsourcing through several real business applications, including iStockphoto for images exchange, InnoCentive for problem solving, and Amazon Mechanical Turk for micro tasks. Not only limited to business applications, crowdsourcing can also be applied in scientific research [34], urban planning [35], and cultural heritage [36]. Through the success of these initiatives, the literature recommends that crowdsourcing can leverage expertise, information, skills, and labour [37-39]. In particular, crowdsourcing is very helpful for tasks that need a large amount of human labour and cognitive abilities that are hard for computers to reproduce [30, 40].

Given the ability of crowdsourcing for achieving tasks that cannot be automated and need large workforce, using the crowd for building anti-virus knowledge bases seems a promising approach. Indeed several studies have reported their success on utilizing crowdsourcing for building knowledge bases in a variety of contexts [19, 20] [41]. To clarify more, here we introduce two interesting cases. The first case is the well-known encyclopaedia, Wikipedia, which is one of the most successful stories of using the crowd to perform tasks in the last decade. Since its introduction in 2001, this encyclopaedia has achieved 21 million users [42], who are both readers and contributors for contents of the encyclopaedia. It would certainly be impossible for Wikipedia to build one of the largest knowledge bases in human history without utilizing the ability of the crowd. The second case is reported by McCoy et al. [41], where crowdsourcing was applied to build a knowledge base in the medical area, which requires high knowledgeable and precise information. Utilizing the crowd for generating problem-medication pairs, these authors concluded that “crowdsourcing is an effective, inexpensive method for generating an accurate, up-to-date problem-medication knowledge base,” (p. 5) [41].

In spite of these promising capabilities, only a few studies have considered crowdsourcing in the context of virus detection, where the crowd inputs can help overcoming high false-positive rates, as discussed in Section 2.1. One of these studies is the work by Burguera et al. [43]. Focusing on the Android platform, these authors proposed a framework collecting application data to detect malwares in Android devices. Utilizing data from end-users for detecting viruses, our framework however is different from the work by Burguera et al. [43] in two aspects. First, we are not limited to Android malware, but address viruses and malware in general. Second, the users' inputs in our architecture will be processed by internal and crowd anti-virus experts, and then a machine learning algorithm is applied for building the anti-virus knowledge base.

3 Utilizing the Crowd for Building an Anti-virus Knowledge Base

In this section, we propose an architecture utilizing the crowd for building an anti-virus knowledge base. As the architecture utilizes crowdsourcing to extend the Machine Learning Approach to Anti-virus System [8], we named it CrowdMAV. This section starts by overviewing on the business process crowdsourcing of CrowdMAV, which involves three main activities (Fig. 1).

First, an *open-call* is delivered to the internet users, asking them for providing inputs to the CrowdMAV. The call is delivered to the crowd through two channels. It is posted in the official D32 anti-virus website [14], along with a delivery of a trial version of the D32 anti-virus software (D32). For existing D32 users, D32 shows a message asking their participation in the crowdsourcing process. Currently, we provide a trial (or extended) period of using D32 as an incentive for user participation.

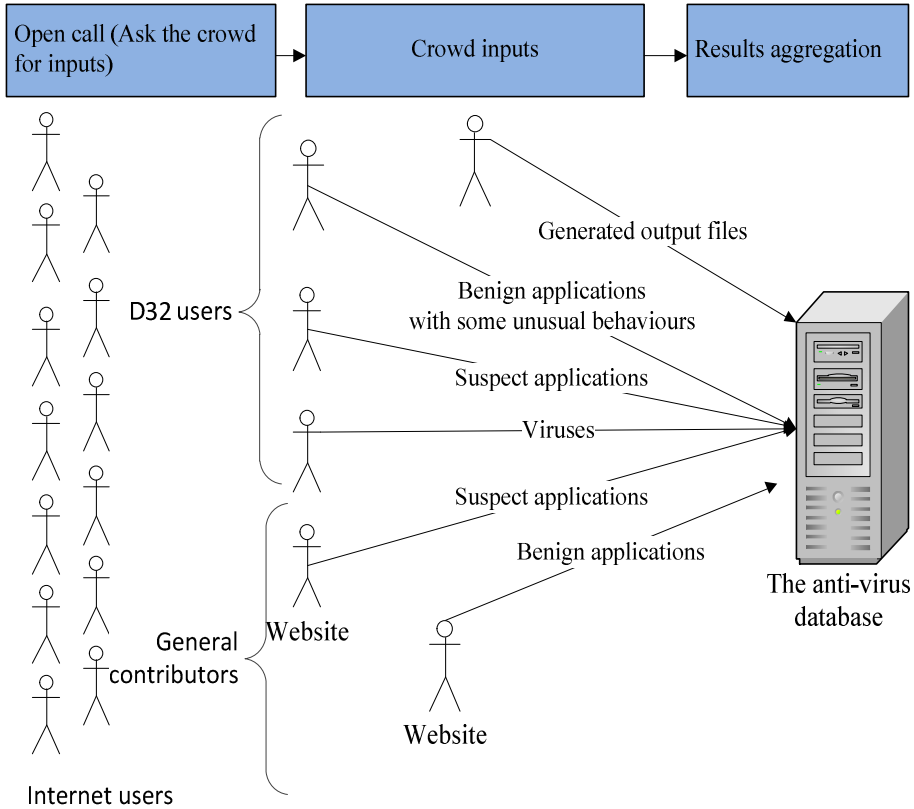


Fig. 1. Business process crowdsourcing of CrowdMAV

Second, when participating in the CrowdMAV, D32 users or general contributors *submit their inputs*. For D32 users, the anti-virus software automatically asks users to confirm the suspect behaviour. Alternatively, D32 users can also manually raise suspect application behaviour and send their inputs when they found virus-related problems, such as benign applications that were detected as viruses, slow response, or unauthorised access to a website. In these cases, these users can activate a form within D32, which then guides their submission. Other general contributors, who may use anti-virus software different from D32, can also submit their inputs through the

CrowdMAV website or send emails. We note that although the inputs can be submitted through different channels, they should include the following information: attached file(s), whether the files should be seen as virus or benign applications that were wrongly detected as a virus, the suspect behaviour, e.g. unauthorized access to a website or continuously reading the hard disk, and user messages that provide more information for the submission. For those who use D32, we additionally collect application behaviour-related data, but personal data will not be collected. We note that each user over time can provide more than one submission to the CrowdMAV.

The final activity is related to *results aggregation*. The users' inputs coming from different channels are sent to the anti-virus server. Receiving these inputs, the server processes each one as a dataset. Thus, the larger numbers of users participate in the CrowdMAV, the more datasets can be collected and processed by the server, leading to more thorough anti-virus knowledge base by the end. Aligning with the classification by [44], the server classifies the datasets into three groups: suspect, benign, and unclassified objects. This classification is based on the users' suggestions included within the inputs. If the users' inputs are consistent on whether an application is suspect or benign, this application is classified to the corresponding group. If there are conflicted user opinions on certain datasets, e.g. new software that is suggested as a malicious application by some users but seen as a benign application by other users, the datasets are grouped as unclassified objects. All groups of datasets are then updated into the database, and the CrowdMAV will use these datasets to develop the anti-virus knowledge base, as presented in the following architecture.

3.1 The Architecture of CrowdMAV

The results from the previous process are aggregated by CrowdMAV. Fig. 2 shows the architecture of the CrowdMAV, which relies on three components. The first component is *virus detection*, which is responsible for analysing the datasets received by the anti-virus server to detect viruses and malware. The virus detection is processed by the Virus Scanning Agent (VSA) that is also embedded in the D32 anti-virus software (D32) [8]. One can argue that the users have already used D32 to scan these files in their own computers before sending them to the server, and therefore rescan is unnecessary. The reason for doing it again in the architecture is that the users' D32 installation is not always up-to-date, while the VSA in CrowdMAV uses the latest anti-virus knowledge base. As a result, the VSA can automatically classify some files without further processing. In particular, the detected and classified objects are checked against the anti-virus knowledge base. Regarding the three aforementioned groups of datasets, we believe that the VSA is effective in processing the suspect group because it is regularly updated with new virus samples collected from different sources, e.g. <http://openmalware.org>. This component can also classify some files in the unclassified object group. However, this component faces difficulties when classifying the last group that were suggested as malicious applications by the anti-virus software but reported as benign files by users.

The objects that cannot be classified by the virus detection are analysed by *expert evaluation* (Component 2). As a crowdsourcing approach often returns a huge amount

of data [22], i.e. unclassified objects in our case, we combined our available experts with the experts in the crowd for evaluation. The decision on who will be considered as experts and invited to participant in the evaluation is made by the D2 company. After invited, these experts are communicated through a web application, similar to the work by [19]. This application then chooses the most frequent unclassified objects received from the users to ask for expert evaluation. It then provides a controlled environment for each expert to execute unclassified objects. Using some heuristics [45], the expert makes his decision on the classified objects. In case the decisions are consensus, it is the final decision on object classification. In case of conflicted decisions, i.e. malicious or benign, the decision supported by the majority of experts will be chosen as the final decision. The results are updated into the knowledge base. We note that the results of this component are not only classified objects, but also new suggested rules and application behaviour for detection and classification.

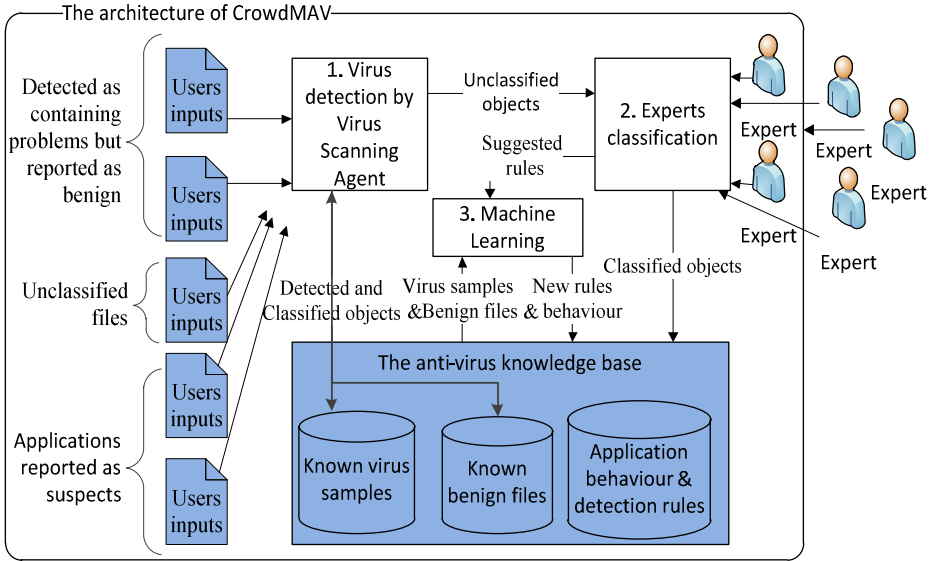


Fig. 2. The architecture of the CrowdMAV for building the anti-virus knowledge base

As a result of these two components, new objects are classified in the knowledge base. Using the updated knowledge base, the *machine learning component* (component 3) is responsible for discovering new detection rules. Extending from the work by Truong and Hoang [8], the machine learning is firstly trained using a list of known virus samples and benign files. As the list evolves, the component updates its detection rules based on the rules suggested by both internal and crowd experts. Finally, the machine learning component will be retrained over the updated knowledge base. As a result, new application behaviour and detection rules are identified and added to the anti-virus knowledge base.

4 Implementation (Work-in-Progress)

A part of this work has already been implemented. In particular, we have already developed the virus detection as an application and collect a number of virus sample that will be used for training the machine learning (Virus Scan function and Database Status in Fig. 3).

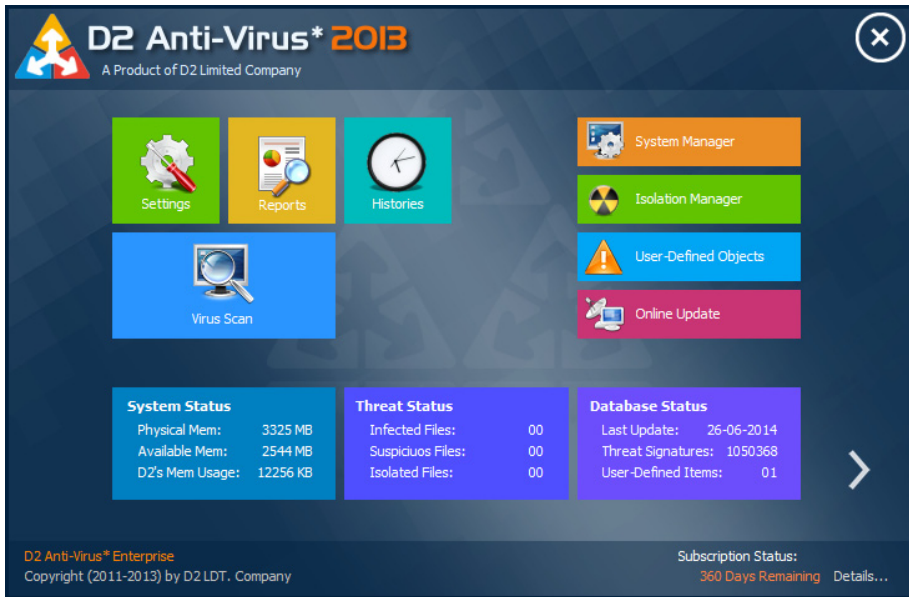


Fig. 3. Virus detection component and known virus database

Within D32, users can access the ‘reports’ function to call the submission form (Reports function in Fig. 3). The detailed submission form is presented in Fig. 4. For internet users, they can use our website for submitting suspect applications [14]. However, the form and website for collecting data need to be extended as currently they are focusing on collecting suspect files. In particular, the extension should focus on gathering data that meet the new structure of the anti-virus knowledge base, and thus should include the following data: suspect files, whether the files are virus or benign, their suspect behaviour, and messages from the users.

We also deployed the anti-virus server in order to collect users input from D32 and the website [14]. We are implementing the process required to aggregate users’ submissions. In the expert evaluation component, although we already have a team of internal experts, we need to engage more experts from the crowd for detecting viruses and classifying suspect objects. Consequently, a web application for these experts to communicate and process their classification activities is needed. Another activity that is receiving our focus is the extension of the machine learning algorithm proposed by [8] to improve its ability of learning, corresponding to the new knowledge base structures. Additionally, some parts of the current user interface in D32 website are presented in Vietnamese, and need to be translated into English.



Fig. 4. A form for submitting users' inputs

5 Conclusion and Future Work

We proposed CrowdMAV, an architecture for updating the anti-virus knowledge base utilizing the crowdsourcing strategy. Addressing the high false-positive rates found in the behaviour-based techniques [1], our architecture utilized users' inputs in order to extend the anti-virus knowledge base. The users' inputs are first classified by the VSA and then by expert evaluation, in which we also suggested a crowdsourcing approach by combining internal and crowd experts. The classified data are mined by the machine learning algorithm [8], resulting in new rules and new objects classified in the knowledge base. By doing so, the knowledge base includes new information learned from the crowd, and as a result, the system can correctly diagnose and classify objects that could not be classified previously.

In comparison to other studies [3, 4, 11, 25], our work has a similar aim to improve the effectiveness of behaviour-base detection techniques. However, we chose a different focus. While other works focus on analysing application behaviour to detect benign and malicious applications [3, 4, 11, 25], our architecture aims at building a comprehensive knowledge base, necessary for the analysis. Although this focus has received attention from several anti-virus vendors, such as Norton anti-virus from Symantec [46] and Malware Protection Centre from Microsoft [47], only a few studies examined the use of crowdsourcing to improve anti-virus knowledge base (e.g. [43]). Our architecture fulfils this gap by proposing an architecture considering the role of crowdsourcing for building the anti-virus knowledge base.

From a practical point of view, our framework, by enriching the anti-virus knowledge base, can improve the performance of the behaviour-based technique. To an extent, although our architecture is targeted to the D32 anti-virus software, we believe that this approach can also be applied to other anti-virus systems. It can also be combined with other detection techniques, i.e. signature-based techniques. As seen via Fig. 2, our architecture also updates the known virus database, which can enrich the virus sample for the signature-based techniques.

Given that this is a work in progress, we acknowledge that future studies are needed to solidify the architecture. In particular, the future work should be seen from two perspectives: crowdsourcing and anti-virus systems. From crowdsourcing point of view, the business process crowdsourcing in the current study addressed three activities (Fig. 1). We understand that business process crowdsourcing involves other aspects, such as incentive mechanism, crowd management, and workflow design [48], which need to be considered in further development of our architecture. From an anti-virus systems' view, we need to complete the machine learning algorithm with the new knowledge base and test the effectiveness of the architecture. This effectiveness should base on not only the number of correctly recognized viruses (true positive and true negative rates) but also false-positive and false negative rates, which reflect the current limitation of the behaviour-based techniques [1].

References

1. Sukwong, O., Kim, H.S., Hoe, J.C.: Commercial antivirus software effectiveness: an empirical study. *Computer* 44(3), 0063–0070 (2011)
2. Rad, B.B., Masrom, M., Ibrahim, S.: Evolution of computer virus concealment and anti-virus techniques: a short survey. arXiv preprint arXiv:1104.1070 (2011)
3. Park, Y., Reeves, D.S., Stamp, M.: Deriving common malware behavior through graph clustering. *Computers & Security* 39, 419–430 (2013)
4. Bayer, U., et al.: Scalable, Behavior-Based Malware Clustering. In: NDSS. Citeseer (2009)
5. Egele, M., et al.: A survey on automated dynamic malware-analysis techniques and tools. *ACM Computing Surveys (CSUR)* 44(2), 6 (2012)
6. Lin, D., Stamp, M.: Hunting for undetectable metamorphic viruses. *Journal in Computer Virology* 7(3), 201–214 (2011)
7. Hu, X., Chiueh, T.-C., Shin, K.G.: Large-scale malware indexing using function-call graphs. In: Proceedings of the 16th ACM Conference on Computer and Communications Security, pp. 611–620. ACM, Chicago (2009)
8. Truong, M.N.Q., Hoang, T.N.: A multi-agent mechanism in machine learning approach to anti-virus system. In: Nguyen, N.T., Jo, G.-S., Howlett, R.J., Jain, L.C. (eds.) KES-AMSTA 2008. LNCS (LNAI), vol. 4953, pp. 743–752. Springer, Heidelberg (2008)
9. Rieck, K., Holz, T., Willems, C., Düssel, P., Laskov, P.: Learning and Classification of Malware Behavior. In: Zamboni, D. (ed.) DIMVA 2008. LNCS, vol. 5137, pp. 108–125. Springer, Heidelberg (2008)
10. Microsoft. Evolution of Malware (2014), http://www.microsoft.com/security/sir/story/default.aspx#!10_year_malware

11. Yin, H., et al.: Panorama: capturing system-wide information flow for malware detection and analysis. In: Proceedings of the 14th ACM Conference on Computer and Communications Security. ACM (2007)
12. Stinson, E., Mitchell, J.C.: Characterizing Bots' Remote Control Behavior. In: Hämmerli, B.M., Sommer, R. (eds.) DIMVA 2007. LNCS, vol. 4579, pp. 89–108. Springer, Heidelberg (2007)
13. Schultz, M.G., et al.: Data mining methods for detection of new malicious executables. In: Proceedings of 2001 IEEE Symposium on Security and Privacy, S&P 2001 (2001)
14. D32. D32 Anti-virus (2014), <http://www.d32av.vn/>
15. Howe, J.: The rise of crowdsourcing. *Wired Magazine*, 1–4 (2006)
16. Muntés-Mulero, V., Paladini, P., Manzoor, J., Gritti, A., Larriba-Pey, J.-L., Mijndhardt, F.: Crowdsourcing for industrial problems. In: Nin, J., Villatoro, D. (eds.) CitiSens 2012. LNCS, vol. 7685, pp. 6–18. Springer, Heidelberg (2013)
17. Chi, E.H., Bernstein, M.S.: Leveraging Online Populations for Crowdsourcing. *IEEE Internet Computing* 16(5), 10–12 (2012)
18. Zhao, Y., Zhu, Q.: Evaluation on crowdsourcing research: Current status and future direction. *Information Systems Frontiers*, 1–18 (2012)
19. Vukovic, M., Laredo, J., Rajagopal, S.: Challenges and experiences in deploying enterprise crowdsourcing service. In: Benattallah, B., Casati, F., Kappel, G., Rossi, G. (eds.) ICWE 2010. LNCS, vol. 6189, pp. 460–467. Springer, Heidelberg (2010)
20. Fraternali, P., et al.: Putting humans in the loop: Social computing for Water Resources Management. *Environmental Modelling & Software* 37, 68–77 (2012)
21. Corney, J., et al.: Putting the crowd to work in a knowledge-based factory. *Advanced Engineering Informatics* 24(3), 243–250 (2010)
22. Doan, A., Ramakrishnan, R., Halevy, A.Y.: Crowdsourcing systems on the world-wide web. *Communications of the ACM* 54(4), 86–96 (2011)
23. Cohen, F.: Computer viruses: theory and experiments. *Computers & Security* 6(1), 22–35 (1987)
24. Clause, J., Li, W., Orso, A.: Dytan: a generic dynamic taint analysis framework. In: Proceedings of the, International Symposium on Software Testing and Analysis. ACM (2007)
25. Willems, C., Holz, T., Freiling, F.: Toward automated dynamic malware analysis using cwsandbox. *IEEE Security and Privacy* 5(2), 32–39 (2007)
26. Hu, Y., et al.: Unknown malicious executables detection based on run-time behavior. In: Fifth International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2008. IEEE (2008)
27. Lanzi, A., Sharif, M.I., Lee, W.: K-Tracer: A System for Extracting Kernel Malware Behavior. In: NDSS (2009)
28. Rouse, A.C.: A preliminary taxonomy of crowdsourcing. In: Proceedings of the 21st Australasian Conference on Information Systems, pp. 1–10 (2010)
29. Kittur, A., Chi, E.H., Suh, B.: Crowdsourcing user studies with Mechanical Turk. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM (2008)
30. Sarasua, C., Simperl, E., Noy, N.F.: CROWDMAP: Crowdsourcing ontology alignment with microtasks. In: Cudré-Mauroux, P., Heflin, J., Sirin, E., Tudorache, T., Euzenat, J., Hauswirth, M., Parreira, J.X., Hendler, J., Schreiber, G., Bernstein, A., Blomqvist, E. (eds.) ISWC 2012, Part I. LNCS, vol. 7649, pp. 525–541. Springer, Heidelberg (2012)
31. Estellés-Arolas, E., González-Ladrón-de-Guevara, F.: Towards an integrated crowdsourcing definition. *Journal of Information Science* 38(2), 189–200 (2012)

32. Brabham, D.C.: *Crowdsourcing*. The MIT Press, Cambridge (2013)
33. Thuan, N.H., Antunes, P., Johnstone, D.: Factors Influencing the Decision to Crowdsource. In: Antunes, P., Gerosa, M.A., Sylvester, A., Vassileva, J., de Vreede, G.-J. (eds.) *CRIWG 2013*. LNCS, vol. 8224, pp. 110–125. Springer, Heidelberg (2013)
34. Mason, W., Suri, S.: Conducting behavioral research on Amazon’s Mechanical Turk. *Behavior Research Methods* 44(1), 1–23 (2012)
35. Brabham, D.C.: Motivations for Participation in a Crowdsourcing Application to Improve Public Engagement in Transit Planning. *Journal of Applied Communication Research* 40(3), 307–328 (2012)
36. Kingston, A.: “Choir attempted that beautiful anthem “Oh, Radiant Morn” – made a hash of it” - Making a hash of the Adkin Diary transcriptions. In: *Workshop on Crowdsourcing for the Digital Humanities and Cultural Heritage Sector*, Wellington, New Zealand (2013)
37. Brabham, D.C.: Crowdsourcing as a Model for Problem Solving: An Introduction and Cases. *Convergence: The International Journal of Research into New Media Technologies* 14(1), 75–90 (2008)
38. Vukovic, M., Bartolini, C.: Towards a research agenda for enterprise crowdsourcing. In: Margaria, T., Steffen, B. (eds.) *ISoLA 2010, Part I*. LNCS, vol. 6415, pp. 425–434. Springer, Heidelberg (2010)
39. Aitamurto, T., Leiponen, A., Tee, R.: The Promise of Idea Crowdsourcing—Benefits, Contexts, Limitations, in *White Paper for Nokia IdeasProject* (June 2011)
40. Franklin, M.J., et al.: CrowdDB: answering queries with crowdsourcing. In: *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, pp. 61–72. ACM, Athens (2011)
41. McCoy, A.B., et al.: Development and evaluation of a crowdsourcing methodology for knowledge base construction: identifying relationships between clinical problems and medications. *Journal of the American Medical Informatics Association* 19(5), 713–718 (2012)
42. Wikipedia. *Statistics* (2014),
<http://en.wikipedia.org/wiki/Special:Statistics> (cited June 2014)
43. Burguera, I., Zurutuza, U., Nadjm-Tehrani, S.: Crowdroid: behavior-based malware detection system for android. In: *Proceedings of the 1st ACM Workshop on Security and Privacy in Smartphones and Mobile Devices*. ACM (2011)
44. Saeed, I.A., et al.: A Survey on Malware and Malware Detection Systems. *Analysis* 3(10), 13–17 (2013)
45. Adkins, F., et al.: Heuristic malware detection via basic block comparison. In: *2013 8th International Conference on Malicious and Unwanted Software: The Americas (MALWARE)*. IEEE (2013)
46. Symantec. *Submit Virus Samples* (June 2014),
http://www.symantec.com/security_response/submit.samples.jsp
47. Microsoft. *Submit a sample* (June 2014),
<https://www.microsoft.com/security/portal/submission/submit.aspx>
48. Thuan, N.H., Antunes, P., Johnstone, D.: Toward a Nexus Model Supporting the Establishment of Business Process Crowdsourcing. In: Dang, T.K., Wagner, R., Neuhold, E., Takizawa, M., Küng, J. (eds.) *FDSE 2014*. LNCS, vol. 8860, pp. 136–150. Springer, Heidelberg (2014)

Two-Way Biometrics-Based Authentication Scheme on Mobile Devices

Duong-Tien Phan¹, Toan-Thinh Truong¹, Minh-Triet Tran¹,
and Anh-Duc Duong²

¹ Faculty of Information Technology, University of Science, VNU-HCM
1112328@student.hcmus.edu.vn,
{ttthinh, tmtriet}@fit.hcmus.edu.vn

² Faculty of Software Engineering, University of Information Technology, VNU-HCM
ducda@uit.edu.vn

Abstract. Online transactions with mobile devices through internet environment have become popular worldwide. Therefore, many authentication schemes have been proposed to protect users from various potential attacks in e-transactions with online service providers from mobile devices. In 2013, Khan et. al. propose a key-hash based scheme on mobile devices to resist known kinds of attacks that previous schemes cannot resist. However, we prove that Khan et. al.'s scheme still cannot withstand impersonation, denial of service, and three-factor attacks. This motivates our proposal of an improved scheme to further overcome the found limitations in Khan's scheme. The main idea of our proposed method is that the user ID and the secret key of the server are hashed together to prevent user impersonation. We also prove that our method can also resist against known attacks, such as server and user impersonation attack, replay attack, password guessing attack, malicious user attack, mobile device loss attack, attacks due to ID theft, attacks using login request.

Keywords: Authentication, identity, mobile device, session key, biometrics.

1 Introduction

Mobile devices allow users to access many different types of online services and data at any time and where. This enhances the convenience for users over traditional means of interaction via personal computers and opens a new trend of smart interactive environment. Users can now receive useful information and assistance corresponding to their current external contexts from their own mobile devices. Thus new generations of applications for mobile devices, as well as wearable devices, have been developed, such as location-based services[1], mobile social networks[2], augmented reality applications[3], mobile commerce systems[4].

To ensure users' privacy and security when users access online resource from their mobile devices, security issues should be carefully considered. Authentication is one of the most important problems in security and privacy. This is the entry step to guarantee that a mobile device or online system only provides personal sensitive information to a valid authorized user. This motivates researchers to propose different

methods and protocols for authentication, especially for mobile devices to access remote systems.

In fact, mobile devices, especially wearable devices, have certain limitations and constraints for authentication processes, such as limited power consumption and processing speed. As a result of this, besides a common approach for authentication with bilinear pairing[5], elliptic curve[6], there is another trend of light-weight authentication schemes[7][8] with low computational cost such as hash functions together with randomness and XOR operations.

In this paper, we analyze the drawbacks and limitations of the scheme proposed by Khan et. al. [9]. The original scheme has several advantages such as providing session key establishment[10], quick wrong password detection[11], freedom to change password[12], user's anonymity[13], withstanding replay attack[14] password guessing attack[15], and using low-cost computations[16] to save energy and increase using-time of mobile devices. However Khan et. al.'s scheme cannot withstand impersonation[17], denial of service[18], and three-factor[19] attacks. Furthermore, we also analyze the limitation of storing redundancy information on mobile device in their scheme. Thus, we propose a new scheme to overcome the drawbacks of Khan et. al.'s scheme.

There are two main principles in our proposed scheme. First, we use random values combined with secret keys that a server shares with users to generate different secret keys at different registration times. Second, a user's identity is hashed with the server's secret value to prevent an arbitrary malicious user from impersonating others. The authentication phase applies the three-way challenge-handshake technique with random values to better withstand replay attacks[20].

The rest of our paper is organized as follows. In section 2, we review Khan et. al.'s scheme, and point out some of the weaknesses existing in this scheme. Then in Section 3 we propose our improved version for the method proposed by Khan et. al. In section 4, we analyze our proposed scheme on two aspects, namely security and efficiency. Finally, the conclusion is presented in Section 5.

2 Review and Cryptanalysis of Khan et. al.'s Scheme

In this section, we review Khan et. al.'s scheme and prove their scheme cannot resist impersonation, denial of service, and three-factor attacks.

2.1 Review of Khan et. al.'s Scheme

Khan et. al.'s scheme includes four phases: registration, login, mutual authentication with session key agreement, and password-change phases. Below are some notations used in this scheme:

- U_i : i^{th} user.
- ID_i : U_i 's unique identity.
- PW_i : U_i 's unique password.
- F_i : U_i 's fingerprint.
- S : Server.

- IDS : S 's secret identity.
- x : S 's secret key.
- $h(\cdot)$: One-way hash function.
- $h_k(\cdot)$: One-way keyed hash function.
- N : Random value.
- \oplus : XOR bit operation.
- \otimes : NOR bit operation.
- \parallel : String concatenation operation.

Registration Phase

When U_i registers with S , U_i submits ID_i , $h(PW \parallel N)$, and F_i to S through a secure channel, where $h(PW \parallel N)$ is a masked password. Figure 1 demonstrates this phase.

- Step 1: Upon receiving U_i 's request, S computes $hpw = h(PW \parallel N) \otimes F_i$, $B_i = hpw \oplus e$, $C_i = hpw \oplus h(x \parallel IDS)$, $E_i = hpw \oplus h(x \parallel e \parallel IDS)$, $R_i = h(ID_i \oplus h(x \parallel e \parallel IDS)) \oplus hpw$, and $V_i = h_{h(ID_i \oplus h(x \parallel e \parallel IDS))}(F_i)$.
- Step 2: S returns $\{B_i, C_i, E_i, R_i, V_i, h(\cdot), h_k(\cdot)\}$ to U_i 's mobile device through a secure channel.
- Step3: U_i stores information from S and $ID_i \oplus N$ into U_i 's mobile device.

In this phase, the scheme has two advantages. First, a user can hide his/her password by sending hash value $h(PW \parallel N)$ instead of PW to S . Second, finger prints are used to enhance security. Therefore, our scheme inherits these ideas of Khan et. al.'s scheme.

However, this scheme still has a potential vulnerability for impersonate attacks because of the design of B_i , C_i , E_i , and R_i . (c.f. Section 2.2). If a malicious legal user U_A knows the ID_i of another user U_i , U_A can then impersonate U_i .

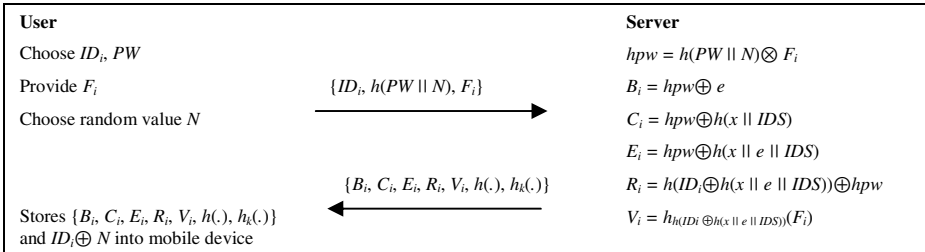


Fig. 1. Registration phase of Khan et. al.'s scheme

Login Phase

As illustrated in Figure 2, after a successful registration, U_i can login to S . U_i inputs ID_i , PW , and F_i into U_i 's mobile device:

- Step1: Mobile device computes $N = (ID_i \oplus N) \oplus ID_i$, $hpw = h(PW \parallel N) \otimes F_i$, and $A_i = R_i \oplus hpw$. If $h_{A_i}(F_i) = V_i$, it allows the user to go to the next step; otherwise, it terminates the session.
- Step2: Mobile device computes $e = hpw \oplus B_i$ and chooses N_U . Then, it computes $C_1 = N_U \oplus E_i \oplus hpw$, $RCID = ID_i \oplus h(N_U) \oplus e$, $C_2 = B_i \oplus C_i$, and $C_3 = h_{A_i}(N_U \parallel e)$.

- Step3: Mobile device sends a login request $\{RCID, C_1, C_2, C_3\}$ to the server S through a common channel.

A mobile device chooses a random value N_U to challenge S , and makes $RCID$ be dynamic in each login to provide anonymity. However, $RCID$ is computed from ID_i , which is not compelled with the secret key. Therefore, a user can easily change his/her identity to impersonate another valid user. By combining ID_i with S 's secret value in a hash function, our proposed method can prevent malicious users from changing their real identity.

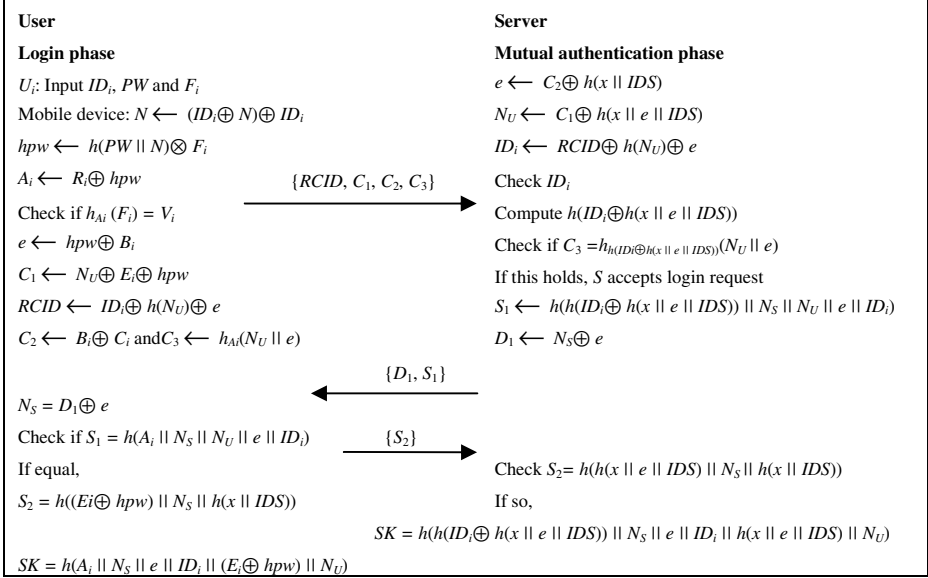


Fig. 2. Login, mutual authentication and session key agreement phases of Khan et. al.'s scheme

Mutual Authentication with Session Key Agreement Phase

After receiving the login request $\{RCID, C_1, C_2, C_3\}$ from U_i , S and U_i perform the following steps for mutual authentication. Figure 2 demonstrates this phase.

- **Step1:** S computes $e = C_2 \oplus h(x \parallel IDS)$, $N_U = C_1 \oplus h(x \parallel e \parallel IDS)$, $ID_i = RCID \oplus h(N_U) \oplus e$, and checks the validity of ID_i . Then, S computes $h(ID_i \oplus h(x \parallel e \parallel IDS))$ and compares C_3 with $h_{h(ID_i \oplus h(x \parallel e \parallel IDS))}(N_U \parallel e)$. If this holds, S accepts U_i 's login request; otherwise, S terminates the session. Next, S chooses N_S and computes $S_1 = h(h(ID_i \oplus h(x \parallel e \parallel IDS)) \parallel N_S \parallel N_U \parallel e \parallel ID_i)$, $D_1 = N_S \oplus e$. S sends $\{D_1, S_1\}$ to U_i .
- **Step2:** When receiving $\{D_1, S_1\}$, U_i computes $N_S = D_1 \oplus e$. Then, U_i compares S_1 with $h(A_i \parallel N_S \parallel N_U \parallel e \parallel ID_i)$. If they are not identical, U_i terminates the session; otherwise, U_i computes $S_2 = h((E_i \oplus hpw) \parallel N_S \parallel h(x \parallel IDS))$ and sends $\{S_2\}$ to S .
- **Step 3:** When receiving $\{S_2\}$ from U_i , S checks if $S_2 = h(h(x \parallel e \parallel IDS) \parallel N_S \parallel h(x \parallel IDS))$. If this does not hold, S terminates the session; otherwise S computes session key $SK = h(h(ID_i \oplus h(x \parallel e \parallel IDS)) \parallel N_S \parallel e \parallel ID_i \parallel h(x \parallel e \parallel IDS) \parallel N_U)$. Similarly, U_i computes session key $SK = h(A_i \parallel N_S \parallel e \parallel ID_i \parallel (E_i \oplus hpw) \parallel N_U)$.

In this phase, the server not only checks login request but also generates a random value to re-challenge the user. Furthermore, this phase has a session key agreement step to establish a secure channel after the successful authentication. Therefore, our scheme also inherits this idea of Khan et. al..

Password Change Phase

This phase help users change their passwords. In this phase, a user can also replace the old values of N and F_i by the new values of N^* and F_i^* (c.f. Figure 3).

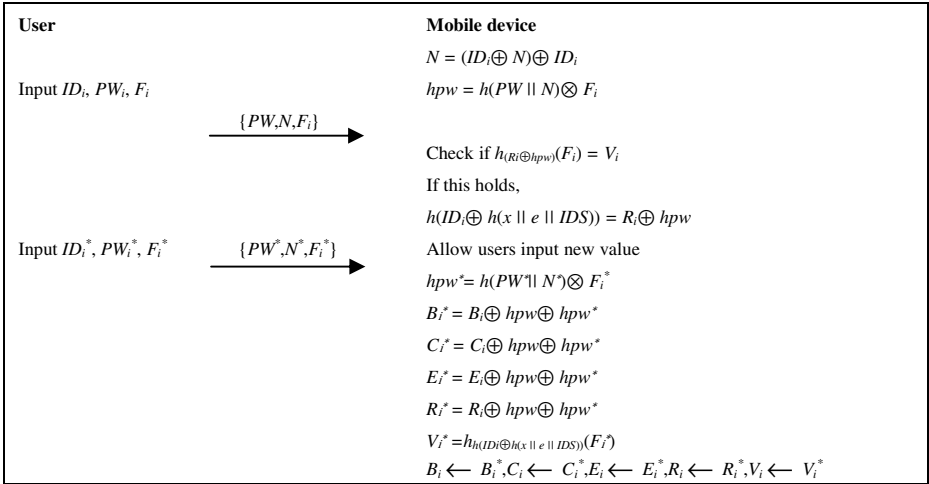


Fig. 3. Password change phase in Khan et. al.’s scheme

- **Step1:** U_i inputs $ID_i, PW_i,$ and F_i into his or her own mobile device and requests to change U_i ’s password, random value N , and fingerprint F_i .
- **Step 2:** Mobile device extracts $N = (ID_i \oplus N) \oplus ID_i$. Then, it computes $h_{pw} = h(PW \parallel N) \otimes F_i$, and checks if $h_{(R_i \oplus h_{pw})}(F_i)$ and V_i are equal. If this does not hold, mobile device terminates this session. Otherwise, it extracts $h(ID_i \oplus h(x \parallel e \parallel IDS)) = R_i \oplus h_{pw}$ and the user U_i can now input new values. U_i chooses new PW^*, N^* and, fingerprint F_i^* (e.g. another finger of U_i). Finally, U_i inputs $\{PW^*, N^*\}$ and F_i^* to mobile device.
- **Step3:** Mobile device computes $h_{pw}^* = h(PW^* \parallel N^*) \otimes F_i^*, B_i^* = B_i \oplus h_{pw} \oplus h_{pw}^*, C_i^* = C_i \oplus h_{pw} \oplus h_{pw}^*, E_i^* = E_i \oplus h_{pw} \oplus h_{pw}^*, R_i^* = R_i \oplus h_{pw} \oplus h_{pw}^*, V_i^* = h_{h(ID_i \oplus h(x \parallel e \parallel IDS))}(F_i^*)$; and replaces B_i, C_i, E_i, R_i, V_i with $B_i^*, C_i^*, E_i^*, R_i^*, V_i^*$.

In this phase, only a legitimate user U_i can change his/her password (together with the random value and fingerprint) because this phase requires $ID_i, PW_i,$ and F_i of U_i . Furthermore, U_i can change password without any server’s assistance. Therefore, our scheme also inherits these ideas from Khan et. al.’s scheme.

2.2 Cryptanalysis of Khan et. al.'s Scheme

In this phase, we prove that Khan et. al.'s scheme is vulnerable to impersonation, denial of service, and three-factor attacks.

User Impersonation due to Leaking Identity

In an authentication scheme, it is commonly that the identity of a user is not necessarily kept secret. Only personal sensitive information for authentication, such as password, should be secure.

However in Khan et. al.'s scheme, a malicious U_A can easily impersonate U_i if the ID_i of U_i is known. The following scenario illustrates the steps for U_A to generate messages for authentication with his or her own information and ID_i to trick the server S without the mobile device, valid password, and fingerprint of U_i .

- U_A computes $N_A = (ID_A \oplus N_A) \oplus ID_A$, $hpw_A = h(PW_A \parallel N_A) \otimes F_A$, $h(x \parallel IDS) = C_A \oplus hpw_A$, $h(x \parallel e_A \parallel IDS) = E_A \oplus hpw_A$ and $e_A = B_A \oplus hpw_A$.
- U_A generates N_U , computes $C_1 = N_U \oplus h(x \parallel e_A \parallel IDS)$, $RCID = ID_i \oplus h(N_U) \oplus e_A$, $C_2 = e_A \oplus h(x \parallel IDS)$ and $C_3 = h_{h(ID_i \oplus h(x \parallel e_A \parallel IDS))}(N_U \parallel e_A)$, and sends $\{RCID, C_1, C_2, C_3\}$ to S for login request.
- S computes $h(ID_i \oplus h(x \parallel e_A \parallel IDS))$ and compares C_3 with $h_{h(ID_i \oplus h(x \parallel e_A \parallel IDS))}(N_U \parallel e_A)$. As the two values are equal, S accepts the login request and computes $S_1 = h(h(ID_i \oplus h(x \parallel e_A \parallel IDS)) \parallel N_S \parallel N_U \parallel e_A \parallel ID_i)$ and $D_1 = N_S \oplus e_A$. S sends $\{D_1, S_1\}$ to U_A .
- U_A computes $N_S = D_1 \oplus e_A$, $S_2 = h(h(x \parallel e_A \parallel IDS) \parallel N_S \parallel h(x \parallel IDS))$, and sends $\{S_2\}$ to S .
- S compares S_2 with $h(h(x \parallel e_A \parallel IDS) \parallel N_S \parallel h(x \parallel IDS))$. As the two values are identical, S and U_A can now compute the common session key $SK = h(h(ID_i \oplus h(x \parallel e_A \parallel IDS)) \parallel N_S \parallel e_A \parallel ID_i \parallel h(x \parallel e_A \parallel IDS) \parallel N_U)$.

Server and User Impersonation When Catching Login Message and Leakage of Mobile Device's Information

When proving security, Khan et. al. claimed that it is impossible to co-relate corresponding login message and lost mobile device. This is not true because any valid user can easily perform this co-relation and impersonate another user and a server. A malicious valid user U_A can extract $h(x \parallel IDS) = C_A \oplus hpw_A$ (this is important information which the server shares for all users). If B_i , C_i , E_i , R_i , V_i and $ID_i \oplus N$ in U_i 's mobile device are leaked and U_A knows this information, U_A easily impersonates user and server. U_A computes $e_i = B_i \oplus C_i \oplus h(x \parallel IDS)$ and $h(x \parallel e_i \parallel IDS) = E_i \oplus C_i \oplus h(x \parallel IDS)$. Now, U_A can co-relate mobile device's information and U_i 's corresponding login message by computing $e^* = C_2 \oplus h(x \parallel IDS)$. If $e^* = e_i$, login message belongs to U_i . As well, U_A computes $N_U = C_1 \oplus h(x \parallel e_i \parallel IDS)$, $ID_i = RCID \oplus h(N_U) \oplus e_i$. Knowing ID_i , U_A can easily impersonate U_i .

Furthermore, when user U_A knows e_i , $h(x \parallel e_i \parallel IDS)$, and $h(x \parallel IDS)$, U_A can also impersonate S to cheat U_i . To set up such attack, a malicious user U_A first captures and drops the legal login message $\{RCID, C_1, C_2, C_3\}$ of U_i . U_A then determines the sender of this login message by computing $e^* = C_2 \oplus h(x \parallel IDS)$ and comparing this

value with the known value of e_i inferred from leaked information of the mobile device of U_i . U_A re-computes N_U and ID_i . Finally, U_A completely re-computes C_3 and S_1 , D_1 and SK to impersonate S to cheat U_i .

Three-Factor Attack

This kind of attack implies that attacker has two of the three factors: password, mobile device, and fingerprint. In Khan et. al.'s scheme, if ID_i is leaked, it is sufficient for the corresponding user U_i to be impersonated even when no more factors are leaked. Therefore, Khan et. al.'s scheme cannot resist this kind of attack.

Denial of Service Attack

Khan proposed storing ID_i until the end of the session to resist parallel session attack. If session remains, server checks identity's existence in temporary list and denies any login request from this identity. In case of attacker's successfully impersonation, he/she can block legitimate user by procrastinating.

3 Proposed Scheme

In this section, we propose an improved scheme to resolve existing limitations in Khan et. al.'s scheme. Our scheme inherits the advantages and modifies some aspects in Khan's design to enhance security. Before more detail, we present main ideas in our scheme. In registration phase, main objective is providing users with two secret key sh ($IDS \parallel h(e) \parallel x$) and $h(ID_i \parallel x \parallel h(e) \parallel IDS)$, where random value e is used to resist re-registration attack and ID_i is combined with the secret key in a hash function to prevent a valid user from impersonating others. In login phase, we use two random values N_U (hidden in C_1) and N_S (hidden in S_1) for mutual challenge between user and server. Furthermore, random value N_U is used to change user's identity at each login. After successfully authentication, both sides share common session key for data encryption. Our scheme includes four phases: registration, login, and mutual authentication and password-change phases. However, we don't present password-change phase because this phase and Khan's scheme are the same.

3.1 Registration Phase

This phase has four basic requirements: channel between users and server must be secure; true password must be kept secret, even with the server; secret keys which the server shares for each user must be different; the true identity ID_i must be compelled with the server's secret key by a hash function. We can see Khan's scheme has first three requirements but not the last. Therefore, our scheme will implement the last requirement to achieve the better phase. When U_i registers with S , U_i chooses ID_i , F_i , and computes $hpw = h(PW \parallel N \parallel F_i)$. Then, U_i sends $\{ID_i, hpw, F_i\}$ to S . Figure 4 demonstrates this phase.

- Step1: When receiving registration message from U_i , S generates random value e to make different secret key at different time.

- Step2: S computes $B_i = h(hpw \parallel F_i \parallel ID_i) \oplus e$, $C_i = h(ID_i \parallel hpw \parallel F_i) \oplus h(IDS \parallel h(e) \parallel x)$, $E_i = h(F_i \parallel hpw \parallel ID_i) \oplus h(ID_i \parallel x \parallel h(e) \parallel IDS)$, $V_i = h_{h(ID_i \parallel x \parallel h(e) \parallel IDS)}(hpw)$, where B_i includes random value e , C_i and E_i includes secret key shared by S , V_i is a password-verification value.
- Step3: S sends $\{B_i, C_i, E_i, V_i, h(\cdot), h_k(\cdot)\}$ to U_i . Once receiving, U_i stores these information and N into mobile device.

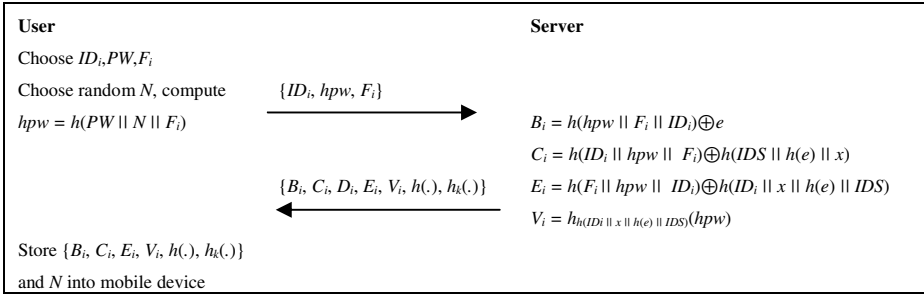


Fig. 4. Registration phase of proposed scheme

3.2 Login Phase

U_i inputs ID_i, PW_i and F_i to login. Then, mobile device performs:

- Step 1: Compute $hpw = h(PW \parallel N \parallel F_i)$, $A_i = h(F_i \parallel hpw \parallel ID_i) \oplus E_i$, compare $h_{A_i}(hpw)$ with V_i . If these two values are not equal, mobile device terminates the session; otherwise, user goes to next step.
- Step 2: Mobile device generates random value N_U and computes $e = h(hpw \parallel F_i \parallel ID_i) \oplus B_i$, $C_0 = h(ID_i \parallel hpw \parallel F_i) \oplus C_i \oplus N_U$, $C_1 = h_{A_i}(N_U)$, $RCID = ID_i \oplus h(N_U)$, where $RCID$ is dynamic identity; C_0 includes random values N_U ; C_1 includes secret key combined with random value for S 's verification of user's validity.
- Step 3: Mobile device sends $\{e, RCID, C_0, C_1\}$ to S to login.

3.3 Mutual Authentication with Session Key Agreement Phase

In this phase, S uses two secret keys x and IDS to extract random value N_U . S not only verifies user's validity through information provided by the user but also generates random value N_S to challenge the user. After successfully authentication, both sides compute common session key for data encryption.

When receiving login message $\{h(e), RCID, C_0, C_1\}$ from U_i , S and mobile device perform the following steps to mutual authenticate. Figure 5 demonstrates this phase.

- Step 1: S extracts $N_U^* = C_0 \oplus h(IDS \parallel h(e) \parallel x)$, $ID_i^* = RCID \oplus h(N_U^*)$. Then, S verifies ID_i^* 's validity. Next, S computes $A_i^* = h(ID_i^* \parallel x \parallel h(e) \parallel IDS)$ and compares $h_{A_i^*}(N_U^*)$ with C_1 . If two values are equal, S generates random value N_S and computes $S_1 = h(A_i^* \parallel N_U^* \parallel N_S \parallel h(e) \parallel ID_i^*)$ and sends $\{N_S, S_1\}$ to U_i .

- Step 2: When mobile device receives $\{N_S, S_1\}$ from S , it computes and compares $h(A_i \parallel N_U \parallel N_S \parallel h(e) \parallel ID_i)$ with S_1 . If two values are equal, it computes $S_2 = h(N_U \parallel ID_i \parallel A_i \parallel N_S \parallel h(e))$ to respond S 's challenge.
- Step 3: When receiving S_2 from U_i , S computes and compares $h(N_U^* \parallel ID_i^* \parallel A_i^* \parallel N_S \parallel h(e))$ with S_2 . If two values are equal, S accepts user's login message. S computes common session key $SK = h(N_U^* \parallel N_S \parallel h(e) \parallel ID_i^* \parallel h(A_i^*))$ and mobile device also computes session key $SK = h(N_U \parallel N_S \parallel h(e) \parallel ID_i \parallel h(A_i))$.

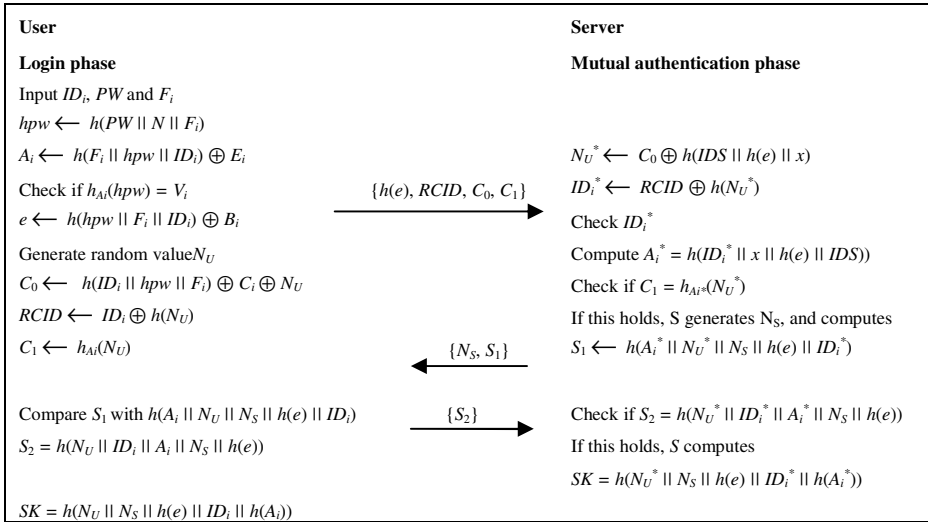


Fig. 5. Login, mutual authentication and session key agreement phases of proposed scheme

4 Security and Efficiency Analysis

In this section, we prove our scheme is more secure than Khan et. al.'s scheme and can resist many kinds of attacks. Moreover, our scheme is also suitable for resource-limited mobile device.

4.1 Security Analysis

Secure authentication scheme not only resists impersonation attack, but also protects communication between users and server. We always assume that information transmitted in a common channel and attacker has total control over this channel. Furthermore, data stored in mobile device may be leaked and attacker can extract information from lost or stolen mobile device. Therefore, we must predict many circumstances to design the secure scheme. In this section, we explain what proposed scheme is different from and why it is more secure than Khan et. al.'s. Afterwards, we analyze some kinds of attacks which Khan et. al.'s scheme cannot resist, as well as some different popular ones in proposed scheme.

If there is much information stored in a mobile device, an attacker has more chance to exploit these clues. Hence, unnecessary information should not be saved into users' mobile devices. In Khan et. al.'s scheme, R_i is redundant because a user can re-compute $h(ID_i \oplus h(x \parallel e \parallel IDS))$ by extracting $h(x \parallel e \parallel IDS)$ from E_i , then combining with ID_i . In our scheme, R_i is totally removed to avoid storing waste data on client-side.

The next weakness in Khan et. al.'s scheme is that all information shared by a server and users is concealed by performing XOR with hpw . With this method, an attacker only needs to perform XOR operation with the two values in B_i, C_i, E_i, R_i to remove hpw and achieve other information. In our scheme, information shared by the server is hidden by different values. We hash hpw with ID_i and F_i , and values' order in hash function is changed to make different hash values. Therefore, performing XOR any two values does not obtain any clues which server shares for users.

Constant information shared with all users may lead to security issues for the authentication process. Khan et. al.'s scheme proposes storing $C_i = hpw \oplus h(x \parallel IDS)$, where $h(x \parallel IDS)$ is constant value shared with all users and concealed by operation of XOR with hpw . If information stored in the mobile device is leaked, an attacker who is another valid user can extract U_i 's hpw by performing XOR C_i with $h(x \parallel IDS)$ —extracted from information shared by server with him/her. Hence, the attacker computes user's hpw without knowing password, fingerprint, and random value. Then, the attacker can extract all other information hidden by operation XOR with hpw in B_i, C_i, E_i, R_i and successfully impersonate as U_i . In our scheme, the information shared by server with users is different. Two secret values x and IDS are combined with random value e to share $h(IDS \parallel e \parallel x)$ instead of sharing $h(x \parallel IDS)$ as in Khan et. al.'s scheme. Therefore, attacker cannot extract information shared by server with him/her to combine stolen information from other valid users.

The final weakness in Khan et. al.'s scheme is that the ID and the server's secret key are not tied in the same hash function. This leads that malicious user can change ID to impersonate other users. In our scheme, value $E_i = h(F_i \parallel hpw \parallel ID_i) \oplus h(ID_i \parallel x \parallel e \parallel IDS)$ is used to hide $h(ID_i \parallel x \parallel e \parallel IDS)$. User must have $h(ID_i \parallel x \parallel e \parallel IDS)$ to authenticate with server. Furthermore, users cannot change ID because they do not know x and IDS .

Eventually, we prove proposed scheme can resist many kinds of attacks. Because our scheme inherits Khan et. al.'s positive points, it also withstands some kinds of attacks which Khan et. al.'s can withstand. Furthermore, our scheme is improved to resist some kinds of attacks which Khan et. al.'s cannot resist.

Replay Attack

In this kind of attack, attacker resends message to impersonate user or server. Our scheme inherits from Khan et. al.'s scheme. We use random values combined with three-way challenge-handshake technique to resist this kind of attack. For example, attacker A can resend $\{h(e), RCID, C_0, C_1\}$ to S . Then, S re-sends $\{N_S, S_1\}$ to A . Because A does not know N_U and $h(ID_i \parallel x \parallel h(e) \parallel IDS)$, A cannot compute S_2 to re-send to S . Therefore, S recognizes who is impersonating U_i and terminates the session.

User Impersonation Attack due to Leakage's Identity

Our scheme can resist this kind of attack. Unlike Khan et. al.'s scheme, our scheme puts ID_i into hash function with two secret values x and IDS , and user need to use hash value to authenticate with server. If user's identity is leaked, attacker cannot compute $h(ID_i || x || h(e) || IDS)$ due to unknowing x and IDS . Therefore, attacker cannot impersonate user.

Impersonation Attack when Catching Login Message and Information Leakage from Mobile Device

In our scheme, values stored in mobile device and messages are careful design. Unlike Khan et. al.'s scheme, secret keys shared by server are completely different and malicious users cannot exploit anything from them. In our scheme, we store $\{B_i, C_i, E_i, V_i, h(\cdot), h_k(\cdot), N\}$ in mobile device. In authentication phase, user and server send $\{h(e), RCID, C_0, C_1, N_S, S_1, S_2\}$ to each other. Although attacker can intercept all those message, he/she cannot infer user's other information. Therefore, attacker cannot impersonate user and server.

Password Guessing Attack

Khan proposed using NOR operation to prevent attacker from extracting F_1 from hpw for password-guessing attack. We do not use NOR operation. Instead, we put F_1 into hash function with PW. Clearly, attacker cannot extract F_1 from hpw and do not have any clues for password-guessing. Therefore, our scheme will be secure even if user's password is weak.

Strong User Anonymity

In our scheme, a user sends $\{h(e), RCID, C_0, C_1\}$ to the server for each login. Therefore, an attacker can intercept and analyze the login message. The attacker cannot extract ID_i without knowing server's two secret values. Furthermore, the login message is dynamic for each login because of using the random value N_U . Therefore, the attacker cannot determine who is logging to the server.

As no part of login messages is identical to any stored value in a mobile device, it is impossible to co-relate a lost mobile device and login requests sent from this device in the past. Therefore, our proposed scheme can also preserve the strong user anonymity property that is considered in the original Khan et al.'s scheme.

Denial of Service Attack

In our scheme, user's ID is stored in the server until the end of the session to resist parallel session attack. If a session is still active, the server checks the existence of user identity in its temporary list and denies any login requests from this identity. Unlike Khan et. al.'s scheme, an attacker cannot impersonate user in our revised method. Thus he or she cannot block a legitimate user by procrastinating.

Mutual Authentication

In registration phase of proposed scheme, S provides U_i with two keys $h(IDS || e || x)$ and $h(ID_i || x || e || IDS)$. In login phase, U_i conceals N_U with secret key to prove his/her ownership of secret key. Together, S must have two secret values to re-extract

for login message’s verification from user. Then, S re-challenges users with N_S . If U_i is legitimate, he/she can compute S_2 to re-send to S .

Session Key Agreement

In our scheme, after successfully authentication, both server and user can compute common session key SK to make a secure channel. Therefore, data transmitted between user and server are encrypted to prevent adversaries from knowing content.

4.2 Efficiency Analysis

To compare efficiency between our scheme and Khan et. al.’s, we re-use Khan’s approach. That is, we count the number of one-way hash function execution. In addition, we ignore exclusive-or operation because it requires very few computations. In table 1, we compare the number of hash operation used in our scheme and Khan et. al.’s. Khan et. al.’s scheme needs $5h(.)$ in registration phase, $4h(.)$ in login phase and $9h(.)$ in authentication phase. Our scheme needs $8h(.)$ in registration phase, $8h(.)$ in login phase and $8h(.)$ in authentication phase.

Table 1. Comparison of computation costs

Schemes	Registration phase	Login phase	Authentication phase
Khan et. al.’s	$5h(.)$	$4h(.)$	$9h(.)$
Ours	$8h(.)$	$8h(.)$	$8h(.)$

Clearly, proposed scheme needs more computational costs than Khan et. al.’s scheme. However, those are necessary to provide user’s anonymity, prevent attacker from performing XOR with values, and generate session key for partners. In short, we see that proposed scheme does not add many additional computational costs and the proposed scheme is more secure than Khan et. al.’s scheme.

Table 2. The comparison between our scheme and the Khan et. al.’s scheme for withstanding various attacks

Attack	Schemes	
	Khan et al.’s	Our
Attacks due to ID theft	No	Yes
Replay attack	Yes	Yes
User impersonation attack	No	Yes
Server impersonation attack	No	Yes
Attacks using login request	No	Yes
Mobile device loss attack	No	Yes
Password guessing attack	Yes	Yes
Malicious user attack	No	Yes
Denial of service attack	No	Yes

In table 2, we list the comparisons between our improved scheme and Khan et. al.’s scheme for with standing various attacks. We see that Khan et. al.’s scheme cannot

resist to server and user spoofing attacks, attacks due to ID theft, attacks using login request and mobile device loss, malicious user attack, and denial of service attack. It can be seen that our proposed scheme is more secure against various attacks.

In table 3, we list the comparisons between our improved scheme and Khan et. al.'s scheme for the numbers of friendly features. Our scheme not only satisfies all Khan et. al.'s does but also supplies one important feature which their scheme lacks. This property is three-factor security.

From the entire comparison, we observe that the minor addition of six hash functions is worth achieving added security features and admired usable features.

Table 3. Comparison for providing admired friendly features

Feature	Schemes	
	Khan et al.'s	Our
User anonymity	Yes	Yes
User un-traceability	Yes	Yes
Mutual authentication	Yes	Yes
Session key establishment	Yes	Yes
No verification table at S	Yes	Yes
Quick wrong password detection	Yes	Yes
Freedom to change password	Yes	Yes
Three-factor security	No	Yes

5 Conclusion

We analyze Khan et. al.'s scheme. Although Khan et. al. propose new ideas, such as using NOR operation to resist password-guessing attack, using two secret values to resist secret key-guessing attack. However, we proved that Khan's design cannot resist some kinds of attacks such as attacks due to id theft, user and server impersonation attack, attacks using login request, mobile device loss attack, malicious user attack, and denial of service attack. Therefore, we propose improved scheme to overcome their limitations.

Compared with Khan et. al.'s scheme, our scheme owns some advantages as follows. (1) Our scheme provides user's strong anonymity and three-factor security. (2) Our scheme has more logic design and eliminates some drawbacks in Khan et. al.'s scheme. Therefore, our scheme is more secure than Khan et. al.'s and still efficient when compared with Khan et. al.'s scheme.

References

1. Ohrt, J., Turau, V.: Building-linked Location-based Instantaneous Services System. *Proceeding Computer Science* 32, 445–452 (2014)
2. Nikou, S., Bouwman, H.: Ubiquitous use of mobile social network services. *Telematics and Informatics* 31, 422–433 (2014)
3. Daponte, P., Vito, L.D., Picariello, F., Riccio, M.: State of the art and future developments of the Augmented Reality for measurement applications. *Measurement* 57, 53–70 (2014)

4. Lee, J.-S., Lin, K.-S.: An innovative electronic group-buying system for mobile commerce. *Electronic Commerce Research and Applications* 12, 1–13 (2013)
5. Oh, J.-B., Yoon, E.-J., Yoo, K.-Y.: An Efficient ID-Based Authenticated Key Agreement Protocol with Pairings. In: Stojmenovic, I., Thulasiram, R.K., Yang, L.T., Jia, W., Guo, M., de Mello, R.F. (eds.) *ISPA 2007*. LNCS, vol. 4742, pp. 446–456. Springer, Heidelberg (2007)
6. Tian, X., Wong, D.S., Zhu, R.W.: Analysis and improvement of authenticated key exchange protocol for sensor networks. *IEEE Communications Letters* 9, 970–972 (2005)
7. Chen, C., He, D., Chan, S., Bu, J., Gao, Y., Fan, R.: Light weight and provably secure user authentication with anonymity for the global mobility network. *International Journal of Communication Systems* 24, 347–362 (2011)
8. Xue, K., Hong, P., Ma, C.: A light weight dynamic pseudonym identity based authentication and key agreement protocol without verification tables for multi-server architecture. *Journal of Computer and System Sciences* 80, 195–206 (2014)
9. Khan, M.K., Kumari, S., Gupta, M.K.: More efficient key-hash based fingerprint remote authentication scheme using mobile device. *Computing* 96, 793–816 (2013)
10. Wang, R.-C., Juang, W.-S., Lei, C.-L.: Robust authentication and key agreement scheme preserving the privacy of secret key. *Computer Communications* 34, 274–280 (2011)
11. Yoon, E.-J., Yoo, K.-Y.: Improving the Dynamic ID-Based Remote Mutual Authentication Scheme. In: Meersman, R., Tari, Z., Herrero, P. (eds.) *OTM 2006 Workshops*. LNCS, vol. 4277, pp. 499–507. Springer, Heidelberg (2006)
12. Liao, I.E., Lee, C.C., Hwang, M.S.: Security Enhancement for a Dynamic ID-based Remote User Authentication Scheme. In: *Proceedings of the International Conference on Next Generation Web Services Practices*, pp. 437–440. IEEE Computer Society (2005)
13. Xu, J., Zhu, W.-T., Feng, D.-G.: An efficient mutual authentication and key agreement protocol preserving user anonymity in mobile networks. *Computer Communication* 34, 319–325 (2011)
14. Li, X., Niu, J.-W., Ma, J., Wang, W.-D., Liu, C.-L.: Cryptanalysis and improvement of a biometrics-based remote user authentication scheme using smartcards. *Journal of Network and Computer Applications* 34, 73–79 (2011)
15. Li, C.-T., Weng, C.-Y., Fan, C.-I.: Two-Factor User Authentication in Multi-Server Networks. *International Journal of Security and Its Applications* 6, 261–268 (2012)
16. Yang, J.-H., Chang, C.-C.: An ID-based remote mutual authentication with key Agreement scheme for mobile devices on elliptic curve cryptosystem. *Computers & Security* 28, 138–143 (2009)
17. Chen, T.-H., Chen, Y.-C., Shih, W.-K., Wei, H.-W.: An efficient anonymous authentication protocol for mobile pay-TV. *Journal of Network and Computer Applications* 34, 1131–1137 (2011)
18. Beitollahi, H., Deconinck, G.: Analyzing well-known countermeasures against distributed denial of service attacks. *Computer Communications* 35, 1312–1332 (2012)
19. Fan, C.-I., Lin, Y.-H.: Provably secure remote truly three-factor authentication scheme with privacy protection on biometrics. *Transactions on Information Forensics and Security* 4, 933–945 (2009)
20. Islam, S.K.H., Biswas, G.P.: A more efficient and secure ID-based remote mutual authentication with key agreement scheme for mobile devices on elliptic curve cryptosystem. *Journal of Systems and Software* 84, 1892–1898 (2011)

Prospective Cryptography in NFC with the Lightweight Block Encryption Algorithm LEA

Ha Van Nguyen, Hwajeong Seo, and Howon Kim

Department of Computer Engineering, Pusan National University,
Busan, Republic of Korea
{vanhax1,hwajeong84,howonkim}@gmail.com

Abstract. Near Field Communication (NFC) technology has been used more and more widely nowadays due to some undeniable interactive advantages. There are variety of NFC applications has been developed and used in business like bank transaction, public transport, commercial services, guided shopping, NFC access control and coupon. As a result, it has raised some new threats like hackers, card crimes and other privacy issues. For that reason, in this paper, we would like to contribute a method for securing NFC communication by applying a simple and lightweight cryptography algorithm. To illustrate the feasibility, we develop an NFC payment program in Android devices using lightweight encryption algorithm (LEA) and advanced encryption standard (AES) to make data confidentially. The performance of these two cryptography algorithms has been estimated and the result shows that LEA is much more advantageous than AES when applying in NFC application in which the small amount of data is saved to the card.

Keywords: NFC, security, mobile payment, LEA.

1 Introduction

Due to rapid advancement of mobile devices, several services are getting more and more popular in anytime at anywhere. These are feasible with modern smartphone supporting various network protocols including Bluetooth, Wi-Fi and NFC. Recently, NFC technology integrated into major smartphone platform like BlackBerry, Android and iPhone onboard bring us many conveniences in interaction between virtual and physical world. As a short-range wireless communication technology, NFC has been become an essential part of many daily use cases such as credit cards, debit cards, coupons, loyalty cards, access keys to car, offices, houses, etc. For its enormous impact on the financial ecosystem, there are many smartphone manufacturers, mobile network operator, financial institutions, research centers and governments are performing R&D activities to facilitate this technology [1]. Since these services are dealing with sensitive financial information, the services should be carried out secretly. The best solution for securing networking communication is encrypting the information with

secure cryptography algorithm. There are many block cipher algorithms exist. Among of them, we chose AES and LEA algorithm. In Republic of Korea, LEA block cipher has been proved as a fast and efficient way to secure data in common processors [2] and hardware implementation [3]. In this paper, we exploit novel ARX structure based block cipher LEA in NFC application and realize that its advantages are of a light-weight, simple structure method with high usability. It is simple and quick, yet exceptionally secure in NFC application in which the payload saved to tag is a small amount of data.

This paper is organized as follows. In Section 2, we introduce related technologies for our implementation. In Section 3, we show the comprehensive requirements of a payment system. In Section 4, we describe our payment proposed system to adopt LEA algorithm and then performance is estimated and compared with AES algorithm. Finally, in Section 5, we conclude the paper.

2 Related Works

2.1 Near Field Communication (NFC)

Near Field Communication (NFC) [1] is a set of short-range wireless, contactless technologies evolved from radio-frequency identification (RFID). NFC allows to write small piece of data from NFC device to a tag which is unpowered chip or read data from tag. NFC also allows to exchange data between two NFC devices, like two NFC-enabled smartphones. Tags can range in complexity. The data stored in the tag can also be written in a variety of formats. However, many of the Android framework API are based on NFC Forum standard called NDEF (NFC Data Exchange Format). There are some differences between NFC and Bluetooth and RFID wireless technology. Two notable differences among of them are close proximity and lower speed of transfer data. The theoretical maximum is about 10 centimeters. In practical, however, you are going to see closer distance which is about 4 centimeters or less typically to initiate a connection. The operating frequency is 13.56 MHz and the data rate that NFC provide is much lower than Bluetooth and higher than RFID. It depends on tag type, the data rate can be 106, 212, or 424 Kbps. The arrival of NFC technology has increased the level of intuitive, automatic interaction between humans and computer and become one of the vital important enablers for ubiquitous computing. It is seen that with recent quick advancement of NFC-enabled mobile phones, many traditional applications like credit cards, public transport, access cards, customer services can be replaced by NFC-based mobile phones because almost every one now carry a mobile phone.

2.2 Lightweight Encryption Algorithm (LEA)

LEA [2] is firstly introduced in WISA 2013 in South Korea which is favorable used in hardware and software implementation because of its simple ARX-architecture property. It is the same with traditional algorithm AES in size of

block cipher and support three key size which has fixed message block size of 128 bits and supports a key size of 128, 192 and 256 bits. The process consists of key schedule and encryption, decryption activity. The key schedule generates a sequence of round keys as follows. The key schedule uses several constants for generating round keys, which are defined as $\Delta[8] = 0xc3efe9db, 0x44626b02, 0x79e27c8a, 0x78df30ec, 0x715ea49e, 0xc785da0a, 0xe04ef22a, 0xe5c40957$. The number of rounds converting from plaintext to ciphertext in LEA is indicated by the key size used. It specifically takes 24 rounds for 128-bit keys, 28 rounds for 192-bit keys and 32 rounds for 256-bit keys. For 24, 28 and 32 rounds, it encrypts a 128-bit plaintext $P = (P[0]; P[1]; P[2]; P[3])$ to a 128-bit ciphertext $C = (C[0]; C[1]; C[2]; C[3])$ with 128-, 192- and 256-bit keys.

In [2], Hong et al. has presented features of LEA in terms of security level, efficiency and size on many platform of Intel, AMD, ARM, ColdFire. The study made broad comparison with other ciphers by former research such as AES, ARX structure block ciphers, lightweight block ciphers, stream cipher families and proved that the new block cipher LEA is meaningful to propose as a secure, fast and sufficient method that can against all the existing attacks. The security of LEA against several main existing cryptanalytic technique based attacks such as differential attack, linear attack, zero correlation attack, boomerang attack, integral attack, differential-linear attack, attacks using weakness of key schedule and others. Therefore, it is ensured that LEA is a secure cryptography algorithm.

2.3 SQLite

In Android environment, there are variety options for saving persistent application data such as: shared preferences, internal storage, external storage, SQLite database and network connection. We use SQLite database to store data in our application. SQLite is an open source, lightweight, standards compliant relational database management system (RDBMS). It is implemented as a compact C library and becomes an integrated part of the application that created it. SQLite has become popular database system of choice for many programs on smartphones due to its reliability and open source property [4].

3 Requirements of a Payment System

3.1 Requirements of Consumer

The achievement of all basic security requirements is essential for any e-payment protocols. The basic security requirements [5] are:

- **Security:** the payment transaction must always be carried out by the payer. There are some mechanisms for authentication such as: Something you know, Something you have, Something you are and Somebody you know. Using personal identification numbers (PINs), password or signature is the most common way.
- **Availability and Reliability:** Each e-Payment protocol have to beat all active or passive attacks of corrupt during the transaction.

- **Anonymity/Privacy:** practically, payers can pay money for anyone without revealing his identity and other personal information from the merchants.
- **Acceptability:** the e-payment system is acceptable widely. Therefore, when many consumers and merchants use it, they can make transaction with each other.
- **Ease of use:** the system should be simple, intuitive as much as it can to make end-user can make transaction easily.
- **Flexibility:** the payment system is better for user when it is flexible. There are some options for user to choose.
- **Price/Cost:** the cost for using system should be low.
- **Person-to-Person (P2P):** the system should provide feature of making transaction between person and another person.

3.2 Requirements of Merchant and Issuer

Here are list of requirements [5] of merchants and issuer:

- **Scalability:** the system can be upgraded in the future to keep up with government law or when changing their policy.
- **Efficiency:** the banks and financial institutions are always to make the services operate efficiently with low cost. The performance of payment system should be considered.
- **Customer Base:** the banks and merchants with good business revenue will care about their customer. So the e-payment system should be focus on their client.
- **Ease of Integration:** the ability of integrate with other system. It is important as almost software system components will frequently be upgraded, integrated with other components with variety of technologies they use.

3.3 Result of Former m-payment Researches

Due to many undeniable advantages, NFC has been rapidly becoming an important part in many areas, especially in mobile access control and payment system. It is more and more convenient for user when it can be integrated seamlessly to smartphone because of huge amount of smartphones user nowadays. And it inherently leads passion and challenges to many researchers and companies. For example, a practical offline payment system using NFC in mobile phone has been proposed by [6]. The result has shown that it is feasible to construct a highly secure offline payment system using NFC without third party, compromising party. But there are some limitations in their solution such as computation speed, available memory and security functionality. The main reason of timing limitation is that the mobile phone used is Nokia phone. In that phone, the secure element (SE) does not support AES and ECDSA (Elliptic Curve Digital Signature Algorithm). They said that AES and ECDSA would had been desirable [6]. Another paper [7] has compared the performances of different mobile payments based on IVR, SMS, WAP and OTP technology with

NFC-based mobile payment method. Looking at the table 2 of the paper, it can be seen that the NFC-based mobile payment get the best performance compare to other-based system. The time average, median, minimum, maximum values and standard deviation all are the smallest amount. So, it has shown how advantageous the state-of-art NFC-based method is in term of performance [7]. In [8], the list of threats which is applicable in NFC has been addressed. And the solution for defending it is establish a secure channel over NFC communication. In this paper, we would like to introduce a light-weight cryptography which is very favorable in NFC application in term of speed. It is hoped that it will be helpful for both companies that use NFC in their business line and NFC researchers.

4 Proposed Methods

In this section, we introduce our proposed method in which the performance is considerably improved with LEA. We applied LEA block cipher to secure NFC communication and database system. It ensure high availability with high throughput.

4.1 Payment Protocol

The general structure of m-payment system is shown in the Figure 1 below. The actors in this scheme include Banker, ATM, Retailer and Client who is beneficiary. They make money exchange transaction with each other through NFC communication.

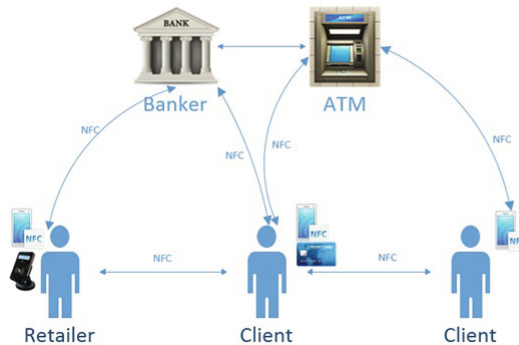


Fig. 1. The m-payment system with NFC

With advancement in smartphone industry development, almost modern android phone support NFC with three mode operations that are read/write mode, peer-to-peer and card mode emulation. So this payment scheme is simple as it contain only one type of transactions between actors which is exchange transaction between NFC reader and NFC tag.

The payment transaction is divided into two parts separately. Firstly, the registration phase is depicted in Figure 2. Secondly, the payment transaction phase is described in Figure 3.

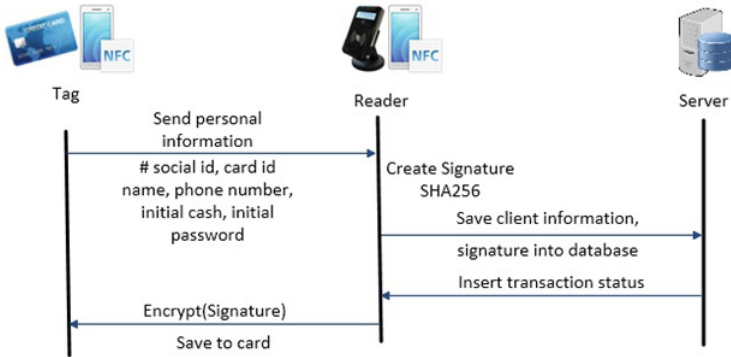


Fig. 2. The registration process in m-payment protocol with NFC

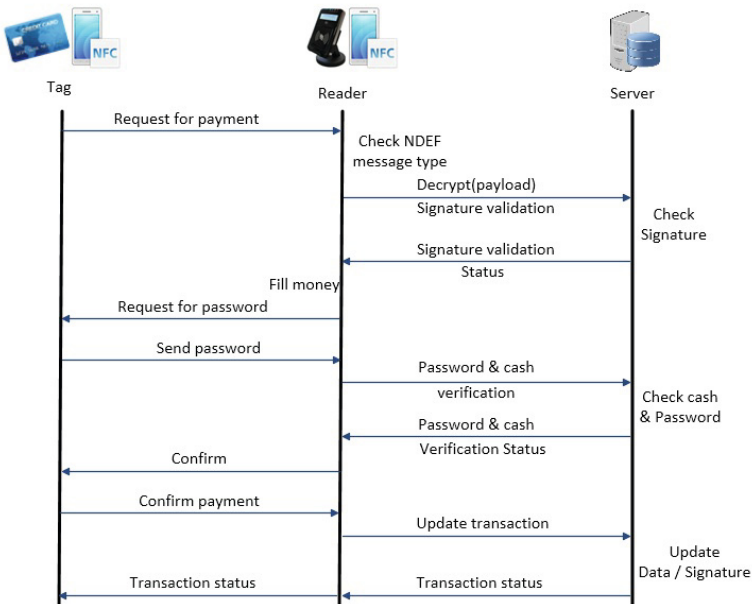


Fig. 3. The payment process of m-payment protocol with NFC

Firstly, the register send their personal information with initial cash and initial password to the banks or merchants. The signature of client is then created with SHA-256 algorithm. After that, the information and signature will be saved in

database. After inserting successfully, the NDEF message is created in which its payload is the encrypted signature of client and `TNF_EXTERNAL_TYPE`. The signature is encrypted with AES or LEA algorithm with three key sizes. Finally, we close the NFC tag which is the card or NFC-enabled smartphone to the reader to write the encrypted signature to tag.

When the customers want to perform a payment transaction, they will do as follow steps. Firstly, clients close their tag (NFC-enabled phone) to the reader (NFC-enable phone) to acquire payment transaction. The following step is checking NDEF message type which is the `TNF_EXTERNAL_TYPE` defined. If the `TNF_EXTERNAL_TYPE` is correct, the payload will be decrypted to get the client's signature. The signature will be checked its validity by comparing with the signature in the database. We can use `TNF_EXTERNAL_TYPE` as the first index layer, the signature as the second index, so the retrieving time in database can be saved significantly. If the signature of tag is correct, the amount of money will be filled and the password will be required. Then, the password and the amount of money will be verified. Finally, the merchants ask confirm from clients then update database and send the transaction status to client.

4.2 Program Implementation

For secure application, we implemented NFC payment system by using LEA block cipher. The demonstration includes tag initialization, payment session and storing the user information. The NFC payment system consists of two main bodies. Firstly, NFC communication between tag and reader. Second, database system is designed to store sensitive user information secretly. The information are encrypted with AES or LEA block cipher.

LEA Cryptography Algorithm Implementation. Java is a form of high level language which enables developer to write program easily and reduce the logical errors. For LEA implementation, we realized basic ADD, ROTATE, and XOR operations in Java because these are basic operations in LEA algorithm. Using Java is simple to implement however it does not have unsigned integer type, so we manually implemented unsigned integer data type. The table 1 below, unsigned integer rotation is realized in Java. Java provides unsigned integer right shift (`>>>`) but not for left shift (`<<`). To convert signed integer into unsigned integer type, we conduct unsigned integer right shift by zero. Remaining operations including addition and bitwise-exclusive-or are simply established with '+' and '^' operation.

AES Cryptography Algorithm Implementation. In Java library, general cryptography algorithm are packaged in `javax.crypto`. The package provides the classes and interfaces for cryptographic operations include key generation, key agreement, message authentication code (MAC) generation, encryption and decryption. Because many cipher classes in this package are provider-based and written by independent third party vendors and plugged in seamlessly, the developers can take advantage of any these implementations without having to

modify source code. In our demonstration, we use traditional algorithm AES to compare with LEA. Our implementation does not add any optimization methods, so it is a fair comparison of Java implementations between AES and LEA block cipher algorithm.

Table 1. Left/Right rotation operation

function ROL(input, offset)	function ROR(input, offset)
input = ((input<<offset)>>>0) (input>>>(32-offset))	input = ((input<<32-offset)>>>0) (input>>>(offset))
return input	return input

Database Implementation. Table 2 below describes the structure of the data table we use in the NFC Payment application. We named the table accounts, store it in SQLite database. This table contains eight columns. The first is columns id (INT) which is PRIMARY KEY. Other columns are name (TEXT) is the name owner of the card, phone_number (TEXT), social_number (TEXT), open_date (TEXT) which is the date open of this account, cash (INT) which is the money this account has, password (INTEGER) which is 4-private number the owner of the card has to fill every time they pay or deposit money, signature (TEXT) which is generated when opening the account. This signature will be encrypted with AES or LEA algorithms with key length 128, 192 or 256 bits then will be stored to the NFC card.

Table 2. Data structure of accounts in database

Field	Type	Key
id	INTEGER	PRIMARY KEY
name	TEXT	NOT NULL
phone_number	TEXT	NOT NULL
social_number	TEXT	NOT NULL
open_date	TEXT	NOT NULL
cash	INTEGER	NOT NULL
password	INTEGER	NOT NULL
signature	TEXT	NOT NULL

In Figure 4, screen of payment program is described. Firstly, users credit card information is filled into blank. Secondly, user can choose some options like algorithm, key size, encrypt performance estimation. When we pay money in store or shop, we should submit correct password to use our card. Whenever our password is correct, transactions are accomplished.

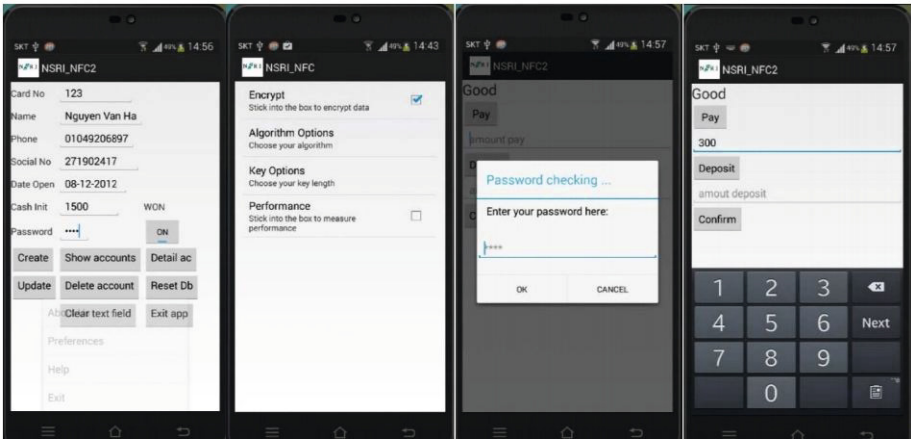


Fig. 4. The screen of demo m-payment program

4.3 Performance Evaluation

In this section, we would show the performance estimated when applying AES and LEA block cipher over NFC Payment application. The android version for running the applications is equal or greater than 4.0 and NFC-enabled smartphone. We prepared two NFC white cards to write and read information on it. The NFC Payment programs have been deployed and tested on Samsung Galaxy Note 2 which its model is SHV-E250S and its Android version is 4.3. The core processor is ARM Cortex-A9 powered at 1.6 GHz and the number of core is quad. The RAM size is 2 Gigabytes. The performance is measured by triggering `System.currentTimeMillis()` to get current time in millisecond. To stabilize results, we iterated the function several hundred times to get rid of delays.

The target devices we use in our project are shown in Figure 5. The NFC white card is used for target NFC tag.



Fig. 5. The target device using for testing

As a pilot project, we simple tested scenario that NFC reader encrypts/decrypts data. However, we can exploit card emulation mode which emulates card function on Android NFC-enable mobile phone. Finally, encrypt/decrypt functions would be conducted on Tag (Smartphone) as well.

We implemented 128-, 192-, 256-bit AES and LEA algorithms and measured performance five times to get average values. The performance of encryption is described in Table 3. It is believed that NDK will further improve the performance but it does not show compatibility on various devices.

Table 3. Performance evaluation of AES and LEA encryptions (clock/byte)

Encrypt-bit	AES-128	AES-192	AES-256	LEA-128	LEA-192	LEA-256
1	4885	5570	5875	985	1150	1390
2	5405	5490	5545	1055	1025	1485
3	5470	5155	5555	1215	1255	1235
4	5205	5365	5660	965	1005	1385
5	5275	5360	5595	1160	1080	1315
Average	5250	5390	5645	1075	1103	1360

And the performance of decryption activity is described in Table 4 below. It can be seen that LEA was not only better in encryption but also in decryption case.

Table 4. Performance evaluation of AES and LEA decryptions (clock/byte)

Decrypt-bit	AES-128	AES-192	AES-256	LEA-128	LEA-192	LEA-256
1	6665	7040	6920	1180	1375	1445
2	6165	6720	6975	980	1360	1325
3	6190	6835	7110	1145	1310	1345
4	6255	6915	7085	985	1210	1465
5	6215	6395	7080	1270	1210	1380
Average	6300	6780	7035	1110	1295	1390

In Figure 6, comparison results on various conditions are drawn. With a glance, we can easily recognize that LEA block cipher is better choice for efficient encryption task in term of performance, yet securely. It improves nearly 80, 80, 76, 82, 81, 80 % for 128-, 192-, 256-bit key encryption and decryption respectively. We can look at the bar graph which is shown in Figure 6 below for more intuitive view.

When we use NFC applications to share data, it goes through two basic processes including writing card and reading card. In writing card activity, the time for writing NDEF message is much bigger than the time for encrypting payload, so choosing the encryption algorithm is not meaningful. On contrast, in reading

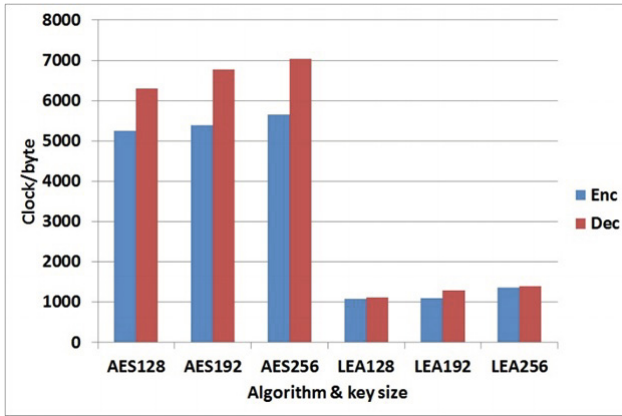


Fig. 6. Performance for AES and LEA algorithm in NFC payment program

card activity, the time for reading NDEF message is smaller than the time for decrypting the payloads, so for better performance the efficient algorithm should be considered. In this perspective LEA algorithm is undeniable good option to construct secure NFC systems. In Table 5, the performance of total costs for NFC payment system is shown.

Table 5. Performance evaluation of AES and LEA decryptions (clock/byte)

Process	128-bit	192-bit	256-bit
LEA-Read	1635	1760	1950
LEA-Write	7301075	7140895	7031360
AES-Read	6825	7245	7595
AES-Write	7305250	7145390	7035645

From the Table 5, we can see that the time for whole reading process is much smaller than the time using for whole writing process. That is because the time for writing NDEF message is much longer than the time for reading NDEF message in NFC. And the bar-graph has been drawn in Figure 7 below. Looking at these two bar charts, the first graph shows about performance for whole writing process. The blue column is when using AES algorithm and the red column is when using LEA algorithm. It can be seen that it is a bit same when compare using AES or LEA. Because the time for encrypting data is very smaller than the time for writing NDEF message. So, in term of performance, choosing which cryptography algorithm in this case is not much meaningful. On the other hand, when we look at the second bar graph, it shows that LEA has a much better performance than AES algorithm. That is as the time for decrypting data is bigger than the time for reading NDEF message. Then the time for decrypting data has affected much in this activity. Therefore, choosing

a good cryptography algorithm is vital important in this case if we want to get a good performance. In this case, LEA has enhanced performance approximately 76, 76, 75 % for 128-, 192-, 256-bit key decryption. In the future, the trend to extend our work is there are various mode of operations of LEA on Android. And if we can apply LEA on NDK environment for Android services, the performance will be improved more.

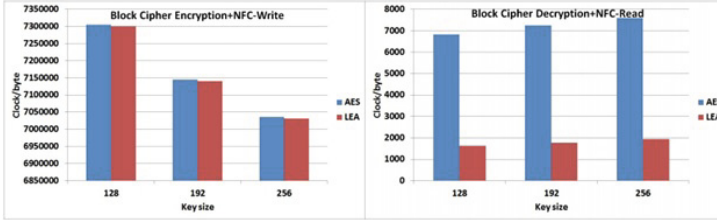


Fig. 7. Performance comparison for whole writing and reading process

5 Conclusion

In this paper, we proposed method to get confidentiality of data in NFC application by exploiting LEA and AES cryptography algorithm. To apply LEA algorithm into NFC case, we implemented a NFC payment example program. The performance of these two algorithms has been estimated and shown in part 4.3. The comparison shows that when using LEA, the speed is improved over 70% than when using AES algorithm in our NFC payment system. It is particularly meaningful in reading tag case and almost NFC transactions operated daily is reading tag, not writing tag. So, it would be really efficient if we can apply such a lightweight cryptography algorithm like LEA in the economically attractive technology NFC application. It is hoped that our method could contribute not only in term of m-payment systems but also in other NFC applications, provide valuable information not only for NFC-enabled smart devices manufacturers but also for other organizations such as bankers, merchants, NFC researchers who develop their system with NFC adoption.

Acknowledgements. This work was supported by the Industrial Strategic Technology Development Program (This work was supported by the ICT R&D program of MSIP/IITP. [10043907, Development of high performance IoT device and Open Platform with Intelligent Software]).

References

1. Coskun, V., Ok, K., Ozdenizci, B.: Professional NFC Application Development for Android, 1st edn. John Wiley & Sons (2013)
2. Hong, D., Lee, J.-K., Kim, D.-C., Kwon, D., Ryu, K.H., Lee, D.-G.: LEA: A 128-bit block cipher for fast encryption on common processors. In: Kim, Y., Lee, H., Perrig, A. (eds.) WISA 2013. LNCS, vol. 8267, pp. 3–27. Springer, Heidelberg (2014)
3. Lee, D., Kim, D.-C., Kwon, D., Kim, H.: Efficient Hardware Implementation of the Lightweight Block Encryption Algorithm LEA. *Sensors* 14(1), 975–994 (2014)
4. Meier, R.: Professional Android 4 Application Development. John Wiley & Sons (2012)
5. Walezuch, R., Duppen, R.: Payment system for the internet - consumer, <http://arno.unimaas.nl/show.cgi?fid=445>
6. Damme, G.V., Vouters, K.M., Karanhan, H., Preneel, B., Offline, N.F.C.: Payments with Electronic Vouchers. In: Proc. of the 1st ACM Workshop on Networking, System and Applications for Mobile Handhelds, pp. 25–30 (2009)
7. Massoth, M., Bingel, T.: Performance of different mobile payment service concepts compared with a NFC-based solution. In: Fourth International Conference on Internet and Web Application and Services, pp. 205–210 (2009)
8. Haselsteiner, E., Breitfub, K.: Security in Near Field Communication (NFC), <http://ece.wpi.edu/~dchasaki/papers/SecurityinNFC.pdf>
9. List of NFC-enabled mobile devices, http://en.wikipedia.org/wiki/List_of_NFC-enabled_mobile_devices

Enhance Fuzzy Vault Security Using Nonrandom Chaff Point Generator

Minh Tan Nguyen, Quang Hai Truong, and Tran Khanh Dang

Faculty of Computer Science & Engineering, HCMC University of Technology,
VNUHCM, Vietnam

tannnguyen@dds-data.vn

{haitruong, khanh}@cse.hcmut.edu.vn

Abstract. Toward the combination of cryptographic and biometric systems, by performing a bidding technique on cryptographic key and biometric template, fuzzy vault framework tries to totally enhance security level of current biometric cryptographic systems, hides secret key and protects the template. Though original schema suggests the use of error-correction techniques (e.g., Reed Solomon (RS) encoder/decoder) to reconstruct the original polynomial, recent implements do not share the same point of view. As a replacement, Cyclic Redundant Code (CRC) is applied to identify the genius polynomial from set of candidates. Within scope of this paper, we address a significant flaw of current CRC-based fuzzy vault schemas which leads to a potential of successful blend substitutions attack. To overcome that problem, we proposed a modification and integration of two novel modules into fuzzy vault's general schema: chaff points generator and chaff points verifier. New modules are designed to integrate easily to current running systems as well as simple to enhance. The proposed schema can detect any of modification in vault and, as a result, eliminate the blend substitution attack, enhance general security.

Keywords: Biometric template security, fuzzy vault, CRC, blend substitutions attack, cyclic hashing, linear projection.

1 Introduction

Nowadays, though biometric are widely applied in authentication as well as authorization systems, there are still several concerns against the use of them. Most of the worries related to the lack of robust security schemas to protect the biometric template and efficiently handle errors caused by noises. Reviewing the state of the arts techniques addressing those concerns lead us to two mains technologies: biometric cryptosystems and cancelable biometrics [1]. Toward protecting biometric template, biometric cryptosystems (BCSs) try to integrate biometric with cryptographic. Specifically, cryptographic key will be bind directly with biometric template in key binding schemas or generated from biometric information in key generator schemas. Meanwhile, cancelable biometrics (CB) chooses to transform and compare biometric within transformed domain. Along with those techniques, there are researches around

hybrid schemas which merge CB with BCSs. Specifically, biometric templates will be transformed before treated as input for BCSs.

Within this paper, we pay attention to fuzzy vault, one of the most popular key binding techniques. Fuzzy vault schema, first introduced by Juels and Sudan [2], is well-designed for unordered set and can handle noisy data by applying error correction code. Within fuzzy vault systems, an unordered set A and secret key k are bind together by applying polynomial projection, yielding a vault, denoted by V_A . The secret key, later, is reconstructed if and only if the system receive an unordered set B which has sufficient overlapping elements with A . Most of the implement of fuzzy vault schema includes two mains phases: the enrollment phase and the authentication phase. During the enrollment phase, secret key k is used to generate coefficients of a predefined mono polynomial p . Projecting each element of set A to 2D-space by this polynomial results a set of genius points. For hiding genius points, random chaff points are added. The set of all points, including genius points as well as chaff points, forms the vault V_A . Considering authentication stage, the input set B needs a tolerable overlap with A in order to locate sufficient amount of genius points for reconstruction of p . For the successful regeneration of p , subsequently secret key k , error correction code are applied on B . The essential security of the whole system is based on the complexity of reconstruction polynomial without knowing genius points hiding among considerable amount of chaff points.

Juels and Sudan suggest the use of RS error correction code within their schema for efficiently decoding vault with noise. As mentioned above, recent researches on fuzzy vault, [3] [4] [5] [6], propose approach of using the combination of CRC and Lagrange interpolation for vault decoding. Within new approach, authentication's conclusion will be evaluated from sort-list of candidates not just based on one specific error correction result. However, significant performance and security problems of CRC approach as reported in [7] are left behind.

This paper proposed a novel method address solving current CRC-based fuzzy vault's problems. Additional to traditional modules of fuzzy vault schema, we introduce a modified chaff point generator and new chaff points verifier modules. Continuously hashing and linear projection will be used to structural generate chaff points at enrollment phase. As the consequence, authentication phase applies the same algorithm to regenerate the same chaff points before giving final decision. Within the proposed schema, chaff points will be treated as signature for the combination of biometric template and secure key. Applying our approach, any modification of decoding vault in compare with originate vault will be detected at authentication phase, preventing the attack of blend substitutions.

Turning to the main structure of this paper, after a brief review of technical approach of fingerprint-based fuzzy vault implements, another significant security flaw of CRC-based fuzzy vault schema will be described in section 1. Assume that attacker has read/write permission to template database, by using blend substitutions attack; attacker can exploit above flaw to gain access to our system without effecting current user working, as a consequence, hiding from detection. Within section 2, proposed approach to overcome that flaw as well as enhance totally security level of whole systems will be introduced. Two additional modules and major changing in chaff points generating

process will be described in details. This section also contains security analysis of the proposed schema. Finally, experiments results will be reported in section 3.

2 Related Work

2.1 Fuzzy Vault Implements

Originally, Juels and Sudan, in [2], discuss problems of using ordered set within biometric system to authentication and cryptography integration. As a consequence, they introduced the schema of fuzzy vault for unordered set and the application of error correction code to deal with the variant of input set. The schema including two main algorithms: *LOCK* and *UNLOCK* basing basically on the complexity of the polynomial reconstruction problem, a special case of the RS decoding problem, to achieve appropriate security level. Nevertheless, there isn't detail implementation as well as real evaluation base on practical data is introduced.

Clancy et al., in [8], provide a practical implement of fuzzy vault on smart card using fingerprint information. Within this implementation, finger print image are assumed to be pre-aligned. User's minutiae are extracted from finger print and form the unordered input set for fuzzy vault schema. Chaff points are, then, randomly inserted with a restriction of having to keep minimum distance d from other points of vault. For the chosen of error correction code, they keep the work of RS, just as in [2]. Two main contributions of [8] are proving of attacking complexity on fuzzy vault schema and the introduction of a practical implement of fuzzy vault with real data set. However, as mentioned above, the system need a pre-aligned finger print, which is not practical within real world and there isn't a details RS code implement is described.

Toward a more practical implement, Udulag and Jain [3] introduce two specific changings on fuzzy vault schema in compare with previous systems. First, at enrolment phase, they apply Orientation Field Flow Curves (OFFC) estimation and Iterative Closet Point (ICP) based alignment on finger print image to extract a specific helper data. Those helper data, later, are used to align the decoding fingerprint image at authentication phase and as the consequence reduce noises and improve accuracy of the whole system. Second, in spite of using RS error correction code, Lagrange interpolation and CRC are used to regenerate the correct polynomial from combination sets of potential genius points chosen from vault. For details implements, CRC of genius key will be calculated and appended directly to the end of original key, this new key will be used for encoding algorithm. At decoding phase, after getting the key, CRC part will be retrieved and used to check whether current key is genius or not and as the consequence user will be gain access permission or not.

CRC is continuously applied in Nandakumar et al. implement [6]. Inherit from the previous research, this paper apply the work of Udulag [3] to reduce environmental noise by extracting helper data as described above. Main enhancement of [6] is the integration of password-based transformation. Biometric template, the fingerprint minutiae, is individually transformed by user-provided password. Particularly, the coefficients and orientation of each minutiae will be transformed by adding user's password (after converted to bit strings) and modulo by maximum value of each

dimensions. Transformed minutiae, then, is used as input for fuzzy vault schema. Within the security perspective, this composite technique is proved to be secure against essential attacks on fuzzy vault schema such as record multiplicity attack, brute force attack and key-inversion attack. However, because of limitation of CRC-based approach, this schema still can be attack via blend substitutions method.

Main schema of CRC-based fuzzy vault system is described in figure 1 as below:

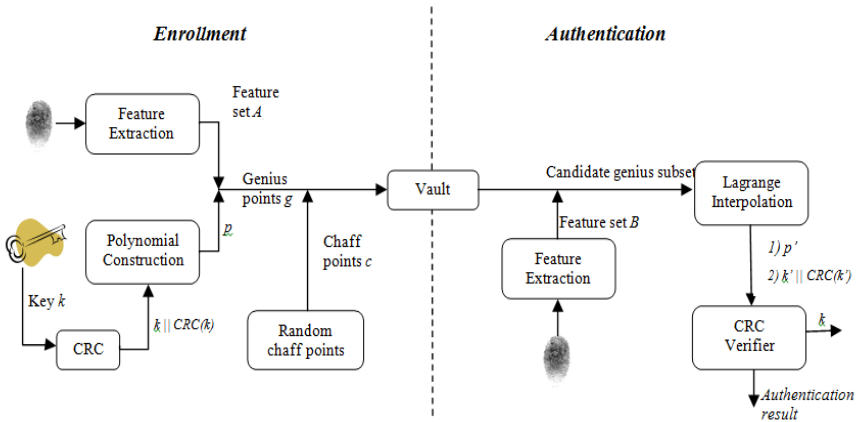


Fig. 1. Original CRC-based fuzzy vault

2.2 Attack against Fuzzy Vault Systems

Bruce Force Attack

Preda, in [9], build a brute force attack model on Clancy’s fuzzy vault’s implement [8]. Providing a vault and information about degree k of chosen polynomial, attacker continuously find $k+1$ points from vault and apply Lagrange interpolation to get candidate key. By verifying CRC of the result key, attacker can identify the genius one. Within this work, mathematic model to appropriate the complexity of attacking algorithm are constructed. From author’s analysis, the attack’s complexity grows when increasing the degree of the polynomial, the number of chaff points per vault or reducing the number of genius points. On the other hand, choosing insufficient combination of configurations can lead to significant decrease of complexity.

Statistical Analysis Attack

Base on the definition of free area, the available space to put in new points without violate minimum distance restriction between points in vault, and the assumption that points are selected to insert one-by-one, Ea-Chien et al. [10] provide a novel method to narrow the size of candidates for genius point when applying brute force attack. From their observation, points that are inserted later to vault are more likely to have smaller free space. As the result, when receiving a vault, attacker firstly tries to compute free area of all points. Brute force attack, as the later step, firstly chooses points which have large free area as input.

Collusion Attack

Collusion attack on fuzzy vaults considers the case when attacker gains access to multiple vaults of same biometric template and/or same secret case. By comparing points across vaults, attacker can identify and remove chaff points, thus, reducing the number of spurious polynomials. Hoi Ting Poon, in [11], examines in details each collusion attack scenarios and analyses the security loss in practical implement.

Key Inversion Attack

Key inversion attack addresses the case when secret key are compromised. Knowing the secret key, attacker can regenerate genius polynomial, thus, identify genius points among chaff points by filter those which have coordinates satisfy the genius polynomial. Because of the ability to directly disclose user's real biometric template, key inversion attack is a critical attack on fuzzy vault systems. Recall that biometric template is almost unique and hard to be replaced in practical.

Blend Substitution Attack

Within this kind of attack, by clever calculation and injection of fake points to vault, attacker can gain access to system without effect to current user's authentication process. By applying this attack, attacker can generate a backdoor to continuously penetrate our authentication mechanism without detected. In [12], blend substitution attack via biometric systems is reported and analyzed with significant notice.

2.3 CRC Problems

Previously Reported Problems

Hoi and Ali, in [7], point out two main security problems related to CRC-based fuzzy vault schema. First, current CRC implement often results a checking code with short length such as 16 or 8 bits in length, thus, gain the probability of CRC collision. The paper's implement shows that, within 16-bits CRC, when working with large number of interpolations to decode, result key can have up to 50% CRC collision. Within a CRC-based fuzzy vault system, a CRC collision can lead to accepting illegal result key and returning authentication successful, thus, increase False Accept Rate (FAR). Second, adding CRC to secret key before apply polynomial generation can lead to requirement of higher degree of polynomial in compare with secret key only. As a result, to assure the security level, our encoding and decoding have to work with high degree polynomial, thus, need more genius points to successful decode same vault leading to increasing of False Reject Rate (FRR).

Blend Substitutions Attack on CRC-based Fuzzy Vault

Number of points per vault of a fuzzy vault system normally is a predefined parameter of encode/decode algorithm. Therefore, any modification lead to the change of this number of specific vault will be easily detected by the system. Within blend substitution attack, main target of attacking is bypassing the authentication system without being detected. To archive this goal, attacker have to apply fake injection on vault without modify total number of points while keeping current authentication behavior of genius users.

Figure 2 is demo of blend substitution attack on vault. In order to perform the attack, attacker, firstly, generate their fake key and corresponding CRC. Applying fuzzy vault encoding algorithm on fake key and biometric, fake genius points with sufficient information to bypass the authentication system will be generated and injected to current vault after randomly deleting some vault's points. Consider vault V with g genius points and c chaff points, n denoted for the level of current polynomials, for successful regenerate key from that vault, from the theory of polynomial reconstruction, genius user need to have at least $(n+1)$ matching genius points from their biometric information. Consequently, attacker will need to delete at least $(n+1)$ points from vault V to inject his own fake points. Let's consider a practical case, suppose that $n = 9$, $g = 24$, within the worst case, 10 genius points are deleted from vault via blend substitution attack. Considering vault still have 14 genius points belonging to genius user allowing genius user to perform successful authenticating action. Recall that to attempt a successful authentication, genius user, as well as attacker, just need 10 points matched.

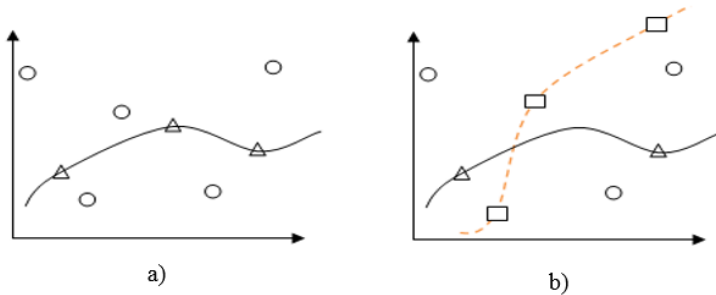


Fig. 2. Blend substitutions attack on CRC-based Fuzzy Vault systems: a) Original vault with only one genius polynomial; b) Blend attacked vault with genius polynomial together with fake polynomial

3 Proposed

3.1 Proposed Schema

In addition to traditional fingerprint-fuzzy vault modules, we modify and integrate two specific modules: chaff points generator and chaff points verifier. New modules are used as a replacement for current CRC mechanism of CRC-based fuzzy vault systems:

- *Chaff points generator module* accepts feature set extracted from biometric template and secrete key as input, then, outputs corresponding set of chaff points.
- *Chaff points verifier module* apply the same algorithm with chaff points generator to verify vault's chaff points after successful regenerate secret key.

Integration of new modules with original fuzzy vault schema can be found in figure 3. Within new schema, chaff points can be seen as a virtual 'CRC' for genius points and

secret key. By checking chaff points at authentication phase, we can not only verify the secret key but also detect any unauthorized modification on vault and, as a result, prevent efficiently blend substitution attack.

Algorithms to implement new modules can be varied however they have to satisfy basic conditions:

1. *Generated chaff points have to depend directly on both biometric information and secret key* in order to prevent attack of reducing candidate set when knowing part of biometric template or secret key
2. *Provide unique output for each combination of biometric and secret key* in order to eliminate collusion attack on different vaults of same biometric.
3. *Low entropy loss*: additional implement mostly leads to increasing of saving data and/or providing additional potential attacking methods. New algorithm need to eliminate all those kinds of entropy loss.

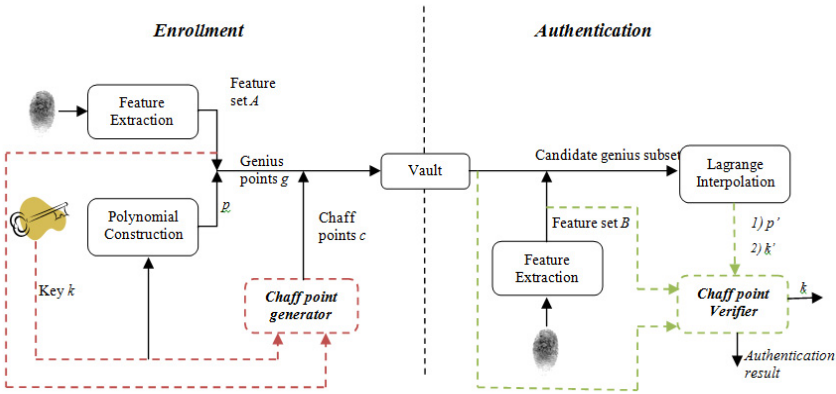


Fig. 3. Proposed schema: Fuzzy vault with chaff point generator and chaff point verifier

3.2 Proposed Algorithm

Within scope of this paper, we introduce a novel method to implement new proposed schema by apply continuous hashing and linear projection.

Consider a fingerprint-fuzzy vault system processing vaults with g genius points and c chaff points. Each genius points have the form as $G(X, Y)$ in which X is an l -bits integer formed by the combination of user biometric and Y is $f(X)$:

- $X = x || y || \Theta$ ($||$: bitwise concatenation)
- $Y = f(X)$

Where:

- x : x-coordinate of user’s minutiae
- y : y-coordinate of user’s minutiae
- Θ : orientation of user’s minutiae
- $f(x)$: the polynomial generated from secret key k

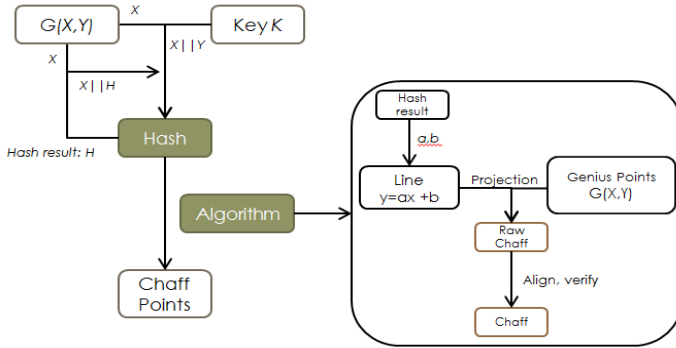


Fig. 4. Proposed algorithm to implement new fuzzy vault modules

Chaff Point Generator

At enrollment phase, each genius points have the responsibilities to generate a fix number of $\lfloor \frac{g}{c} \rfloor$ chaff points. For generating one specific chaff point, x -coordinate of current genius point X will be concatenated with secret k to form the input of a hashing module. The first two non-overlap l -bits strings of hashing result are retrieved to form two coefficients a, b of a linear function $y = ax + b$ within the 2D-space. Corresponding raw chaff point will be the result of projecting current genius points onto $y = ax + b$. The projection has responsibility to remove the link between hash values while keeping point generation to be random. In order to make new raw chaff points be uniform with the others in same vault, modulo function to maximum value of each genius points' coordinates is applied. This candidate chaff point will be considered as real chaff points if it satisfies all chaff point's conditions as:

1. Satisfy minimum distance requirement with other points in vault in order to prevent statistical analysis attack.
2. Not belong to the main polynomial of current vault to guarantee an accurate decoding.
3. Not have the same x -coordinator value with other points in vault because x - coordinator of genius points is generated by combining information of fingerprint minutiae. Recall that there will be no two fingerprint minutiae of a user that have identical information.

New conditions or transformations based on specific implements can be applied easily at this phase.

The hashing result will, continuously, be concatenated with secret key and put in hashing module again to get the new chaff points, following the same procedure. The loop for current genius point will be stopped and continued with the next one whenever sufficient number of chaff points is gotten or maximum loop per genius points is reached.

Within each genius points, number of tries to generate chaff points is counted and saved to a specific list. At the end of chaff points generation phase, this list will be sorted descending based on the value of appropriate x -coordinates. After removing all

x-coordinates data, result list will be saved as helper data for verification phase. Summary of proposed algorithm is described in figure 4.

Algorithm CHAFF_GENERATE.

Public parameters: an integer number *MaxLoop*, a SHA hashing algorithm *SHA*, 2-D space linear projection algorithm *LinearProjection*

Input: a fingerprint feature set *G*, secret key *K*

Output: a vault *V*, helper data *RCT*

Begin

Input *G*, *K*

$V \leftarrow G$;

$CT \leftarrow \{\emptyset\}$

for $i = 1$ to g do

begin

$X \leftarrow G[i].X$;

$H_i \leftarrow X \parallel K$;

$H_o \leftarrow SHA(H_i)$;

$j \leftarrow 0$;

$count \leftarrow 0$;

while $j < t$ AND $count < MaxLoop$ do

begin

$a \leftarrow GetNext16bits(H_o)$;

$b \leftarrow GetNext16bits(H_o)$;

$tempChaff \leftarrow LinearProjection(G[i], f(x) = ax + b)$;

if $(Verify(tempChaff) == true)$

$V \leftarrow V \cup \{tempChaff\}$;

$j \leftarrow j + 1$;

$H_o \leftarrow H_o \parallel K$;

$count \leftarrow count + 1$;

end;

$CT \leftarrow CT \cup \{(G.X, count)\}$;

end;

$RCT \leftarrow RemoveX(Sort(CT))$

End;

Chaff Point Verifier

At authentication phase, after generate candidate key from user-provided biometric, appropriate polynomial as well as candidate genius points can be retrieved easily from vault. Next, genius points list is sorted base on x-coordinates. By combining with previous helper data, we can know exactly number of tries that need to apply on each genius points to get their chaff points. Knowing key as well as genius points, chaff

points generation algorithm can be reapplied to regenerate chaff points. Modification of vault can be detected by one of the following cases:

1. Number of genius points less than length of helper list
2. For specific genius point:
 - (a) The number of tries less than maximum loop per chaff points and the number of chaff points generated less than required
 - (b) Or number of chaff points generated larger than required
 - (c) Or there are points that are not belong to genius points set and chaff points set

In the case of combination of candidate key and candidate genius points regenerates exactly the other points of vault, authentication will be successful.

Algorithm CHAFF_VERIFIER.

Public parameters: a SHA hashing algorithm *SHA*, 2-D space linear projection algorithm *LinearProjection*

Input: a vault *V*, helper data *RCT*, secrete key *K*

Output: a bool value result of authentication process *result*

Begin

Input *V*, *K*, *RCT*

G • *GeniusFilter(V,K)*;

G • *OrderDescByX(G)*;

CountInVault • *Count(G)*

for *i* = 1 to *Count(G)* do

begin

X • *G[i].X*;

Hi • *X || K*;

Ho • *SHA(Hi)*;

j • 0;

while *j* < *RCT[i]* do

begin

a • *GetNext16bits(Ho)*;

b • *GetNext16bits(Ho)*;

tempChaff • *LinearProjection(G[i], f(x) = ax + b)*;

if(*InVault(V,tempChaff)* == true)

CountInVault • *CountInVault + 1*

j • *j + 1*;

Ho • *Ho || K*;

end;

end;

if *CountInVault* == *Count(V)*

result • true;

End;

3.3 Security Analysis

Blend Substitution Attack Problem

Supposed that attacker have permission to access directly to template database server. To successfully attack our system via blend substitution method, attacker has to randomly delete at least $n+1$ points and injects his own fake points into our vault.

- In case deleted points are genius points, real chaff points that belong to those genius points cannot be regenerated at verification phase. As a result, legitimate user fail to authenticate to our system and problem will be reported.
- In other case, if deleted points are chaff points, without the knowledge of secret key, attacker cannot apply our chaff point generation algorithm to generate accurate chaff points. So, fake points will be realized as not belong to any genius points and lead to legitimate user's authentication fail.

Previous CRC Problems

By removing CRC from general schema, all previous reported problem related to CRC are eliminated. There will be no CRC collision that leads to increasing of FAR. Furthermore, input of polynomial construction will be reduced in length while still keeping sufficient degree to ensure the security level, as a result, decreasing FRR.

Entropy Loss Evaluation

Hash function, in general, is sensitive with bit change. So, by continuous concatenating hash result with original value, we can vary the linear function and chaff points, as consequence, keep the "random characteristic" of chaff points.

In case attacker try a brute force attack by investigate all possible hashing value of current hash function, the linear projection will be an additional protection layer. Linear projection, as mentioned above, helps to remove all links between hash values. Recall that chaff points are the result of project genius points to 2D-line but we save nothing related to the 2D-line.

In case attacker have information about a genius points or a chaff points in a chains, the other chaff points of current chain still be safe because he need to get the secret key to figure them out.

Considering the helper data, the helper list contains only the number of tries for each genius points. By applying sorting base on x-coordinator value of genius points at enrollment phase, list of appropriate genius points and its corresponding number of tries can be retrieved easily without saving information about x-coordinator in the helper data.

4 Experiments

As mentioned above, Nandakumar et al. in [6], provide a composite protection schema from fuzzy vault attacks based on the combination of fuzzy vault with helper data and password schema. In inheritance from previous achievements, we choose to represent our new modules built upon the foundation of Nandakumar et al. schema.

4.1 Nandakumar’s System

For general ideal of Nandakumar implement, there are two main phases: *enrollment* and *authentication*.

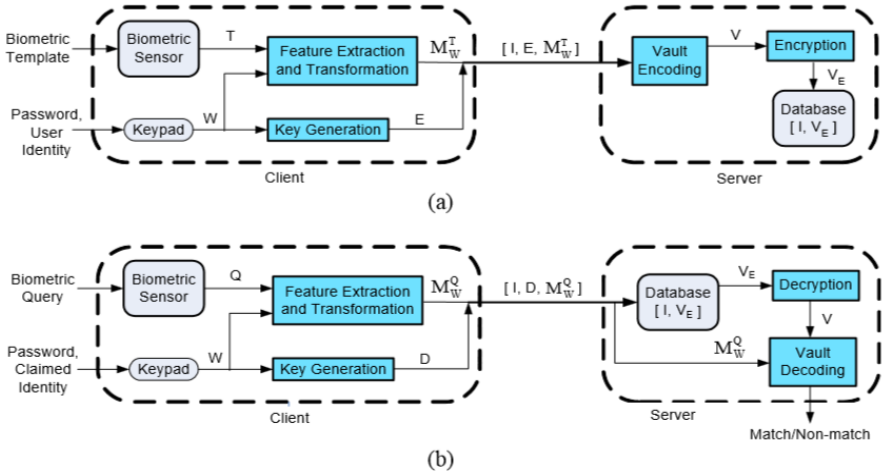


Fig. 5. Nandakumar’s schema

Enrollment Phases

User’s genius minutiae, receiving from fingerprint preprocessing module, are applied transformation by user-provided password when going through *Feature Extraction* and *Transformation* module. Transformed minutiae and secret key will be used as input for *Vault Encoding* to generate vault.

Authentication Phases

Within the case of genius users, when receiving query biometric as well as query password, transformation will be reapplied to generate nearly identical transformed minutiae. By using Lagrange interpolation and CRC check, vault decoding modules can output the previous secret key, thus, lead to authentication successful.

In *Figure 5*, *Key Generation* module generates key for encrypting vault. Generated keys are separated with secret key using in fuzzy vault schema. Furthermore, because of independence between password and secret key, applying password transformation appear to be an additional protection layer to current system without effecting current security level of fuzzy vault.

4.2 Nonrandom Chaff Point Generator Integration

For the encoding phase, after apply password transformation on user minutiae, random chaff points will be generated by apply chaff points generation algorithm on output of *Feature Extraction* and *Transformation* module and secret key k .

For the authentication phase, after receiving vault V from *Decryption* module and decoding template from *Feature Extraction* and *Transformation* module, fuzzy vault decoding algorithm will be applied. All result candidate keys will be verified by our chaff points verifier module.

4.3 Experiment Result

In order to testing our proposed nonrandom chaff point-based fuzzy vault system, we choose the FVC2004-DB1 as our main testing database. FVC2004-DB1 [15] is a public domain database with 800 images (100 fingers x 8 impressions/fingers) of size 640x480 and resolution 500 dpi. Evaluation of result base on two main criteria genius accept rate (GAR) and false accept rate (FAR). GAR represents the percentage of successful authentication of genius users. For testing GAR, one impression is used as encoding template and the other of same fingerprint are the decoding ones (totally we have $7 \times 8 \times 100 = 5600$ cases). On the other hand, FAR shows the amount of successful authentication of illegal users. Within this case, one impressions of each fingerprint will be chosen as encoding template. The decoding template will be random selected from impressions of other users. We will have $99 \times 100 = 9900$ cases for FAR testing.

Turning to basic parameter of fuzzy vault schema, we fix the number of genius points for each fingerprint as $g = 25$. The number of chaff point c is 10 times the number of genius points. Minimum distance of points in vault is $d = 10$. Finally, length of secret key k , which has significant effect to GAR and FAR, is chosen in order to result main polynomial p with degree n within the range of [8, 11].

Table 1. Genius Accept Rate (GAR), False Accept Rate (FAR) of Original Nandakumar's system and proposed system for FVC2004-DB1. Here n represent the degree of the polynomial used in vault encoding.

	n = 8		n = 9		n = 10		n = 11	
	GAR(%)	FAR(%)	GAR(%)	FAR(%)	GAR(%)	FAR(%)	GAR(%)	FAR(%)
<i>Original</i>	80%	9%	77%	6%	73%	2%	69%	0%
<i>Propose</i>	92%	7%	81%	13%	79%	3%	74%	0%

From *Table 1*, it can be proved that the proposed schema have significant enhancement on GAR, or decreasing FRR. The additional 16 bits CRC on secret key of original Nandakumar's system can be seen as the main reason for the different. Within same start secret key, polynomial generation module of Nandakumar's schema need to work with longer input (original secret key plus 16 bits CRC) and output a higher degree of polynomial. As the consequence, the system required additional genius points for a successful authentication.

Within current experiments, beside traditional evaluation of GAR and FAR, we also try to analyze space distribution between points in vaults of traditional and new schema. Mean value of distance between all points, genius points versus chaff points of both schema are calculated and result of evaluation for each schema are reported in table 2.

Particularly, within each system (*Nandakumar* and *Proposed* schemas), we calculate mean of distance between points of same vault by the below formula:

$$M = \frac{\sum_{1 \leq i \leq g+c} \sum_{1 < j < g+c} d(i, j)}{n(g+c)^2}$$

And mean of distance between genius points and their corresponding chaff points by:

$$m = \frac{\sum_{1 \leq i \leq g} \sum_{1 < j < c} d(i, j)}{n \times g \times c}$$

Where

- *n*: total number of vault (within our experiments, *n* = 100)
- *d(i,j)*: distance between two point *i*, *j* of vault
- *g*: number of genius points for each vault (within our experiments, *g* = 25)
- *c*: number of chaff points belonging to specific genius point (within our experiments, *c* = 10)

Table 2. a) Mean value of distance between all points in vaults generated from Nandakumar’s system and proposed system; b) Mean value of distance between all points in vaults and between genius points together with their corresponding chaff points of proposed system. Here we choose polynomial with degree as *n* = 8.

	n = 8			n = 8
<i>Nandakumar et al.</i>	209.33		<i>Vaults</i>	190.67
<i>Proposed system</i>	190.67		<i>Chaff per Genius</i>	189.47

a)

b)

As can be seen from Table 2, proposed system does not take effect on current distribution of points in vaults. Mean values of all points within vault are around 200. There is just a soft decrease when comparing proposed and original schemas. Considering proposed system, mean distance between chaff points and there corresponding genius points are almost the same with the one of all vault points. From that observation, it is quite difficult for attacker to apply statistical attack on vault generated by our proposed system.

5 Conclusions

Our contribution is addressing a significant problem of current CRC-based fuzzy vault system, the blend substitution attack. As a consequence, additional nonrandom chaff point generator and verifier modules are introduced. By applying continuously hashing and linear projection, new modules can detect any of modification from orig-

inal vault, thus eliminate blend substitution attack on traditional CRC-based fuzzy vault. Security analysis of new schemas and experiments result on FVC2004-DB1 are provided. The analysis and practical result show that the proposed algorithm does not lead to changing in current FRR and FAR of previous schema meanwhile total security level have significant enhancement. On the other hand, by demonstrating the integration of proposed modules with current Nandakumar et al. system, we show that new modules can be easily integrated with previous system and open for enhancement.

Acknowledgements. This research is funded by Vietnam National University Ho Chi Minh City (VNU-HCM) under grant number B2013-20-02. We also want to show a great appreciation to each member of D-STAR Lab (www.dstar.edu.vn) for their enthusiastic supports and helpful advices during the time we have carried out this research.

References

1. Rathgeb, C., Uhl, A.: A survey on biometric cryptosystems and cancelable biometrics. *EURASIP Journal on Information Security* 2011(1), 1–25 (2011)
2. Juels, A., Sudan, M.: A fuzzy vault scheme. *Designs, Codes and Cryptography* 38(2), 237–257 (2006)
3. Uludag, U., Pankanti, S., Jain, A.K.: Fuzzy vault for fingerprints. In: Kanade, T., Jain, A., Ratha, N.K. (eds.) *AVBPA 2005*. LNCS, vol. 3546, pp. 310–319. Springer, Heidelberg (2005)
4. Uludag, U., Jain, A.: Securing fingerprint template: Fuzzy vault with helper data. In: *Conference on Computer Vision and Pattern Recognition Workshop, CVPRW 2006*. IEEE (2006)
5. Nandakumar, K., Jain, A.K., Pankanti, S.: Fingerprint-based fuzzy vault: Implementation and performance. *IEEE Transactions on Information Forensics and Security* 2(4), 744–757 (2007)
6. Nandakumar, K., Nagar, A., Jain, A.K.: Hardening fingerprint fuzzy vault using password. In: Lee, S.-W., Li, S.Z. (eds.) *ICB 2007*. LNCS, vol. 4642, pp. 927–937. Springer, Heidelberg (2007)
7. Poon, H.T., Miri, A.: On efficient decoding for the Fuzzy Vault scheme. In: *2012 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA)*. IEEE (2012)
8. Clancy, T.C., Kiyavash, N., Lin, D.J.: Secure smartcard based fingerprint authentication. In: *Proceedings of the 2003 ACM SIGMM Workshop on Biometrics Methods and Applications*. ACM (2003)
9. Mihailescu, P.: The fuzzy vault for fingerprints is vulnerable to brute force attack. arXiv preprint [arXiv:0708.2974](https://arxiv.org/abs/0708.2974) (2007)
10. Chang, E.-C., Shen, R., Teo, F.W.: Finding the original point set hidden among chaff. In: *Proceedings of the 2006 ACM Symposium on Information, Computer and Communications Security*. ACM (2006)
11. Poon, H.T., Miri, A.: A Collusion Attack on the Fuzzy Vault Scheme. *ISeCure* 1(1) (2009)

12. Scheirer, W.J., Boulton, T.E.: Cracking fuzzy vaults and biometric encryption. In: Biometrics Symposium. IEEE (2007)
13. Hartloff, J., et al.: Security analysis for fingerprint fuzzy vaults. In: SPIE Defense, Security, and Sensing. International Society for Optics and Photonics (2013)
14. Örencik, C., et al.: Improved fuzzy vault scheme for fingerprint verification, pp. 37–43 (2008)
15. Maio, D., Maltoni, D., Cappelli, R., Wayman, J.L., Jain, A.K.: FVC2004: Third fingerprint verification competition. In: Zhang, D., Jain, A.K. (eds.) ICBA 2004. LNCS, vol. 3072, pp. 1–7. Springer, Heidelberg (2004)

Smart Card Based User Authentication Scheme with Anonymity

Toan-Thinh Truong¹, Minh-Triet Tran¹, and Anh-Duc Duong²

¹University of Science, Hochiminh city, VNU-HCM
{ttthinh, tmtriet}@fit.hcmus.edu.vn

²University of Information Technology, Hochiminh city, VNU-HCM
ducda@uit.edu.vn

Abstract. Mobile devices (e.g., PDA, mobile phone, tablet, and notebook PC) become necessary for a convenient and modern life. So, we can use them to access services, for examples online shopping, internet banking. In such insecure environment, we see that communications are more and more essential because they defend users and providers against illegitimate adversaries. Recently, Shin et al have proposed scientific paper entitled 'A Remote User Authentication Scheme with Anonymity for Mobile Devices' to enhance security for remote user authentication. They claimed that their scheme is truly more secure than previous ones and it can resist various attacks. However, it is not true because their scheme's vulnerable to insider, impersonation and replay attacks. In this paper, we present an improvement to their scheme to isolate such problems.

Keywords: Authentication, Dynamic ID, Impersonation, Session key.

1 Introduction

With rapid growth of mobile devices, such as smart-phones, tablets, or even wearable devices, electronic transactions are more and more widely deployed on them. Therefore, the users easily access online services at anytime and anywhere [6,8,13,4,7]. In reality, when the users upload some data to cloud storage servers, they want to keep secret the content. Furthermore, in telecare-medicine services, servers need assuring security and integrity for customers information when communicating with them. Or in mobile-banking services, user anonymity is the most important aspect servers need to achieve [3,11]. Clearly, while the users communicate with online services through mobile devices, such as smart-phones or wearable devices, both servers and users need to be protected from popular attacks, such as impersonation or information leaking [9,15,10]. Authentication scheme must be carefully designed to ensure security criteria and have low-computational cost. Some power and capacity-limited devices need light-weight protocols which do not have modular exponentiation to decrease energy-using and enhance time-using [2,16,18,19].

In this paper, we propose new method for user authentication while accessing online services. Also, our scheme supports for energy-limited wearable devices

because of having fixed-computational cost which Shin et al's scheme [14] cannot provide. Instead of following traditional approaches such as bilinear-pairing or elliptic curve operations, we take advantage of hash function combined with cryptography algorithms for withstanding popularly-known attacks, such as replay [5], impersonation [1], reflection [17], or parallel session attacks [12], etc... It is said that proposed scheme not only overcomes some limitations existing Shin et al.'s, but also enhances security and is suitable for resource-limited mobile devices.

The remainder of this paper is organized as follows: section 2 quickly reviews Shin et al.'s scheme and discusses its weak points. Then, our proposed scheme is presented in section 3, while section 4 discusses the security and efficiency of the proposed scheme. Our conclusions are presented in section 5.

2 Review and Cryptanalysis of Shin et al.'s Scheme

In this section, we review Shin's scheme [14] and analyze it on security aspect.

2.1 Review of Shin et al.'s Scheme

Their scheme includes four phases: registration, login and authentication, key agreement and secure password update phases. Some important notations in this scheme are listed as follows:

Table 1. The notations used in the proposed scheme

Notations	Description
U_i	User i .
PW_i	Unique password of U_i .
S	The remote server.
K_S	The secret key of server.
K_U	The common key of user for S .
TID_i	The transformed identity of U_i .
$CTID_i$	The changed identity of U_i .
DID_i	The dynamic identity of U_i .
DID_S	The dynamic identity of S .
SK_i	The generated session key of U_i .
SK_U	The generated session key of S .
$h(.)$	A one-way hash function.
$h^k(A)$	Perform hash function of k times.
\oplus	The bitwise exclusive-or operation.
\parallel	The string concatenation operation.
$A \Rightarrow B: M$	A sends M to B through a secure channel.
$A \rightarrow B: M$	A sends M to B through a common channel.
$E_{SK_i}\{M\}$	Encrypted message by the session key, SK_i .

2.2 Registration Phase

When U_i wants to access a remote server from a service legally, U_i performs the following registration steps before the access. The procedure is as follows:

- Step 1. $U_i \Rightarrow S: ID_i, h(PW_i)$.
 U_i freely chooses his/her ID_i and password PW_i for registration and submits ID_i and $h(PW_i)$ hashed value of PW_i to S via secure channel.
- Step 2. After receiving ID_i and $h(PW_i)$ from U_i , S performs:
 1. Generate $TID_i = h(ID_i \parallel h(PW_i))$, and check the existence of TID_i in the database. If the identity already exists in the database, S requests U_i to re-initiate the registration procedure with different ID_i or PW_i . Otherwise, S stores TID_i in the database. This process ensures the uniqueness of the user's TID_i .
 2. Compute $A_i = h(K_U) \oplus K_S$, where K_S is a secret key of S and K_U is user's common key for S . K_U is used to generate DID_i in the login and authentication phase.
 3. Compute $B_i = (g^{A_i} \text{ mod } p) \oplus h(PW_i)$, where g is primitive element in Galois field $GF(p)$ and p is large prime positive integer.
 4. Store $TID_i, B_i, h(\cdot)$ and K_U in smart card and issue it to U_i .

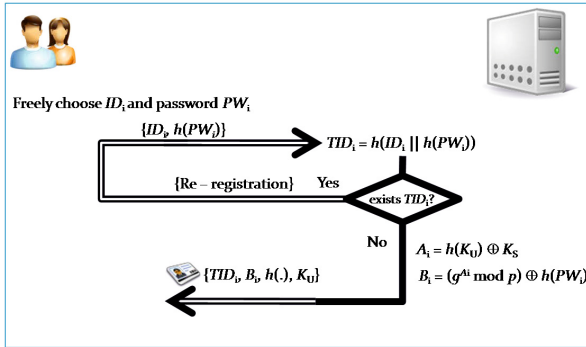


Fig. 1. Shin et al.'s registration phase

In their registration phase, we see that there are two advantages: another user can choose PW_i and ID_i freely. Furthermore, user also can hide his/her password from server by sending a hash value $h(PW_i)$ instead of only PW_i . And at these points, our scheme proposed later completely inherits them. However, A_i in their scheme is a combination of K_U and K_S , which are common for all different users. And this is a dangerous point we need to consider. Therefore, our scheme will supply a random value e for each user's registration.

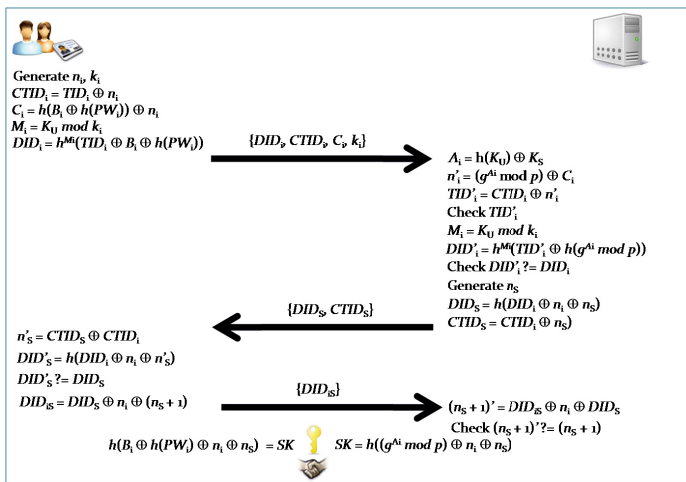


Fig. 2. Shin et al's login and authentication phase

Login and Authentication Phase. After registering to S , U_i sends a login message to S . The login message contains DID_i to protect anonymity. After verification of the login message, U_i can authenticate S and vice versa.

- Step 1. $U_i \rightarrow S$: $DID_i, CTID_i, C_i, k_i$
 U_i inserts his/her smart card to a card reader. He/she inputs ID_i and PW_i . The smart card performs the following steps
 1. Generate nonce, n_i and k_i
 2. Compute $CTID_i = TID_i \oplus n_i$, $C_i = h(B_i \oplus h(PW_i)) \oplus n_i$, $M_i = K_U \text{ mod } k_i$ and $DID_i = h^{M_i}(TID_i \oplus B_i \oplus h(PW_i))$
 3. U_i sends $DID_i, CTID_i, C_i$ and k_i with the login request message to S
- Step 2. $S \rightarrow U_i$: $DID_S, CTID_S$
 S does the following steps to authenticate U_i
 1. Compute $A_i = h(K_U) \oplus K_S$, $n_i' = C_i \oplus h(g^{A_i} \text{ mod } p)$ and $TID_i' = CTID_i \oplus n_i'$.
 2. Then S checks that TID_i' is the registered transform identity in the database. S terminates the connection if TID_i' is not valid; otherwise, S continues the process.
 3. Then, S computes $M_i = K_U \text{ mod } k_i$ and $DID_i' = h^{M_i}(TID_i \oplus B_i \oplus h(PW_i))$
 4. S compares the received value, DID_i and the generated value, DID_i' , S authenticates the legal user, U_i . Otherwise, S fails authentication of U_i and S terminates the connection with U_i
 5. S generates nonce n_s and computes $DID_S = h(DID_i \oplus n_i \oplus n_s)$ and $CTID_S = CTID_i \oplus n_s$
 6. Finally, S sends DID_S and $CTID_S$ to U_i

– Step 3. $U_i \rightarrow S$: DID_{iS}

U_i authenticates S and mutual authentication is completed according to the following steps.

1. Compute $n_{S'} = CTID_S \oplus CTID_i$.
2. U_i computes $DID_{S'} = h(DID_S \oplus n_i \oplus n_{S'})$ and compares the received value, DID_S . If $DID_{S'} = DID_S$, U_i authenticates S . Otherwise, U_i fails server authentication and terminates connection with S .
3. U_i computes $DID_{iS} = DID_S \oplus n_i \oplus (n_S + 1)$ and sends DID_{iS} to S
4. S computes $(n_S + 1)' = DID_{iS} \oplus n_i \oplus DID_S$, compares the value, $(n_S + 1)$ and the generated value $(n_S + 1)'$. If $(n_S + 1)' = (n_S + 1)$, mutual authentication is complete. Otherwise, S terminates connection with U_i

In their login and authentication phase, we see that user generates two random values n_i and k_i to challenge S . This guarantees no one except S can know transformed identity of U . However, drawback of this phase is that anyone can also compute TID_i of U from message packages transmitted in this phase, so we will fix this weak point of their phase.

2.3 Key Agreement Phase

After authenticating successfully in login and authentication phase, U_i and S compute common SK . U_i computes $SK_i = h(B_i \oplus h(PW_i) \oplus n_i \oplus n_k)$ and S computes $SK_S = h((g^{A_i} \bmod p) \oplus n_i \oplus n_S)$. We see that SK_i and SK_S are equal to each other because $(g^{A_i} \bmod p)$ and $(B_i \oplus h(PW_i))$ are the same.

Secure Password Update Phase. When U_i wants to change his/her password, he/she can change it freely. Following are steps to change.

1. $U_i \rightarrow S$: $DID_i, CTID_i, C_i, k_i, M_{request-change-PW_i}$
 U_i inserts the smart card into card reader and sends $DID_i, CTID_i, C_i$ and k_i with $M_{request-change-PW_i}$, the request message.
2. Mutual authentication is acted between U_i and S , as in the login and authentication phase.
3. U_i types new password, PW_i^* and computes $TID_i^* = h(ID_i \oplus h(PW_i^*))$.
4. $U_i \rightarrow E_{SK_i}\{TID_i^*\}$
 U_i encrypts new TID_i^* using SK_i and sends this message to S .
5. S decrypts the received message using SK_S and then replaces the value TID_i with the received value TID_i^* . S sends the response message to U_i .
6. After receiving the response message from S , U_i computes $B_i^* = B_i \oplus h(PW_i) \oplus h(PW_i^*)$ and replaces stored values in the smart card, TID_i and B_i with TID_i^* and B_i^* with each other.

We see that the security of their secure password update phase is completely based on the security of login and authentication phase. In the next section, we will prove their login and authentication phase is insecure. Therefore, their password update phase is insecure too.

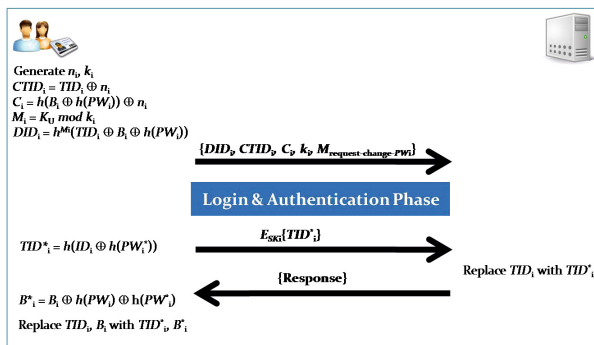


Fig. 3. Shin et al.'s secure password update phase

3 Cryptanalysis of Shin et al.'s Scheme

In this subsection, we present our results on Shin et al.'s scheme. We will show that their scheme is vulnerable to replay, stolen-verifier, impersonation from other valid user's attacks. Besides, their scheme does not provide forward secrecy.

Replay Attack. Shin et al claimed that their scheme is secure against replay attack due to secrecy of two random values n_i and n_S . This is not true because we will show the way of computing these two random values from message packages eavesdropped. Assuming that $\{DID_i, CTID_i, C_i, k_i\}$, $\{DID_S, CTID_S\}$ and DID_{iS} are spied. Attacker A will perform following steps to obtain n_i and n_S .

- A gets n_S by performing $CTID_S \oplus CTID_i$ and n_i by performing $DID_{iS} \oplus DID_S \oplus (n_S + 1)$.
- A has n_i , so A can use the previous message $\{DID_i, CTID_i, C_i, k_i\}$ constantly. For example, A re-sends $\{DID_i, CTID_i, C_i, k_i\}$ to S . Then, S replies $\{DID_S, CTID_S\}$ and A easily computes n_S by performing $CTID_S \oplus CTID_i$. Finally, A computes $DID_{iS} = DID_S \oplus n_i \oplus (n_S + 1)$.

Clearly, we see that everything's valid. So we conclude that Shin et al.'s scheme is insecure against to replay attack.

Stolen-Verifier Attack. Shin et al claimed that their scheme is secure against stolen-verifier attack because only malicious insider or intruder gets the table of the user's transformed identity. This is not true because we will show the way of computing the value TID_i from message packages eavesdropped. Assuming that $\{DID_i, CTID_i, C_i, k_i\}$, $\{DID_S, CTID_S\}$ and DID_{iS} are spied. Attacker A will perform following steps to obtain TID_i .

- A computes $n_S = CTID_S \oplus CTID_i$ and $n_i = DID_{iS} \oplus DID_S \oplus (n_S + 1)$.
- A has n_i , so A can achieve $TID_i = CTID_i \oplus n_i$.

Obviously, we see that U_i 's TID_i will be leaked by computing information of packages eavesdropped. So we conclude that Shin et al.'s scheme is insecure against to stolen-verifier attack.

Impersonation from Other Valid Users Attack. Shin et al claimed that their scheme is secure against impersonation attack. This is not completely true because we will show the way of attack from other valid users. Assuming that $\{DID_i, CTID_i, C_i, k_i\}, \{DID_S, CTID_S\}$ and DID_{iS} are spied. Another U_i will perform following steps to obtain TID_i .

- U_i computes $n_S = CTID_S \oplus CTID_i$ and $n_i = DID_{iS} \oplus DID_S \oplus (n_S + 1)$.
- U_i has n_i , so U_i can achieve TID_i of victim by performing $CTID_i \oplus n_i$.
- With victim's TID_i , U_i uses it to start computing $\{DID_i, CTID_i, C_i, k_i\}$ to masquerade victim. Firstly, U_i generates two random values n_i and k_i . U_i computes $CTID_i = TID_i \oplus n_i$. Secondly, U_i computes $C_i = h(B_i \oplus h(PW_i))$ with PW_i and B_i belonging to U_i but not victim. Finally, U_i computes $M_i = K_U \text{ mod } k_i$. So U_i has a valid login message to fake victim.

Therefore we conclude Shin et al.'s scheme is insecure against to impersonation from other valid users' attacks.

Inability to Provide Forward Secrecy. Shin et al claimed that their scheme provide forward secrecy due to the fact that no one can derive two random values n_i and n_S . This is not true because in above section we show the way of detecting them. So, if the long-term secret key material is revealed to an adversary, all previous session keys will be easily leaked.

4 Proposed Scheme

In this section, we will propose revised scheme of Shin et al's scheme that removes the security problems described in the previous section. Our improved scheme not only inherits the advantages of their scheme, it also enhances the security. Before entering into each phase, we will present general ideas in our scheme more detailed. In registration phase, our main goal is achieving $h(K_U \parallel e)$ and $h(K_S \parallel TID_i \parallel e)$. Random value e helps to register with the same identity but various authentication keys at different time. In login and authentication phases, we only use two random values n_i and n_S for server and user to challenge each other. Furthermore, we employ three-way challenge-response handshake technique to resist replay or impersonation attacks. And it is very important to have the same session key for user and server after authenticating successfully. Our scheme is also divided into the four phases of registration, login, mutual authentication and secure password update phase.

4.1 Registration Phase

Before we continue to present, we list three requirements for a registration phase: secrecy for information transmitted between user and server, the true password

of user should not shown to anyone even the server, and difference between keys provided for each time of registration by server. Easily, we see that Shin et al.'s scheme achieved first two requirements but not the last. So, we will recover this point to accomplish a good registration phase. When U_i wants to register to S , he/she has to submit his/her $ID_i, h(PW_i \parallel N)$, where N is a nonce chosen by U_i . Figure 4 illustrates the steps of registration phase.

1. U_i freely chooses ID_i, PW_i and generates a random value N .
2. U_i sends $\{ID_i, h(PW_i \parallel N)\}$ to S via secure channel.
3. S computes $TID_i = h(ID_i \parallel h(PW_i \parallel N))$ and checks its existence. If TID_i does not exist, S goes to next step. Otherwise, S requests U_i to re-register.
4. S computes $T_i = h(h(PW_i \parallel N) \parallel TID_i)$ and generates e .
5. S computes $A_i = h(K_U \parallel e) \oplus T_i, B_i = h(K_S \parallel TID_i \parallel e) \oplus T_i$ and $K_i = h(h(K_U \parallel e) \parallel h(K_S \parallel TID_i \parallel e))$.
6. S issues a smart card $\{K_i, A_i, B_i, h(\cdot), e\}$ to user U_i via secure channel.
7. U_i inserts N into smart card.

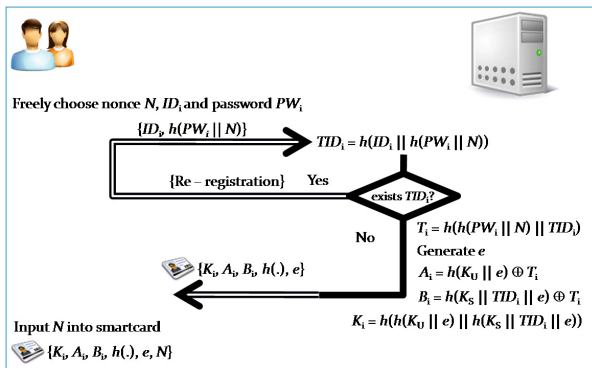


Fig. 4. Proposed registration phase

4.2 Login Phase

The U_i types his/her ID_i, PW_i to login S , and then the smart card performs:

1. Compute $TID_i = h(ID_i \parallel h(PW_i \parallel N)), R_i = h(h(PW_i \parallel N) \parallel TID_i), K_1 = R_i \oplus A_i$ and $K_2 = R_i \oplus B_i$.
2. Check $K_i \stackrel{?}{=} h(K_1 \parallel K_2)$. If the equation holds, smart card goes to the next step. Otherwise, it terminates the session.
3. Generate a random value n_i .
4. Compute $CTID_i = n_i \oplus TID_i, C_i = K_1 \oplus n_i$ and $DID_i = h(K_2 \parallel TID_i \parallel K_1 \parallel n_i)$.
5. Send $\{CTID_i, DID_i, C_i, e\}$ to S via common channel.

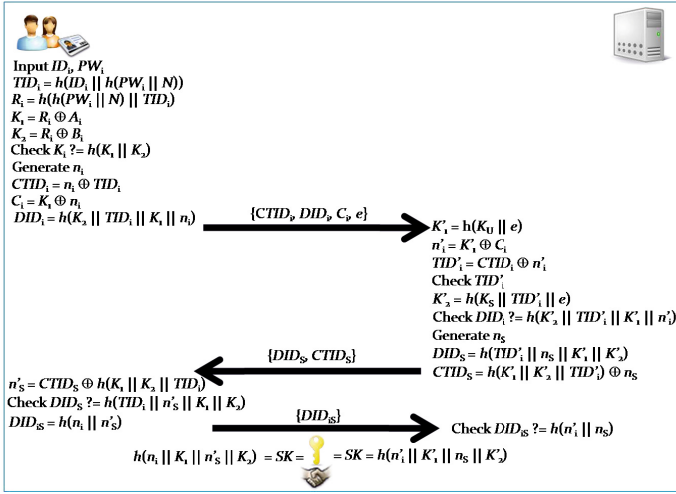


Fig. 5. Proposed login, mutual authentication and session key agreement phase

4.3 Mutual Authentication and Session Key Agreement Phase

Similarly, we propose three requirements that help authentication be more secure: user must use a random value to challenge server, server must use a random value to re-challenge user. And user and server share a secret session key. In Shin et al.'s scheme, their design is vulnerable to break and anyone can detect these random values. In this section, S will receive the login request message $(CTID_i, DID_i, C_i, e)$ from U_i in the login phase. Figure 5 illustrates the steps that S authenticates U_i .

1. S computes $K'_1 = h(K_U || e)$, $n'_i = K'_1 \oplus C_i$ and $TID'_i = CTID_i \oplus n'_i$.
2. S checks the validity of TID'_i in database. If everything is alright, S goes to the next step. Otherwise, S terminates the session.
3. S computes $K'_2 = h(K_S || TID'_i || e)$
4. S checks $DID_i \stackrel{?}{=} h(K'_2 || TID'_i || K'_1 || n'_i)$. If the equation holds, S goes to the next step. Otherwise, S terminates the session.
5. S generates a random value n_s .
6. S computes $DID_s = h(TID'_i || n_s || K'_1 || K'_2)$ and $CTID_s = h(K'_1 || K'_2 || TID'_i) \oplus n_s$.
7. S sends $\{DID_s, CTID_s\}$ to U_i via common channel.
8. After receiving $S'\{DID_s, CTID_s\}$ to U_i , U_i computes $n'_s = CTID_s \oplus h(K_i || K_s || TID_i)$.
9. U_i checks $DID_s \stackrel{?}{=} h(TID_i || n'_s || K_i || K_s)$. If the equation holds, U_i accepts S and computes $SK = h(n_i || K_i || n'_s || K_s)$. Otherwise, U_i terminates the session.
10. U_i computes $DID_{iS} = h(n_i || n'_s)$, sends DID_{iS} to S via common channel.
11. After receiving $\{DID_{iS}\}$, S checks $DID_{iS} \stackrel{?}{=} h(n'_i || n_s)$. If the equation holds, S accepts U_i and computes $SK = h(n'_i || K'_1 || n_s || K'_2)$. Otherwise, S terminates the session.

4.4 Secure Password Update Phase

In Shin et al’s scheme, we see that the security of this phase is completely based on the security of login and authentication phase. So, their password update phase is not enough confident. Consequently, we will recover this phase with many adjustments to improve security. If U_i would like to change password for some reasons, he/she must perform login and authentication phase firstly. After authenticating successfully, S will replace old TID_i with new TID^*_i . Then S sends $E_{SK_i}\{h(K_S \parallel TID^*_i \parallel e), Response\}$ to U_i . After receiving package from S , U_i will perform following steps:

- The smart card computes $h(K_U \parallel e) = A_i \oplus T_i$.
- The smart card decrypts message from S by using session key SK_i to obtain $h(K_S \parallel TID^*_i \parallel e)$.
- Then, it derives $K^*_i = h(h(K_U \parallel e) \parallel h(K_S \parallel TID^*_i \parallel e))$.
- Smart card continues to computes $T^*_i = h(h(PW^*_i \parallel N) \parallel TID^*_i)$, $A^*_i = h(K_U \parallel e) \oplus T^*_i$ and $B^*_i = B_i \oplus T_i \oplus T^*_i$.
- Finally, smart card replaces B_i, A_i, K_i with B^*_i, A^*_i, K^*_i

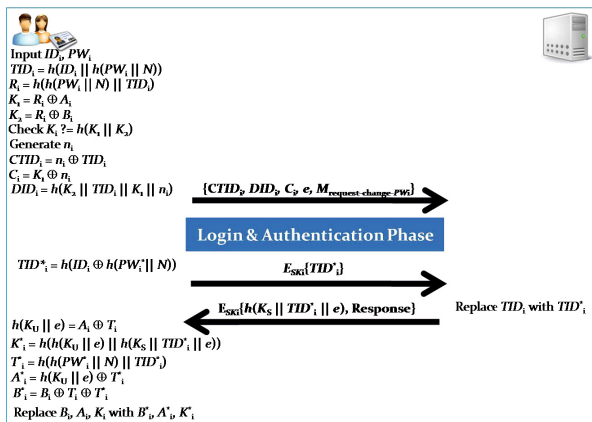


Fig. 6. Proposed secure password update phase

5 Security and Efficiency Analysis

In this section, we analyze our scheme on two aspects: security and efficiency.

5.1 Security Analysis

In this subsection, we present these security analyses of our scheme and show that proposed scheme can resist many kinds of attack. Assume that wireless communications are insecure and there exists an attacker. He/she has capability to intercept all messages communicated between server and user. Furthermore, we assume the attacker can obtain or steal information of user’s mobile device.

Replay Attack. The replay attack is replaying the same message of the receiver or the sender again. Our scheme uses nonce and three-way challenge-response handshake technique instead of time stamp to withstand replay attacks. For example, another attacker A resends $\{CTID_i, C_i, DID_i, e\}$ to S . Then, S will send $\{DID_S, CTID_S\}$ to attacker A . Without knowing ID_i, PW_i and nonce N of valid user, attacker A cannot compute K_1 and K_2 to obtain DID_{iS} to send to S . So, S recognizes someone is impersonating U_i and S will terminate the session. Therefore, our scheme can frustrate this kind of attack.

Impersonation Attack. In our scheme, if another attacker A would like to fake U_i to cheat S , he/she must have user's ID_i, PW_i and random value N . In addition, A must also have smart card of user to extract $h(K_U \parallel e)$ and $h(K_S \parallel TID_i \parallel e)$ to construct $\{CTID_i, C_i, DID_i, e\}$. Clearly, attacker A has no chance to do this. And if attacker A would like to fake S to cheat U_i , A must have K_U and K_S of server to build $\{DID_S, CTID_S\}$. Obviously, there aren't also no chance for A to perform this task. Consequently, our scheme can resist this kind of attack.

Stolen Verifier Attack. In our scheme, S maintains users' TID_i . If an attacker A would like to steal TID_i , A only computes information based packages between U_i and S . But in our scheme, A needs to have $h(K_U \parallel e)$ to know random value n_i . And from this random value n_i , A can compute $TID_i = C_i \oplus h(K_U \parallel e)$. Clearly, A has no way to have $h(K_U \parallel e)$. So our scheme can withstand this kind of attack.

Stolen Information from Smart Card. Assuming that users lose their smart card SC , an attacker A can extract information from it to harm users. In our scheme, if SC containing $\{K_i, A_i, B_i, h(\cdot), e, N\}$ of another user is stolen by A . He/she cannot extract any information from this SC . Hence, our scheme is immune from this kind of attack.

Password Guessing Attack. In our scheme, the hash value of password of user consists of PW_i and random value N . So it is very difficult for server to predict user's password. Towards people from outside, we see that they achieve nothing from packages transmitted from user and server. In case of losing smart card, people from outside also have no way to infer password even the hash value of it. Thereupon, our scheme can counteract this kind of attack.

Our scheme is a revised version of Shin et al.'s scheme, so it can also resist known-key, insider attacks. Additionally, our scheme provides mutual authentication and user anonymity. However, we still cannot find the way to provide forward secrecy if long-term secret key material is revealed to adversary. If K_S and K_U are leaked and adversary A has previous packages, A easily re-computes two random values n_i, n_S and $h(K_U \parallel e), h(K_S \parallel TID_i \parallel e)$ to know SK .

5.2 Efficiency Analysis

To compare efficiency between our scheme and the previous scheme proposed by Shin et al, we reuse approach used in that previous scheme to analyze

computational complexity. That is, we calculate the number of one-way hash function execution. Let H be one-way hash function, C be concatenation, \oplus be bitwise exclusive, M be modular exponentiation, E be encryption, D be decryption, A be arithmetic operation, O be comparison. In table 2, there are our scheme and Shin et al’s scheme. Shin et al.’s scheme needs $1C, 3H, 2\oplus$ and $1M$ in registration phase, and $8H+, 16\oplus, 1M, 2A, 3O$ in login and authentication phase. Our scheme needs $7C, 6H$ and $2\oplus$ in registration phase, and $14H, 8\oplus, 24C, 4O$ in login and authentication phase.

Table 2. A comparison of computation costs

Schemes	Registration	Login and Authentication
Shin et al [14]	$1C, 3H, 2\oplus, 1M$	$8H+, 16\oplus, 1M, 2A, 3O$
Our	$7C, 6H, 3\oplus$	$24C, 14H, 8\oplus, 4O$

Proposed scheme almost needs more computational amount than Shin et al.’s scheme. However, our scheme has stable quantity of operations while Shin et al’s scheme must depend on random value $M_i = K_U \text{ mod } k_i$ to determine how many operations do their scheme has. Furthermore, their scheme uses modular exponentiation which cost too much. So, our scheme is still better than their scheme. In short, proposed scheme not only reduce computational costs but also enhances security.

In table 3, we list the comparisons between our improved scheme and Shin et al.’s scheme for withstanding various attacks. We see that Shin et al.’s scheme cannot resist to impersonation, stolen verifier and replay attacks. It can be seen that our proposed scheme is more secure against various attacks.

Table 3. Comparison between our scheme and the Shin’s for withstanding various attacks

	Shin et al [14]	Our
Replay attack	No	Yes
Impersonation attack	No	Yes
Stolen verifier attack	No	Yes
Password guessing attack	Yes	Yes
Insider attack	Yes	Yes
Known-key attack	Yes	Yes
Mutual authentication	Yes	Yes
Session key agreement	Yes	Yes
Forward secrecy	No	No
Stolen information from SC	No considering in scheme	Yes

6 Conclusions

In this paper, we review remote user authentication scheme with anonymity for mobile device scheme of Shin et al. Although their scheme can withstand some attacks, such as insider, known-key attacks. However, we see that their scheme is still vulnerable to replay attack, impersonation and stolen verifier attacks. Moreover, their scheme cannot guarantee forward secrecy. Wherefore, we propose an improved scheme to eliminate such problems.

References

1. Chen, T.H., Chen, Y.C., Shih, W.K., Wei, H.W.: An efficient anonymous authentication protocol for mobile pay-tv. *Journal of Network and Computer Applications* 34(4), 1131–1137 (2011)
2. Debiao, H., Jianhua, C., Jin, H.: An id-based client authentication with key agreement protocol for mobile client-server environment on ecc with provable security. *Information Fusion* 13(3), 223–230 (2012)
3. Hsiang, H.C., Shih, W.K.: Improvement of the secure dynamic id based remote user authentication scheme for multi-server environment. *Comput. Stand. Interfaces* 31, 1118–1123 (2009), <http://dl.acm.org/citation.cfm?id=1595894.1596057>
4. Hwang, M.S., Lee, C.C., Tang, Y.L.: A simple remote user authentication scheme. *Mathematical and Computer Modelling* 36, 103–107 (2002)
5. Islam, S.H., Biswas, G.P.: A more efficient and secure id-based remote mutual authentication with key agreement scheme for mobile devices on elliptic curve cryptosystem. *Journal of Systems and Software* 84(11), 1892–1898 (2011)
6. Lamport, L.: Password authentication with insecure communication. *Communications of the ACM* 24, 770–772 (1981)
7. Lee, C.C., Hwang, M.S., Yang, W.P.: Flexible remote user authentication scheme using smart cards. *IEEE Transactions on Neural Network* 36(3), 46–52 (2002)
8. Li, L.H., Lin, I.C., Hwang, M.S.: A remote password authentication scheme for multi-server architecture using neural networks. *IEEE Transactions on Neural Network* 12(6), 1498–1504 (2001)
9. Liao, I.E., Lee, C.C., Hwang, M.S.: Security enhancement for a dynamic id-based remote user authentication scheme. *IEEE Transactions on Consumer Electronics* 50, 629–631 (2004)
10. Liao, I.E., Lee, C.C., Hwang, M.S.: Security enhancement for a dynamic id-based remote user authentication scheme. In: *International Conference on Next Generation Web Services Practices*, vol. 6(2), pp. 517–522 (2005)
11. Liao, Y.P., Wang, S.S.: A secure dynamic id based remote user authentication scheme for multi-server environment. *Comput. Stand. Interfaces* 31, 24–29 (2009), <http://dx.doi.org/10.1016/j.csi.2007.10.007>
12. Khan, M.K., Kumari, S., Gupta, M.K.: More efficient key-hash based fingerprint remote authentication scheme using mobile device. *Computing* 96(9), 793–816 (2013)
13. Shen, J.J., Lin, C.W., Hwang, M.S.: A modified remote user authentication scheme using smart cards. *IEEE Transactions on Consumer Electronics* 49(2), 414–416 (2003)
14. Shin, S., Kim, K., Kim, K.H., Yeh, H.: A remote user authentication scheme with anonymity for mobile devices. *International Journal of Advanced Robotic Systems* 9, 1–7 (2012)

15. Wang, Y.Y., Kiu, J.Y., Xiao, F.X., Dan, J.: A more efficient and secure dynamic id-based remote user authentication scheme. *Computer Communications* 32, 583–585 (2009)
16. Yang, J.H., Chang, C.C.: An id-based remote mutual authentication with key agreement scheme for mobile devices on elliptic curve cryptosystem. *Computers and Security* 28(3-4), 138–143 (2009)
17. Yoon, E.-J., Yoo, K.-Y.: Improving the dynamic ID-based remote mutual authentication scheme. In: Meersman, R., Tari, Z., Herrero, P. (eds.) *OTM 2006 Workshops*. LNCS, vol. 4277, pp. 499–507. Springer, Heidelberg (2006)
18. Yoon, E.J., Yoo, K.Y.: Robust id-based remote mutual authentication with key agreement scheme for mobile devices on ecc. In: *IEEE International Conference on Computational Science and Engineering*, vol. 2, pp. 633–640 (2009)
19. Zhang, J., Deng, F.: The authentication and key agreement protocol based on ecc for wireless communications. In: *International Conference on Management and Service Science*, pp. 1–4 (2009)

Cloud-Based ERP Solution for Modern Education in Vietnam

Thanh D. Nguyen¹, Thanh T. T. Nguyen², and Sanjay Misra³

¹ HCMC University of Technology, Vietnam

thanh.nguyenduy@gmail.com

² Hoa Sen University, Vietnam

thanh.nguyenthithanh@hoasen.edu.vn

³ Covenant University, Nigeria

sanjay.misra@covenantuniversity.edu.ng

Abstract. Enterprise Resource Planning (ERP) and cloud computing are becoming more and more important in the World of Information Technology (IT) and Communication. These are two different sectors of modern information systems, and there are several in-depth investigations about ERP and also cloud computing. Recently, there have been some studies on ERP in cloud computing, but not much work as regards its applications in the field of education has been done. Besides that, deploying traditional ERP systems can be challenging and often costly and resource intensive. However, with the emergence of cloud-based ERP solutions, it has lower cost implications than traditional ERP. In this scenario, implementing a study on cloud-based ERP for modern education is an important and beneficial work. By considering this point, the objective of this research is to approach the relevant concepts multi-dimensionally; illustrating the advantages and challenges of cloud-based ERP on education and proposing a cloud-based ERP solution that can apply to any educational institution in Vietnam.

Keywords: Cloud computing, Enterprise Resource Planning, Information System, Information Technology, modern education, Vietnam.

1 Introduction

Cloud computing is a concept that has been researched more and more in some recent years. However, it is not an entirely new technology. The services and applications of cloud computing are growing steadily at a rate of about 40% per year [4]. Several authors have suggested that cloud computing would represent the future of IT usage in organizations or enterprises. For instance, Barnatt [3]; Velte et al. [40]; Zhang et al. [42] showed that the power of computer in cloud computing will profoundly have an influence on the IT industry. Organizations will not need to install software on their Information System (IS), and will not need to purchase hardware or software as these will be available for rent online. Besides that, Enterprise Resource Planning (ERP) is integrated into the packaged software with a common database, which support the operating procedures of the organizations [37]. Since ERP systems support the core

processes of organizational structures with many different scales and sectors, ERP has got the customization process to harmonize with a specific organization or enterprise, and often associated with other software systems. Therefore, this expedience must be resolved before deployment on cloud computing. While there are no doubt about the fact that cloud computing can bring many benefits in the field of the offices and the collaboration of group works [3], it is also interesting to consider the different active forms of a complex IS as ERP in cloud computing. Overall, both ERP and cloud computing can meet many requirements of the modern IS, and they are the development trend of the future [12].

On the other hand, most of the education systems in the countries around the World continue to research the optimal scale and educational method appropriate. In some countries, education is seen as a national policy. For example, in Vietnam, there are over 290 universities and colleges having training in IT and telecommunications. In 2012, the enrollment rate of this sector is about 11% of the target for the academic year; the students are studying at universities and colleges being about 170,000. Besides that, the Ministry of Education and Training is implementing several projects that relate to IT (e.g., Project of educational networks and IT applications in education and project of educational management computerization) [24]. These are favorable conditions for the modern educational organizations to perform the deployment of IT applications. Although many authors have researched on education (e.g., Marquez et al. [22]; Franc et al. [8]; Nguyen et al. [25, 26]; Deshmukh [6]; and Garcia et al. [10]), the majority did not propose any solution as regard ERP on cloud computing. Hence, the research of cloud-based ERP for modern education is a necessary work. The objective of this study is to approach the relevant concepts of cloud computing and ERP, discusses the advantages and challenges when deploying cloud-based ERP for education, and some case studies of the model of ERP in cloud computing were applied in some educational institutions. The research brings benefits not only to the educational institutions but also to the participants who work in the field of education. In addition, the study also contributes to the knowledge about ERP in cloud computing.

The paper is organized in 4 sections. The backgrounds of cloud computing, ERP, ERP in cloud computing, advantage and challenges for application of ERP in cloud computing, and cloud-based ERP vendors are summarized in section 2. Application of cloud-based ERP in modern education with reasons, a case study, and the E-EVN model are presented in section 3. Conclusions drawn is given in section 4.

2 Background

2.1 Cloud Computing

Cloud computing is becoming more popular in the ICT world and is bound to revolutionize it. The most significant reason for this is the search for an adaptive and dynamic IT infrastructure, which does not prevent business development [17]. If an organization wants to prepare their cloud computing strategy, they will need to know inherent capabilities that are supplied by cloud computing. By creating opportunities

for saving cost and organizational quickness, these capabilities can make the competitive advantage; these capabilities are determined such as interface, location dependence, sourcing independence, ubiquitous access, virtual business environments, addressability and traceability, and rapid elasticity. Not only does an organization save cost, but also reacts to changes in the environment quickly by implementing an optimal merge of these capabilities [14].

IT developers, IT managers and end users all understand cloud computing in different ways [19]. The National Institute of Standards and Technology (NIST) has defined that on demand self-service, broad network access, location independent resource pooling and payment for resources used, rapid elasticity, and measured service are the characteristics of the cloud-computing model [20]. Instead of installation and storage for each computing device, users only have to gain access to organize applications, personal applications, data storage, and remote processing power from the cloud, so there is no need for the data center. It is the trend of computing [3]. Figure 1 shows cloud components, which include applications, storage and networking, databases, and server. There are 3 main models of cloud computing, namely *Platform as a Service (PaaS)*, *Infrastructure as a Service (IaaS)*, and *Software as a Service (SaaS)* [34]. In the SaaS model, customer approaches applications as services offered via the Internet and payment is determined based on use. SaaS is a model based on multi users where each customer owns separate storage space but share the same program code. Software that needs to have many customizations and integration with other applications is not suitable for SaaS [38]. Mell & Grance [20] also described four cloud computing deployment models: private cloud, public cloud, community cloud, and hybrid cloud.

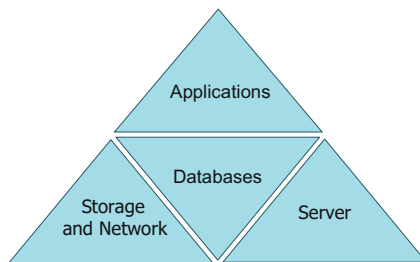


Fig. 1. Cloud Components (Source: Goel et al. [12])

The typical service providers of cloud computing, namely Google: Google Apps (SaaS), EMC: VMware (IaaS), Microsoft: Azure services (PaaS); Windows Live (SaaS), IBM: Lotus Live (SaaS), Amazon: Amazon DB (PaaS); Amazon EC2 (IaaS), and Salesforce: Cloud computing business (SaaS) [37].

2.2 ERP

The term ERP has appeared 1990s, and has gained popularity since then. It is packaged software, which integrates all information about finance, accounting, human resource and supply chain into one database. Vendors built up ERP systems and sold

it as standard software that met the needs of many organizations [5]. The structural model of ERP system is shown in Figure 2.

The ERP vendors, with the estimated market share are: SAP (30%), Oracle (21%), Sage (18%), Microsoft Dynamics (14%), and SSA (7%) [27]. Besides this, there are hundreds of ERP vendors for Small and Medium Enterprises (SME). The big companies have the models and different channels; for example, Microsoft has used indirect sales channels and built the networks of the partners to deploy Dynamics system [1]. In addition, there are many SaaS providers for individual customers and SMEs such as Zoho (database), Employ Ease (HRM), Clarizen (project management), Net Suite (business applications), and Web Office (collaboration tool) [3]. Meanwhile, some companies in Vietnam also have introduced solutions such as Fast: Fast Business Online, Neo: Cloud SME, VNPT: Mega ERP, and Viami: RVX Cloud.

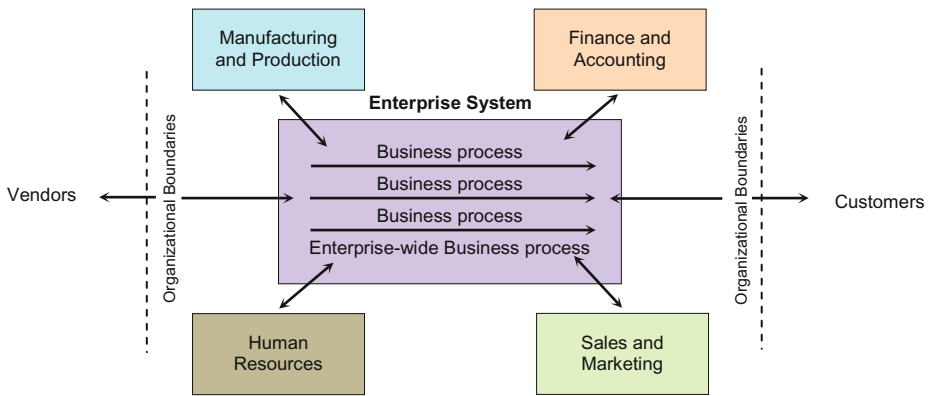


Fig. 2. The Structural Model of ERP System (Source: Laudon & Laudon [19])

2.3 ERP in Cloud Computing

ERP systems support the core business processes and have to reflect the organizational structure of the company. Thus, ERP solutions are tailored to meet the need of each industry [17]. SAP offered more than 25 industry-specific solutions, which provide the pre-customized system that adapted to the standard process used in the selected industries [30]. If an ERP system is deployed in the cloud, it will become an ERP in cloud computing. Virtualization and load balancing technologies allow applications to be deployed across multiple servers and database resources that are used in cloud environments [29]. Figure 3 illustrates 3 types of cloud-based ERP services:

(1) *IaaS for ERP*: Organization leases the computing resources from a cloud service provider, but the organization can still select the ERP vendor and pay for the software licenses. This characteristic makes a viable operational model, linking with ERP vendors or third parties and cloud service providers can make a combination to offer the users of the organizations. If an ERP vendor offers IaaS by themselves, they will create a vertical

integration [32]. (2) *PaaS for ERP*: This level provides an environment that accommodates for software development, testing, distribution software, but is not suitable for an ERP system [32]. (3) *SaaS for ERP*: The roles of ERP vendors and cloud service providers are combined as vertical integration in this model. ERP is provided as a service by the cloud service providers. Recently, cloud-based ERP versions have been developed and supplied by many ERP vendors. The organizations can select a solution that is appropriate for them (e.g., Running the ERP software in their own internal cloud or an external private cloud) [32]. Flexibility and effect on cost and in further consequence on the total cost of ownership (TCO) - which represents the difference between the classical ERP system and the SaaS solution [11].

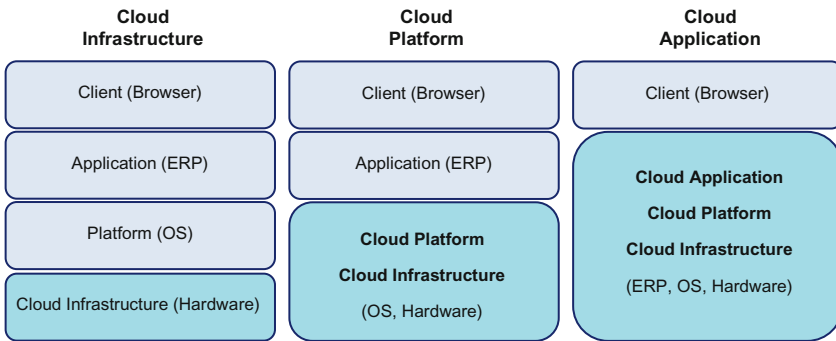


Fig. 3. 3 Types of Cloud-Based ERP Services (Source: Johnson [15])

Iyer & Hednderson [14] mentioned that the advantages of an organization make connections to their partners in what is known as the extended enterprise. The cloud strategy can also have industry structure or ecosystem impacting it in an increasingly network-based competition. Hence, cloud-based ERP will make big advantages in exchange of data with the extraneous organizations (B2B model) [4]. According to Panorama [28], the low cost of the cloud-based ERP solutions is a common selling point for cloud providers. The risk of security breaches is still barriers, which make the solicitude for organizations. However, Panorama has shown that the cloud-based ERP vendors typically provide secure and reliable solutions. When the board of management should envisage this factor during the selection of the software solutions. In fact, more than a half of the organizations, which use cloud-based ERP, confirm that they had saved roughly 40% of the cost [28].

The Advantages and Challenges of ERP in Cloud Computing

Of the three types of cloud computing services, SaaS is most interesting as relating to our subject matter. The benefits of SaaS are low cost, fast installation, and less hassle in maintenance. An ERP system runs on the web as a SaaS running in the cloud, in which organizations will be having the benefits while deployment. Accordingly, the cloud-based ERP providers can accommodate three types of services with the different benefits for the organization in Table 3. In addition,

Johnson [15] noted that besides the benefits that cloud-based ERP offers, drawbacks also exist. If the vendors do not have the ability to customize the applications, users will not have the option of moving an application to other providers. Moreover, educational institutions will precede many difficulties in the infrastructure management, because the vendors provide cloud computing being difference with ERP application service providers.

According to a survey of IMA (Management Accounting Institute of Montvale, USA), the main benefits of cloud-based ERP is total cost being low (about 30%), access data anytime and any anywhere (28%), streamline business processes (21%), easy to upgrade (9%), low storage volume (7%), and fast deployment speed (5%) [39]. Cloud-based ERP allows cost reduction of IT and increases usage efficiency, users only need a personal computer that has Internet access to maximize the efficiency of the applications.

Table 1. The Advantages and Challenges of SaaS

Offering	Advantages	Challenges
1 SaaS uses a Cloud Application	Maximizes efficiencies for “cookie cutter” applications.	Vendor “lock-in”, the customer does not have the option to move an application to a different provider.
2 SaaS uses a Cloud Platform	Mix of flexibility and savings.	Coordination challenges - vendor manages the application while a service provider manages infrastructure.
3 SaaS uses a Cloud Infrastructure	Maximizes flexibility to switch providers or move on-premise.	Some would argue this is nothing more than a hosted service with a slightly lower pricing structure.

Source: Johnson [15]

Besides that, some specific benefits of cloud-based ERP in modern education, such as (1) Reduced cost: The entirety of the hardware used is hosted on the cloud and belongs to the service providers. Thus, an organization does not need to own the hardware. The software costs are also reduced, as organizations using cloud-based ERP do not need to purchase the user licenses, but only pay for cloud-based ERP services [12]. (2) Scalability: Cloud-based ERP solutions can be easily customized and expanded; users can customize the interfaces or reports easily and add more database indexes [12]. (3) Unfettered access: Allowing faculty members, students, parents, administrators and vendors to log on to the systems with the rights of the users. This access can be made through any specific protocol via wired or wireless devices of their choice [12]. (4) Increased innovation: It is conducted through the use of open source software, all fields relating to education have benefited from the acceleration of innovation that is brought about from cloud computing [32]. (5) Less staff: it needs less business analysis specialists and IT staff as most of the ERP services including technical support can be handled by experts outside of the organization [34]. (6) Mobility: this allows users to access the ERP services through mobile devices (e.g., Laptops, tablets and smartphones) [12]. The cloud-based ERP benefits are summarized in Figure 4.

On the other hand, issues of IS security are major challenges for cloud-based ERP applications. However, currently there are many security solutions for both hardware and software applying to the Internet platform, and cloud computing has security standards that are of higher level than the Internet. Therefore, cloud computing service providers must provide the latest technologies together with the commitment to ensuring security. Also, considering the correlation between the benefits and the security risks, the organization can fully designate the secure cloud computing applications. Because the potential of cloud-based ERP services are stentorian [18].

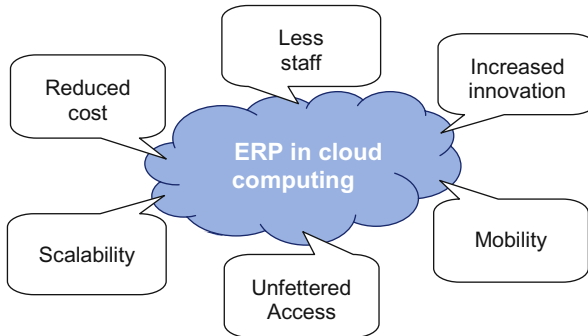







Fig. 4. The Benefits of ERP in Cloud Computing

Cloud-Based ERP Vendors

Cloud-based ERP vendors are becoming increasingly more and more competitive as the “SaaS wars” continue to expand across the World of enterprise software. The cloud is quickly becoming a dominant choice for deployment of ERP. As a result, the cloud-based ERP vendors have warmed up in recent years and it can be hard for users to separate, compare, and contrast all the different offerings [7]. Table 2 is the list of some cloud-based ERP vendors, some top vendors are Acumatica, Netsuite, Plex Systems, Ramco, and Workday.

Table 2. Top 5 Cloud-Based ERP Vendors

	Vendor		Operation	Deployment	Product
1	 Acumatica <small>THE CLOUD ERP</small>	Acumatica	Windows, Mac OS	Cloud or On-Premise	Distribution, Accounting
2	 NETSUITE	Netsuite	Windows, Mac OS	Cloud	Services
3	 PLEX <small>ONLINE</small>	Plex Systems	Windows, Mac OS	Cloud	Manufacturing
4	 ramco	Ramco	Windows, Mac OS	Cloud or On-Premise	Manufacturing
5	 workday.	Workday	Windows, Mac OS	Cloud	Human Resources, Payroll

Source: erpcloudnews.com [7]

3 Cloud-Based ERP for Modern Education

3.1 The Reasons for Implementing Cloud-Base ERP in Education Institutions

One of the important reasons for implementing cloud-based ERP for education is the service improvement for students from the beginning to the end of the courses, increasing the quality and transparency of data, and being controlled by the functional departments in the organization [6]. Besides that, helping the organizations to increase educational efficiency, improve curriculum quality and other resource activities can be effectuated in cloud computing [6], which will help the educational institutions be appreciated from the agencies of educational management.

In addition, the teaching and learning would be better to serve the number of applications for education as installed in cloud computing. Some reasons for implementing cloud-based ERP in educational institutions are shown in Table 3. Accordingly, although the replacement of the old system accounts approximately for 30%, this illustrates the need for innovation in education to meet the increasing development of modern technology; in fact it contributes about 70% to the raising of quality of education. Moreover, the improvement of services for teachers and learners; transforming institution operations in educational institutions; modernizing campus IT environment, can create competition among educational institutions while also improving the quality of modern education. On the other hand, increased efficiency and better teaching /learning process are also necessary reasons for the implementation of cloud-based ERP in educational institutions. The accountability or regulatory compliance has influence on, at least the deployment of IS in educational institutes.

Table 3. Reasons for Implementing ERP Systems in Education Institutions

	Elements	Percentage (%)
1	Replacing aging legacy systems	30
2	Improve service to customers	21
3	Transforming institution operations	16
4	Modernize campus IT environment	12
5	Keep institution competitive	7
6	Increase efficiency	5
7	Better teaching learning process	5
8	Accountability/Regulatory compliance	4
	Total	100

Source: Deshmukh [6]

3.2 Cloud-Based ERP in Education

The university is a site to create the prerequisites for application development and IT support. Cloud computing is a natural technical progression of standards and

architectures Internet-based IT that fully exploits the economic scale and this change is strategic. However, those factors will slow adoption based on culture, organization, and management, rather than technical issues. Adoption will follow a standard technology model [22, 16].

Before that, deploying traditional ERP systems resist many challenges and often take much cost to deploy, from the investment in IT infrastructure to the cost in the software license. Thus, ERP is still far away for SMEs in Vietnam. However, with the emergence of cloud-based ERP solutions, SaaS has lower cost than traditional ERP. On this basis, universities and research institutes can offer cloud-based ERP into widespread usage in modern education. Evolution of ERP implementation in higher educational institutions includes four stages: (1) Traditional ERP implementation, (2) On-campus implementation, (3) Implementation with host Internet provider, and (4) cloud implementation. Besides these, improving the cost, maintenance and technical efficiency are helpful. In addition, there are many problems that need to be resolved (e.g., Proper planning, combination of the right resources and the right timing) [12].

In implementing campus ERP systems or on-premise ERP systems, organizations must overcome many challenges. Managing all these processes can be nightmarish [6]. SaaS and cloud-based ERP trends will surpass these challenges. However, if an organization selects a cloud-based ERP solution, they will face the attendant challenges. Thus, the ERP life cycle model discovered challenges of cloud-based ERP for each of four phases [13, 21]: (1) A cost-benefit analysis is useful for educational institutions, which helps organization see their future with total cost (e.g., Customization, integration, and maintenance) has perceived benefits. (2) Organizations need to evaluate the level of customization and integration because every organization is unique, and a software solution must support their process. The organization should consider the level of customization provided by cloud-based ERP providers. If an educational institution manages following a standard processes such as ISO, it will be advanced. There are two approaches to cloud-based ERP adjustment called configuration and customization. Configuration is normal set-up of the software, and does not require changes to the source code. Customization requires changes to the source code. (3) Organizations should plan system performance measures. (4) User adoption is an important phase to get full cloud-based ERP benefits. Users need to have some knowledge about cloud-based ERP technology. Fortunately, most users in educational institutions are students, lecturers, and researchers who have knowledge, skill and capability to cope with new trends quickly, so their adoption is more advanced.

Nguyen et al. [25] based on the Unified Theory of Acceptance and Use of Technology (UTAUT2) proposing a model of the acceptance and use of E-learning based on cloud computing in Vietnam. After that, another model with the roles of consumer innovativeness had been added into the acceptance and use of E-learning based on cloud computing [26]. Educational institutions and research institutes, the three-layer model of cloud-based ERP can be helpful for supplying general solutions. It is useful not only in teaching, learning, and research, but also managing education. Cloud-based ERP models in China and India bring significant benefits to educational institutions and users [37, 41]. Virtual laboratory model in Ho Chi Minh city university of technology showed that important problems that need to be resolved are cost and

payment, application store, software license, and cross-cloud platforms. Teaching requirement and research requirement in developing countries have several different characteristics [38].

3.3 Cloud-Based ERP Solution for Education

Hoa Sen Case Study

Hoa Sen, a private university in Ho Chi Minh city - Vietnam, one of the first universities to deploy and use ERP. Firstly, with available IT infrastructure, Hoa Sen has chosen and deployed the *Peoplesoft* solution since 2011. Over the time and effort to custom the ERP system appropriate to the needs of the university. Currently, *Peoplesoft* system is used online for two goals that are academic and management. *Peoplesoft* allows students to organize themselves, plan individual learning in a flexible way, and shape the roadmap and also desiderata the cost of the semester. Based on *Peoplesoft* database, the faculties, training department and accounting department can easily elaborate and plan the necessary resources to open classes in accordance with the needs of students. Compared with the old solution, *Peoplesoft* helps to eliminate prodigal exertion, free time when the classes do not open and cancel the courses without registration students and the course registration of students also happening faster and simpler than, because of everything has been planned in advance. For these objects, such as training department, faculties, program chair, all essential data are available on the system for query, statistics or drill down to see details of a specific case certain. Preceptors can muster, manage, control, evaluate and review students online via E-learning system is integrated into the ERP system. Next, in this solution, the interaction between teachers and students is very superior. Although the activity for both purposes are improved from course registration, student management to training management, administration, a measurable evaluation of costs and benefits has not been implemented clearly to look more consistent. Finally, the *Cloud Campus*, accordingly, the hourly rate of the traditional classes hourly is roughly 40%, the online classes hourly is about 60%. *Cloud Campus* provides the animation, video, audio, digital libraries, in particular, the Lab is directly put into the cloud that helps learners or teachers can easily interact with IS via mobile devices, such as laptop, tablet, and smartphone users can connect to *Cloud Campus* and use the services of online education anytime and anywhere. In addition, training and learning in *Cloud Campus* would be increasing the interactivity between the trainers and the learners in the cloud. Specifically, lecturers can closely monitor the learning processes of students, and the students can evaluate their knowledge and levels, thereby proactively design pathways of appropriate learning for their capacity.

In addition, with the developed infrastructure by themselves, the difficulty of technical issues and bandwidth issues are faced a regular basis, which reduce the gap between the effective and the cost of the system. For example, at the time of course registration, the network congestion frequently occurs, though students were more convenient for course registration. This is also the common challenge of traditional ERP solutions. With cloud-based ERP, these problems will be supported from providers (cloud and

ERP), Hoa Sen only focus on management and academic operations. Hence, this is a solution that the university should also deliberate to vanquish the current difficulties.

Cloud-Based ERP Solution for Education in Vietnam

Based on the theoretical basis, the relevant models were implemented in the educational institutions in some countries in the world, combining with the actual conditions of the IT infrastructure in Vietnam. Therefore, a solution of cloud-based ERP for education in Vietnam (E-EVN) has been proposed as Figure 5. Accordingly, E-EVN is designed for E-learning, E-researching, and functional management of the educational institution. E-EVN model has 3 main components: (1) *Clouds*: Private cloud and public cloud, which are based on the standard configuration of the cloud computing. (2) *ERP application modules*: Academic function, administration, human resource, and finance will be built and developed based on the technological platform of SaaS. (3) *End users*: Students, lecturers, researchers, and managers. In which, faculty can install exercise applications on the private cloud, and students can access and practice the exercises. Besides that, researchers can manipulate on the private cloud and external cloud for their feasibility studies. In addition, educational administrators can log in to different service modules as authorized by each respective roles and specific management functions in the universities or educational institutions.

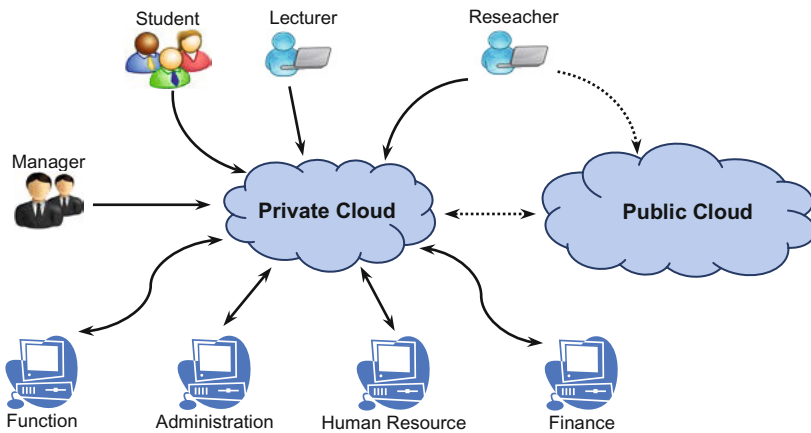


Fig. 5. The Model of Cloud-Based ERP for Education

E-EVN is an integrated model based on ERP applications and cloud computing services. E-EVN solutions will help educational institutions in Vietnam reduce deployment cost to a minimum, no need to invest more on IT infrastructure and software license. Hence, the E-EVN model can be applied widely in universities or educational institutes of large scale, especially the educational institutions of engineering where the experts and researchers specialize in ERP and cloud computing. For SMEs of education, cloud-based ERP solution can also be executed through the advice and support of the university's leading engineering, ERP application providers, and cloud

computing service vendors. Moreover, the adoption of the users is also a critical factor to evaluate the success of cloud-based ERP solutions. In the educational environment, most people use IS (e.g., Pupils/students, teachers, researchers) having the best skills and knowledge as well as approach with the trends of modern educational technology. Thus, the acceptance and use of cloud-based ERP solutions will intervene more easily and fairly.

4 Conclusions

The term of cloud computing is a new synopsis of ITC in the world. The construction of infrastructure for cloud computing application development is realized with multiple methods and solutions, which are the reasons why, cloud computing is becoming increasingly popular. Besides that, ERP is also a new research field of modern IS, as many companies provide the total solutions. These systems contain important data and are integrated into the ERP system to support the operating procedures of the organizations or enterprises. On the other hand, cloud-based ERP brings many benefits in reduced costs, scalability, unfettered access, increased innovation, less staff, and mobility. Besides these, there are many challenges relating to IT infrastructure, ERP applications, cloud computing services and information security. Deploying traditional ERP systems resist many challenges and often take a lot of cost to deploy, from the investment in IT infrastructure to the cost in the software license. However, with the emergence of cloud-based ERP solutions, SaaS has lower cost than traditional ERP. The implementation of cloud-based ERP for modern education has also been synthesized and developed in several stages and different IS. In addition, in this work, the case study of Hoa Sen is shown as an example of cloud-based ERP in Vietnam. Finally, E-EVN solution is proposed as a typical model for the educational institutions in Vietnam. This model is not only minimizing the cost, but also optimizing the applications for the universities or educational institutions, and appreciating the teaching, learning, research, and management in modern education.

Acknowledgment. The authors would like to say thank to 5 blind reviewers for their comments on this paper. We also acknowledge the helpful comments of Thi H. Cao - PhD, vice rector of Saigon Technology University for this study.

References

1. Antero, M., Riis, P.H.: Strategic Management of Network Resources: A case study of an ERP Ecosystem. *International Journal of Enterprise Information Systems* 7(2), 18–33 (2011)
2. Al-Johani, A.A., Youssef, A.E.: A Framework for ERP Systems in SME based on Cloud Computing Technology. *International Journal on Cloud Computing: Services & Architecture* 3(3), 1–14 (2013)
3. Barnatt, C.: *A Brief Guide to Cloud Computing: An Essential Introduction to the Next Revolution in Computing*. Robinson (2010)

4. Bilbao-Osorio, B., Dutta, S., Lanvin, B.: The Global Information Technology Report 2013. Growth and Jobs in a Hyperconnected World (2013)
5. Davenport, T.H.: Putting the Enterprise into the Enterprise System. *Harvard Business Review* 76(4) (1998)
6. Deshmukh, S.: Implementing Cloud ERP systems in Higher Educational Institutes and Universities. *Indian Journal of Research* 3(2), 199–201 (2014)
7. Erpcloudnews: Top 5 Cloud ERP Vendors (2013), <http://erpcloudnews.com>
8. Franc, I., Stankovic, I., Branovic, I., Popovic, R.: Ontology Based Model of Digital Forensic Virtual Lab and Curriculum Design. *International Journal of Engineering Education* 30(4), 964–976 (2014)
9. Houdeshell, R.: Analysis of Cloud ERP Applications, *ERP Cloud News* (2012), <http://erpcloudnews.com>
10. Garcia, J.M., Soriano, E., Garcia, I., Rubio, H.: Implementation of Service-Learning Projects in Engineering Colleges. *International Journal of Engineering Education* 29(5), 1119–1125 (2013)
11. Gerhardtter, A., Ortner, W.: Flexibility and Improved Resource Utilization Through Cloud based ERP Systems: Critical Success Factors of SaaS Solutions in SME. In: *Innovation and Future of Enterprise Information Systems*, pp. 171–182. Springer, Heidelberg (2013)
12. Goel, M.S., Kiran, D.R., Garg, D.D.: Impact of Cloud Computing on ERP Implementations in Higher Education. *International Journal of Advanced Computer Science & Applications* 2(6), 146–148 (2011)
13. Iqbal, U., Uppstrom, E., Juell-Skielse, G.: Cloud ERP Implementation Challenges: A Study based on ERP Life Cycle Model. *Advances in Enterprise Information Systems II*, 389–345 (2012)
14. Iyer, B., Henderson, J.C.: Preparing for the Future: Understanding the seven Capabilities Cloud Computing. *MIS Quarterly Executive* 9(2), 117–131 (2010)
15. Johnson, D.: Different Types of Cloud ERP, *ERP Cloud News* (2010), <http://erpcloudnews.com>
16. Katz, R., Goldstein, P., Yanosky, R., Rushlo, B.: Cloud Computing in Higher Education. *Educause* 10 (2010)
17. Keshwani, B., Sharma, R.: Study & Analysis of Cloud ERP Solutions. *International Journal of Engineering Research and Applications (IJERA)* 3(3), 1130–1136 (2013)
18. Lancon, F.: ERP in Cloud Computing and Information Security Issues (2013), <http://tuvanphanmem.vn>
19. Laudon, K., Laudon, J.: *Management Information Systems: International Edition*, 11/E (2010)
20. Lenart, A.: ERP in the Cloud – Benefits and Challenges. In: Wrycza, S. (ed.) *SIGSAND/PLAIS 2011. LNBIP*, vol. 93, pp. 39–50. Springer, Heidelberg (2011)
21. Markus, M.L., Tanis, C.: The Enterprise Systems Experience - from Adoption to Success. In: *Framing the Domains of IT Research: Glimpsing the Future through the Past*, pp. 173–207 (2000)
22. Marquez, C.Y., Perez, M.A., Yanez, I.L., Nieto, O.C.: Emerging Computational Tools: Impact on Engineering Education and Computer Science Learning. *International Journal of Engineering Education* 30(3), 533–542 (2014)
23. Mell, P., Grance, T.: The NIST definition of Cloud Computing. *NIST Special Publication 800(145)*, 7 (2011)
24. Ministry of Information and Communications: *Information Technology and Communications*. White Paper (2013)

25. Nguyen, T.D., Nguyen, D.T., Cao, T.H.: Acceptance and Use of Information System: E-learning based on Cloud Computing in Vietnam. In: Linawati, Mahendra, M.S., Neuhold, E.J., Tjoa, A.M., You, I. (eds.) *ICT-EurAsia 2014*. LNCS, vol. 8407, pp. 139–149. Springer, Heidelberg (2014)
26. Nguyen, T.D., Nguyen, T.M., Pham, Q.T., Misra, S.: Acceptance and Use of E-learning based on Cloud Computing: The Role of Consumer Innovativeness. In: Murgante, B., et al. (eds.) *ICCSA 2014, Part V*. LNCS, vol. 8583, pp. 159–174. Springer, Heidelberg (2014)
27. Pang, C.: *Dataquest Insight: ERP Software Market Share Analysis, Worldwide (2007, 2008)*, <http://www.gartner.com>
28. Panorama Consulting: *ERP report - 2014*, Panorama Consulting Solutions Research Report (2014)
29. Raihana, G.F.H.: Cloud ERP – A Solution Model. *International Journal of Computer Science and Information Technology & Security* 2(1), 76–79 (2012)
30. SAP: *Industries & Solutions (2014)*, <http://www.sap.com>
31. Sharma, A.K., Ganpati, A.: Cloud Computing: An Economic Solution to Higher Education. *International Journal of Application or Innovation in Engineering & Management* 2(3), 200–206 (2013)
32. Schubert, P., Adisa, F.: Cloud Computing for Standard ERP Systems: Reference Framework and Research Agenda. *Fachbereich Informatik* 14, 36 (2011)
33. Shukla, S., Agarwal, S., Shukla, A.: Trends in Cloud-ERP for SMB's: A Review. *International Journal of New Innovations in Engineering and Technology* 1(1), 7–11 (2012)
34. Staehr, L.: Understanding the role of managerial agency in achieving business benefits from ERP Systems. *Information Systems Journal* 20(3), 213–238 (2010)
35. *Strategic: Moving to a Cloud ERP (2014)*, <http://www.strategic.com>
36. *Sun Systems: Cloud Computing Guide*. Sun Microsystems (2009)
37. Singh, G., Singh, H., Sodhi, N.K.: Cloud Computing-future Solution for Educational Systems. *International Journal of Enterprise Computing and Business Systems* 2(1), 1–17 (2012)
38. Truong, H.L., Pham, T.V., Thoai, N., Dustdar, S.: Cloud Computing for Education and Research in Developing Countries. In: Chao, L. (ed.) *Cloud Computing for Teaching and Learning: Strategies for Design and Implementation*, pp. 64–80 (2012)
39. Turner, P.: *The IMA Survey results are in - What the Cloud means to Finance (2010)*, <http://www.netsuiteblogs.com>
40. Velte, T., Velte, A., Elsenpeter, R.: *Cloud Computing, A Practical Approach*. McGraw-Hill, New York (2009)
41. Yang, Z.: Study on an Interoperable Cloud framework for e-Education. In: *2011 International Conference on E-Business and E-Government*, pp. 1–4. IEEE (2011)
42. Zhang, X., Ma, H., Wu, Y., de Pablos, P.O., Wang, W.: Applying Cloud Computing Technologies to Upgrade the Resource Configuration of Laboratory Course: The Case of Quality Engineering Education Platform. *International Journal of Engineering Education* 30(3), 596–602 (2014)

Heuristics for Energy-Aware VM Allocation in HPC Clouds

Nguyen Quang-Hung¹, Duy-Khanh Le², Nam Thoai¹,
and Nguyen Thanh Son¹

¹Faculty of Computer Science and Engineering,

HCMC University of Technology, VNUHCM

268 Ly Thuong Kiet Street, Ho Chi Minh City, Vietnam

{hungnq2,nam,sonsys}@cse.hcmut.edu.vn

²Department of Computer Science, National University of Singapore

leduykha@comp.nus.edu.sg

Abstract. High performance computing (HPC) clouds have become more popular for users to run their HPC applications on cloud infrastructures. Reduction in energy consumption (kWh) for these cloud systems is of high priority for any cloud provider. In this paper, we first study the energy-aware allocation of virtual machines (VMs) in HPC cloud systems along two dimensions: multi-dimensional resources and interval times of virtual machines. On the one hand, we present an example showing that using bin-packing heuristics (e.g. Best-Fit Decreasing) to minimize the number of physical servers could not lead to a minimum of total energy consumption. On the other hand, we find out that minimizing total energy consumption is equivalent to minimizing the sum of total completion time of all physical machines. Based on this finding, we propose the MinDFT-ST and MinDFT-FT algorithms to place the VMs onto the physical servers in such a way that minimizes the total completion times of all physical servers. Our simulation results show that MinDFT-ST and MinDFT-FT could reduce the total energy consumption by 22.4% and respectively 16.0% compared with state-of-the-art power-aware heuristics (such as power-aware best-fit decreasing) and vector bin-packing norm-based greedy algorithms (such as VBP-Norm-L1, VBP-Norm-L2, VBP-Norm-L30).

1 Introduction

High Performance Computing (HPC) clouds have been popularly adopted [12,14,16,19] and are provided by industrial companies such as Amazon Web Service Cloud [2]. A HPC cloud is a cloud system that provides users with computing resources in terms of virtual machines (VMs) to run their HPC applications [10,14]. These cloud systems are often built from virtualized data centers [18,4]. Powering these cloud systems is very costly and is increasing with the increasing scale of these systems. Therefore, advanced scheduling techniques for reducing energy consumption of these cloud systems are highly concerned

for any cloud providers. Energy-aware scheduling of VMs in HPC cloud is still challenging [10,12,17,20,21].

There are several works that have been proposed to address the problem of energy-efficient scheduling of VMs in cloud data centers. A large body of works [4,5,15] presents methodologies for consolidating virtual machines in cloud data centers by using bin-packing heuristics (such as FFD, BFD). They attempt to minimize the number of running physical machines and to turn off as many idle physical machines as possible. In this paper, we find out that using a minimum number of physical machines is not necessarily a good solution to minimize total energy consumption. For example, consider a d -dimensional resource allocation where each user requests a set of virtual machines (VMs). Each VM requires multiple resources (such as CPU, memory, and IO) and a fixed quantity of each resource at a certain time interval. Under this scenario, using a minimum of physical machines may not be a good solution. For example, given five virtual machines (VMs) with their resource demands described in Table 1. In the example, a bin-packing-based algorithm could result in a schedule S_1 in which two physical servers are used: one for allocating VM1, VM3, and VM4, and another one for allocating VM2 and VM5. The resulted total completion time is 20 hours. However, in another schedule S_2 where VMs are placed on three physical servers: VM1 and VM2 on the first physical server, VM3 and VM4 on the second physical server, and VM5 on the third physical server, then the total completion time of the five VMs is only 12 hours. In a homogeneous environment where all physical servers are identical and the power consumption of each physical server is linear to its CPU utilization, a schedule with longer working time (such as S_1) will consume more energy consumption than another schedule with shorter working time (such as S_2).

Table 1. Example showing that using a minimum number of physical servers is not optimal. (*: normalized demand resources to physical server’s capacity resources)

Virtual machine	CPU*	RAM*	Network*	Start-time	Duration (hour)
VM1	0.5	0	0	1	10
VM2	0.5	0	0	1	10
VM3	0.2	0.5	0	1	1
VM4	0.2	0.5	0	1	1
VM5	0.1	0.1	0.1	1	1

The resource allocation of VMs with multiple resources is a hard problem. Each VM requires multiple resources such as CPU, memory, I/O to execute its applications. The resource allocation problem can be seen as a d -dimensional ($d = 1, 2, 3, \dots$) Vector Bin Packing problem (VBP_d) [15], in which each physical server with multiple resources is considered as a d -dimensional bin, and each virtual machine is a d -dimensional item with various sizes of the requested resources (e.g. CPU, memory). The VBP_d problem is known as NP-hard for $\forall d \geq 1$ and APX-hard for $\forall d \geq 2$ [15]. Therefore, the resource allocation of VMs with multiple resources is also NP-hard. In this paper, we propose

MinDFT-ST and MinDFT-FT heuristics for the power-aware VM allocation problem with d-dimensional resources and interval times. Using numerical simulations, we compare the MinDFT-ST and MinDFT-FT algorithms with a popular power-aware best-fit decreasing (PABFD) [5] and vector bin-packing norm-based greedy (VBP-Norm-L1/L2/L30) [15]. Simulation results showed that our proposed MinDFT-ST, MinDFT-FT can reduce the total energy consumption of the physical servers more than the PABFD and VBP-Norm-L1/L2/L30 algorithms on a real workload (LPC log) on the Parallel Workload Archive [1].

The rest of this paper is structured as follows. Section 2 discusses related works. Section 3 describes the energy-aware virtual machine allocation with multiple requested resources, fixed starting and duration time. We also formulate the objective of scheduling and present our scheduling algorithms in Section 3. Section 4 discusses our performance evaluation using simulations. Section 5 concludes this paper and introduces future works.

2 Related Works

Cloud computing has been developed as a utility computing model [6] and is driven by economies of scale. Sotomayor et al. [19,18] proposed a lease-based model and implemented First-Come-First-Serve (FCFS) and back-filling algorithms [9] to schedule best effort, immediate and advanced reservation VM-based leases. The FCFS and back-filling algorithms consider only one performance metric such as waiting time and slowdown, without mentioning energy efficiency. To maximize performance, these scheduling algorithms tend to choose free load servers (i.e. those with the highest-ranking scores) when allocating new VMs.

There are several works that have been proposed to address the problem of energy-efficient scheduling of VMs in cloud data centers. Some works [3,11] proposed scheduling algorithms to flexibly change processor speed in such a way that meets user requirements and reduces power consumption of processors when executing user applications. Some other works proposed algorithms that consolidate VMs onto a small set of physical servers in virtualized datacenters [5,4,20] such that power consumption of physical servers is minimized.

Albers et al. [3] reviewed some energy-efficient algorithms which are used to minimize flow time by changing processor speed according to job size. Laszewski et al. [11] proposed scheduling heuristics and presented application experience for reducing power consumption of parallel tasks in a cluster with the Dynamic Voltage Frequency Scaling (DVFS) technique. We did not use the DVFS technique to reduce energy consumption on data centers. We propose software-based VM allocation algorithms to independent to vendor-locked hardware.

Many works have considered the VM placement problem as a bin-packing problem. They use bin-packing heuristics (e.g. Best-Fit Decreasing (BFD)) to place virtual machines (VMs) onto a minimum number of physical servers to minimize energy consumption [5,4]. Microsoft research group [15] has studied first-fit decreasing (FFD) based heuristics for vector bin-packing to minimize

number of physical servers in the VM allocation problem. Beloglazov et al. [5,4] have proposed VM allocation problem as bin-packing problem and presented a power-aware modified best-fit decreasing (denoted as PABFD) heuristic. PABFD sorts all VMs in a decreasing order of CPU utilization and tends to allocate a VM to an active physical server that would take the minimum increase of power consumption. However, choosing a host with a minimized increasing power consumption does not necessarily imply minimizing total energy consumption in VM allocation problems where all physical servers are identical and the power consumption of a physical server is linear to its CPU utilization. The PABFD will assign new VM to any physical server. The PABFD also does not consider the starting time and finishing time of these VMs. Therefore, it is unsuitable for the power-aware VM allocation considered in this paper, i.e. the PABFD may not result in a minimized total energy consumption for VM placement problem with certain interval time while still fulfilling the quality-of-service. In contrast, our proposed MinDFT-ST and MinDFT-FT consider the case where each user VM has a certain interval time (i.e. started at a starting time in non-preemptive duration).

Some other research [10,12,20] considers about HPC applications/jobs in HPC cloud. Garg et al. [10] proposed a meta-scheduling problem to distribute HPC applications to cloud systems with distributed N data centers. The objective of scheduling is minimizing CO_2 emission and maximizing the revenue of cloud providers. Le et al. [12] distribute VMs across distributed cloud virtualized data centers whose electricity prices are different in order to reduce the total electricity cost. Takouna et. al., [20] presented power-aware multi-core scheduling and their VM allocation algorithm selects a host which has the minimum increasing power consumption to assign a new VM. The VM allocation algorithm, however, is similar to the PABFD's [5] except that they concern memory usage in a period of estimated runtime for estimating the host's energy. The work also presented a method to select optimal operating frequency for a (DVFS-enabled) host and configure the number of virtual cores for VMs. Our proposed MinDFT-ST and MinDFT-FT, which are VM allocation algorithms, differ from the these previous works. Our algorithms use the VM's starting time and finished time to minimize the total working time on physical servers, and consequently minimize the total energy consumption in all physical servers.

Mämmelä et. al., [13] presented energy-aware First-In, First-Out (E-FIFO) and energy-aware Backfilling First-Fit and Best-Fit (E-BFF, E-BBF) scheduling algorithms for non-virtualized high performance computing system. The E-FIFO puts new job at the end of job-queue (and dequeue last), finds out an available host for the first job and turns off idle hosts. The E-BFF and E-BBF are similar to E-FIFO, but the E-BFF and E-BBF will attempt to assign jobs to all idle hosts. Unlike our proposed MinDFT-ST and MinDFT-FT, the Mämmelä's work do not consider power-aware VM allocation.

3 Problem Description

3.1 Notations

We use the following notations in this paper:

vm_i : The i^{th} virtual machine to be scheduled.

M_j : The j^{th} physical server.

S : A feasible schedule.

P_j^{idle} : Idle power consumption of the M_j .

P_j^{max} : Maximum power consumption of the M_j .

$P_j(t)$: Power consumption of a single physical server (M_j) at a time point t .

$U_j(t)$: CPU utilization of the M_j at a time point t .

ts_i : Fixed starting time of vm_i .

dur_i : Duration time of vm_i .

T : The maximum time of the scheduling problem.

$r_j(t)$: Set of indexes of all VMs that are assigned to the physical machine M_j at time t .

T_j : The working time of a physical server.

e_i : The energy consumption for running the vm_i in the physical machine that the vm_i is allocated.

3.2 Power Consumption Model

In this paper, we use the following energy consumption model proposed in [8] for a physical server M_j . The power consumption of each physical server denoted as $P_j(\cdot)$ formulates as:

$$P_j(t) = P_j^{idle} + (P_j^{max} - P_j^{idle}) \times U_j(t) \quad (1)$$

The CPU utilization, denoted as $U_j(t)$, of the physical server at time t formulates as:

$$U_j(t) = \sum_{c=1}^{PE_j} \sum_{i \in r_j(t)} \frac{mips_{i,c}}{MIPS_{j,c}} \quad (2)$$

The energy consumption of the server at time t formulates as:

$$E_j = \int_{t_1}^{t_2} P_j(U_j(t)) dt \quad (3)$$

where:

$U_j(t)$ is CPU utilization of the server M_j at time t and $0 \leq U_j(t) \leq 1$.

PE_j : Number of processing elements (i.e. cores) of the physical server M_j .

$mips_{i,c}$: Allocated MIPS of the c^{th} processing element to the vm_i by M_j .

$MIPS_{j,c}$: Total MIPS of the c^{th} processing element on the M_j .

3.3 Problem Formulation

Given a set of virtual machines vm_i ($i = 1, 2, \dots, n$) to be scheduled on a set of physical servers M_j ($j = 1, 2, \dots, m$). Each VM is represented as a d -dimensional vector of demand resources, i.e. $vm_i = (x_{i,1}, x_{i,2}, \dots, x_{i,d})$. Similarly, each physical machine is denoted as a d -dimensional vector of capacity resources, i.e. $M_j = (y_{j,1}, y_{j,2}, \dots, y_{j,d})$. We consider types of resources such as processing element (PE), computing power (Million instruction per seconds -MIPS), physical memory (RAM), network bandwidth (BW), and storage. Each vm_i is started at a fixed starting time (ts_i) and is non-preemptive during its duration time (dur_i).

We assume that the power consumption model is linear to CPU utilization. Even if all physical servers are identical and all VMs are identical too, the scheduling is still NP-hard with $d \geq 1$ [15]. In the problem, when all physical servers are identical and their power consumption models are linear to their CPU utilization as can be seen in the two equations (1) and (3). The energy consumption of a physical server in a unit of time is denoted as E_0 and is the same for all physical servers since the servers are identical. The energy consumption of each VM, denoted as e_i , is independent with mapping to any physical server. A feasible schedule S indicates a successful mapping of all VMs to physical servers, i.e. $\forall i \in \{1, 2, \dots, n\}, \exists j \in \{1, 2, \dots, m\} : allocated(vm_i, M_j)$ where $allocated(vm_i, M_j)$ holds when vm_i is allocated on the physical server M_j . The objective is to find out a feasible schedule S that minimizes the total energy consumption, denoted as $Energy(S)$ in the equation (4) as following with $i \in \{1, 2, \dots, n\}, j \in \{1, 2, \dots, m\}, t \in [0; T]$:

$$\text{Minimize } (E_0 \times \sum_{j=1}^m T_j + \sum_{i=1}^n e_i) \quad (4)$$

where the working time of a physical server, denoted as T_j , is defined as union of interval time of all VMs that are allocated to a physical machine j^{th} at time t .

$$T_j = \bigcup_{vm_i \in r_{j,t}} [ts_i; ts_i + dur_i] \quad (5)$$

The union of two time intervals $[a;b]$ and $[c;d]$ is defined as: $[a;b] \cup [c;d] = \{x \in \mathbb{R} \mid x \in [a;b] \text{ or } x \in [c;d]\}$

The scheduling problem has the following hard constraints:

- Constraint 1: Each VM is run by a physical server (host) at any time.
- Constraint 2: VMs do not request any resource larger than total capacity resource of their hosts.
- Constraint 3: Let $r_j(t)$ be the set of VMs that are allocated onto a host M_j . The sum of total demand resources of these allocated VMs is less than or equal to total capacity of the resources of the M_j .

$$\forall c = 1, \dots, d : \sum_{vm_i \in r_j(t)} x_{i,c} \leq y_{j,c} \quad (6)$$

where:

- $x_{i,c}$ is resource of type c (e.g. CPU core, computing power, memory, etc.) requested by the vm_i ($i=1,2,\dots,n$).

- $y_{j,c}$ is capacity of type c (e.g. CPU core, computing power, memory, etc.) of the physical machine M_j ($j = 1, 2, \dots, m$).

Our key observation is that, according to the objective function described in (4), E_0 is constant while e_i is independent with any mapping (i.e. any schedule). Therefore, minimizing total energy consumption is equivalent to minimizing the sum of total completion time of all physical machines ($\sum_{j=1}^m T_j$).

$$\text{Minimize } (E_0 \times \sum_{j=1}^m T_j + \sum_{i=1}^n e_i) \sim \text{Minimize } (\sum_{j=1}^m T_j) \quad (7)$$

Based on the above observation, we propose our energy-aware algorithms as presented in the next section.

3.4 Scheduling Algorithm

Algorithm 1. MinDFT-ST and MinDFT-FT: Heuristics for energy-aware VM allocation in HPC Clouds

```

1: function MinDFT-ST
2:   Input: vmList - a list of virtual machines to be scheduled
3:   Input: hostList - a list of physical servers
4:   Output: mapping (a feasible schedule) or null
5:   vmList = sortVmListByStartTime( vmList ) ▷ 1
6:   return MinDFT( vmList, hostList )
7: end function
8: function MinDFT-FT
9:   Input: vmList - a list of virtual machines to be scheduled
10:  Input: hostList - a list of physical servers
11:  Output: mapping (a feasible schedule) or null
12:  vmList = sortVmListByFinishTime( vmList ) ▷ 2
13:  return MinDFT( vmList, hostList )
14: end function

```

In this section, we present our energy-aware scheduling algorithms. We present two algorithms, MinDFT-ST and MinDFT-FT, which are best-fit decreasing heuristics to allocate a new VM to a physical server in such a way that minimizes the total completion times of all physical servers. Algorithm 1 shows the pseudocode for our proposed MinDFT-ST and MinDFT-FT, both are based on the core algorithm MinDFT presented in Algorithm 2.

The MinDFT uses an array T whose element is the total completion time of each physical server. Initially, T is initialized to zeros (lines 5-7). In the i^{th} iteration of the loop at line 8, the i^{th} VM (denoted as vm) in the set of n VMs is selected (line 9). At the beginning, vm is not allocated onto any host or physical

Algorithm 2. MinDFT: Core Algorithm for energy-aware VM allocation in HPC Clouds

```

1: function MINDFT
2:   Input: vmList - a sorted list of virtual machines to be scheduled
3:   Input: hostList - a list of physical servers
4:   Output: mapping (a feasible schedule) or null
5:   for  $j = 1$  to  $m$  do
6:      $T[j] = 0$ 
7:   end for
8:   for  $i = 1$  to  $n$  do
9:      $vm = vmList.get(i)$ 
10:     $allocatedHost = null$ 
11:     $T1 = sumTotalHostCompletionTime( T )$  ▷ Calculating the total
    completion time of all active physical servers:
12:     $minDiffTime = +\infty$ 
13:    for  $j = 1$  to  $m$  do
14:       $host = hostList.get(j)$ 
15:      if  $host.checkAvailableResource( vm )$  then
16:        ▷ host's available resources has enough the vm's requested resources
17:        if  $host.isOverUtilizedAfterAllocationVm( vm )$  then
18:          ▷ host is over utilized after allocation the vm
19:          continue
20:        end if
21:         $preTime = T[ host.id ]$ 
22:         $host.vmCreate(vm)$  ▷ begin test
23:         $T[ host.id ] = host.estimateHostTotalCompletionTime( vm )$ 
24:         $T2 = sumTotalHostCompletionTime( T )$ 
25:         $host.vmDestroy(vm)$  ▷ end test
26:         $diffTime = T2 - T1$ 
27:        if  $(minDiffTime > diffTime)$  then
28:           $minDiffTime = diffTime$ 
29:           $allocatedHost = host$ 
30:        else
31:           $T[ host.id ] = preTime$ 
32:        end if ▷ Next iterate over hostList and choose the host that
    minimize the different time
33:      end if
34:    end for ▷ End for host list
35:    if  $(allocatedHost \neq null)$  then
36:      allocate the vm to the host
37:      add the pair of vm (key) and host to the mapping
38:    end if
39:  end for ▷ end for vm list
40:  return mapping
41: end function
42:  $sumTotalHostCompletionTime(T[]) = \sum_{j=1}^m T_j$  ▷  $T[1...m]$ : Array of total
    completion times of  $m$  physical servers

```

server (line 10). The two variables $T1$ and $T2$ are used to compute the sums of total completion times of physical servers before (line 11) and respectively after (line 24) the VM scheduler assigns a new VM onto a physical server. The j^{th} iteration at line 13 is used to find out the best-fit physical server to assign the i^{th} VM (vm) such that the sum of total completion times increases minimally. The algorithm chooses any host $host$ (line 14) and checks whether the host has enough resource to provision for the requested resources of vm and is not over-utilized after allocating the vm onto it (lines 15-17). The available resources (such as CPU cores, computing power (in MIPS), memory (i.e. RAM), network bandwidth, and free size of in storage systems) of the selected host $host$ should meet the requested resources of vm . Otherwise, MinDFT fails to allocate vm onto $host$. If the selected $host$ has enough resources and is not over-utilized after the allocating the allocating vm , the algorithm stores the current completion time of the host (line 21) and attempts to allocate vm onto the selected $host$ (line 22). The algorithm then calculates the new total completion time of $host$ after allocating vm onto it (line 24). The increase $diffTime$ between total completion times $T2$ and $T1$ is computed (line 26). The MinDFT algorithm will assign a new VM vm to any physical server $host$ in such a way that minimizes the difference between $T1$ and $T2$ (lines 27-32). This, consequently, minimizes the sum of total of completion times of physical servers as described in the formula (7). The MinDFT is a best-fit-based heuristic.

The MinDFT-FT differs from the MinDFT-ST in the sorting order of the list of VMs. While MinDFT-ST sorts the list in the ascending order of the VMs' start time at line 5, MinDFT-ST sorts the list of VMs in the ascending order of the VMs' finish time at line 12. The MinDFT-ST and MinDFT-FT solve the scheduling problem in time complexity of $\mathcal{O}(nm)$ where n is the number of VMs to be scheduled and m is the number of physical servers.

4 Experimental Study

In this section, we study the following VM allocation algorithms:

- PABFD, a power-aware and best-fit decreasing heuristic presented in [5]. The PABFD sorts the list of VM_i ($i=1, 2, \dots, n$) by their total requested CPU utilization, and assigns new VM to any host that has a minimum increase in power consumption.
- VBP-Norm-LX, a family of vector packing heuristics that is presented as Norm-based Greedy with degree $X=1, 2, 30$ [15]. Weights of these Norm-based Greedy heuristics use $FFDAvgSum$ which are $\exp(x)$, which is the value of the exponential function at the point x , where x is average of sum of demand resources (e.g. CPU, memory, storage, network bandwidth, etc.). VBP-Norm-LX assigns new VM to any host that has minimum of these norm values.
- EPOBF-ST and EPOBF-FT, presented in [17]. The EPOBF-ST and EPOBF-FT algorithms sorts the list of VM_i ($i=1, 2, \dots, n$) by their starting time (ts_i) and respectively by their finished time ($ts_i + dur_i$).

- MinDFT-ST and MinDFT-FT, our proposed algorithms discussed in Section 3.4. The MinDFT-ST sorts the list of VMs by VM’s starting time. The MinDFT-FT sorts the list of VMs by VM’s finished time ($ts_i + dur_i$).

The algorithms are evaluated using CloudSim [7], a popular simulation tool for modeling cloud systems and evaluates VM allocation algorithms. We used the LPC log-trace from a real HPC system in the Parallel Workloads Archive [1]. Each row of the LPC log-trace simulates a HPC job that requests a VM. In our simulation, there are four type of VMs with configuration as the Table 2 and all physical servers are homogeneous with host configuration of 2660 MIPS, 4 cores, 8192 MB of physical memory, 1 Gb/s of network bandwidth, 1 TBytes of storage. Power model of physical servers is shown in Table 3.

Table 2. Four VM type in simulations

Type	Cores	MIPS	Memory (MB)	Network (Kb/s)	Storage (MB)
VM1	1	2500	871	100000	5000
VM2	1	2000	3840	100000	5000
VM3	1	1500	1536	100000	5000
VM4	1	1000	613	100000	5000

Table 3. Host power consumption model with CPU utilization of an IBM server x3250 (1 x [Xeon X3470 2933 MHz, 4 cores], 8GB).

CPU Utilization (%)	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Host power (Watts)	41.6	46.7	52.3	57.9	65.4	73.0	80.7	89.5	99.6	105.0	113.0

Table 4. Simulation results with 10K jobs and 5K hosts.

Algorithm	#Hosts	#VMs	Energy (kWh)	Energy Saving (%)
PABFD (baseline) [5]	5000	10000	5,677.294	0.0%
VBP-Norm-L1 [15]	5000	10000	5,677.294	0.0%
VBP-Norm-L2 [15]	5000	10000	5,677.294	0.0%
VBP-Norm-L30 [15]	5000	10000	5,677.294	0.0%
EPOBF-ST [17]	5000	10000	4,969.455	12.5%
EPOBF-FT [17]	5000	10000	4,766.789	16.0%
MinDFT-ST	5000	10000	4,405.329	22.4%
MinDFT-FT	5000	10000	4,766.789	16.0%

The simulation results are shown in Table 4 and Figure 1. Table 4 shows simulation results of scheduling algorithms solving scheduling problems with 10K jobs and 5K hosts. The Figure 1 shows a bar chart comparing energy consumption

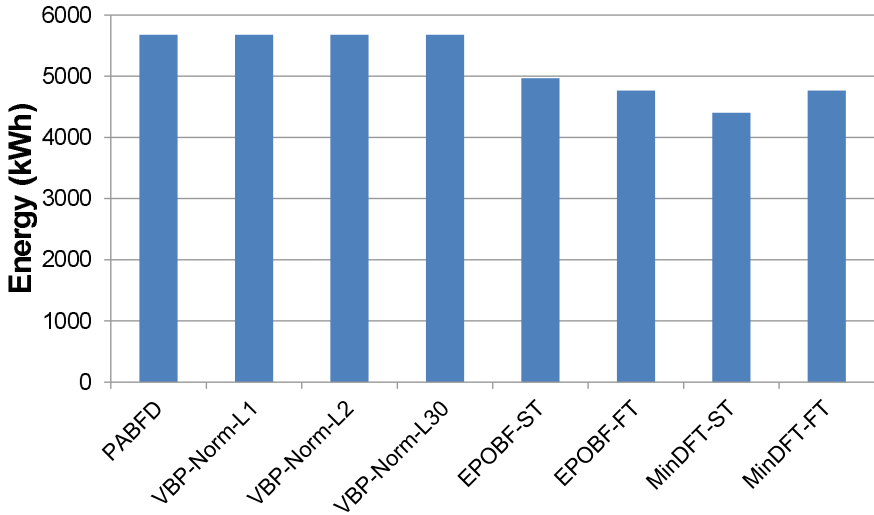


Fig. 1. Energy (kWh)

of VM allocation algorithms. None of the algorithms use VM migration techniques, and all of them satisfy the Service-Level Agreement (e.g. the scheduling algorithm provisions maximum of user VM's requested resources). We use total energy consumption as the performance metric for evaluating these VM allocation algorithms. The energy saving shown in Table 4 is the reduction of total energy consumption of the corresponding algorithm compared with the baseline PABFD [5] algorithm.

We choose PABFD [5] as the baseline algorithm because the PABFD is a famous power-aware best-fit decreasing in the energy-aware scheduling research community. We also compare our proposed VM allocation algorithms with three vector bin-packing algorithms to show the importance of with/without considering VM's starting time and finish time in reducing the total energy consumption of VM placement problem. Table 4 shows that, compared with PABFD [5], our MinDFT-ST and MinDFT-DL algorithms reduce the total energy consumption by 22.4% and respectively 16.0%. MinDFT-ST and MinDFT-DL also reduce the total energy consumption by 22.4% and 16.0% compared with norm-based vector bin-packing algorithms such as VBP-Norm-L1, VBP-Norm-L2, VBP-Norm-L30. The PABFD generates a schedule that uses higher energy consumption than the MinDFT-ST and MinDFT-FT because of following main reasons. First, our hypothesis in this paper that each VM consumes the same energy consumption in any physical server (e_i) and all physical servers are identical. In consequence, the PABFD will choose a random physical server to map a new VM. Instead of that, our proposed MinDFT assigns a new VM to a physical server in such a way that has minimum increase of completion times.

Moreover, sorting the list of VMs by starting and finished times is important because they can reduce total energy consumption in the VM allocation problem

with HPC jobs. In Table 5, the MinDFT-DescU is the MinDFT with the list of VMs sorted in a decreasing order of requested CPU utilizations. The MinDFT-DescU has total energy consumption is 5327.430 KWh; it reduces the energy consumption

Table 5. Sorting the list of VMs by decreasing VM's requested CPU utilization. Energy consumption (KWh).

Algorithm	#Hosts	#Jobs	Energy (kWh)	Saving Energy (%)
PABFD	5000	10000	5677.294	0.0%
VBP-Norm-L1	5000	10000	5677.294	0.0%
VBP-Norm-L2	5000	10000	5677.294	0.0%
VBP-Norm-L30	5000	10000	5677.294	0.0%
MinDFT-DescU	5000	10000	5327.430	6.2%

Table 6. Sorting the list of VM's by their starting time. Energy consumption (KWh).

Algorithm	#Hosts	#Jobs	Energy (kWh)	Energy saving (%)
BFD-ST (base line)	5000	10000	4969.455	0.0%
VBP-Norm-L1-ST (*)	5000	10000	4969.455	0.0%
VBP-Norm-L2-ST (*)	5000	10000	4969.455	0.0%
VBP-Norm-L30-ST(*)	5000	10000	4969.455	0.0%
MinDFT-ST	5000	10000	4405.329	11.4%

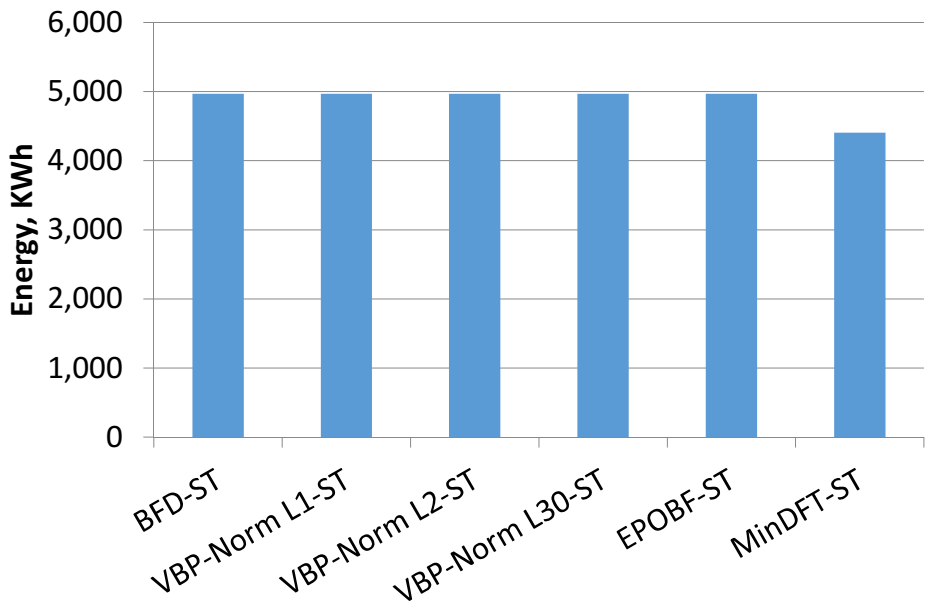


Fig. 2. Energy (kWh)

to about 6.2% compared with the based line algorithms (e.g. PABFD, VBP-Norm-L1/2/30). In addition, we also study on variety of the bin-packing heuristics such as BFD-ST, VBP-Norm-L1-ST, VBP-Norm-L2-ST, VBP-Norm-L30-ST, which are the PABFD, VBP-Norm-L1, VBP-Norm-L2, VBP-Norm-L30 algorithms except that they sort the list of VMs by starting time instead of CPU utilization. Simulation results are shown in Table 6. The results show that the MinDFT-ST is better than the BFD-ST, VBP-Norm-L1-ST, VBP-Norm-L2-ST, VBP-Norm-L30-ST, i.e. the MinDFT-ST reduces 11.4% total energy consumption when compared with the BFD-ST.

5 Conclusion and Future Work

In this paper, we formulated an energy-aware scheduling problem and discussed our key observation in the VM allocation problem: minimizing total energy consumption is equivalent to minimizing the sum of total completion time of all physical machines. Based on the observation, we proposed MinDFT-ST and MinDFT-FT algorithms to solve the scheduling problem in time complexity of $\mathcal{O}(nm)$. Evaluation showed that our proposed MinDFT-ST and MinDFT-DL reduce the total energy consumption by 22.4% and respectively 16.0% compared with well-known PABFD [5] and norm-based vector bin-packing algorithms [15].

In future, we plan to integrate our MinDFT-ST and MinDFT-FT into a cloud resource management software (e.g. OpenStack Nova Scheduler). Additionally, we are interested in cloud systems with heterogeneous physical servers and job requests consisting of multiple VMs. The cloud systems can provide resources to many types of VM-based leases [18] including best-effort, advanced reservation, and immediate leases at the same time. We are also studying Mixed Integer Linear Programming models and meta-heuristics (e.g. Genetic Algorithms, Particle Swarm Optimization (PSO)) of the VM allocation problem.

References

1. The LPC log from the Parallel Workloads Archive, http://www.cs.huji.ac.il/labs/parallel/workload/1_lpc/LPC-EGEE-2004-1.2-cln.swf.gz (retrieved on January 30, 2014)
2. AWS - High Performance Computing - HPC Cloud Computing, <http://aws.amazon.com/hpc/> (retrieved on August 31, 2014)
3. Albers, S.: Energy-efficient algorithms. *Commun. ACM* 53(5), 86–96 (2010)
4. Beloglazov, A., Abawajy, J., Buyya, R.: Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing. *Future Generation Comp. Syst.* 28(5), 755–768 (2012)
5. Beloglazov, A., Buyya, R.: Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in cloud data centers. *Concurrency and Computation: Practice and Experience* 24(13), 1397–1420 (2012)
6. Buyya, R., Yeo, C., Venugopal, S., Broberg, J., Brandic, I.: Cloud computing and emerging it platforms: Vision, hype, and reality for delivering computing as the 5th utility. *Future Generation Comp. Syst.* 25(6), 599–616 (2009)

7. Calheiros, R.N., Ranjan, R., Beloglazov, A., De Rose, C.A.F., Buyya, R.: Cloudsim: A toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms. *Softw. Pract. Exper.* 41(1), 23–50 (2011)
8. Fan, X., Weber, W.D., Barroso, L.: Power provisioning for a warehouse-sized computer. In: *ISCA*, pp. 13–23 (2007)
9. Feitelson, D.G., Rudolph, L., Schwiegelshohn, U.: Parallel job scheduling — A status report. In: Feitelson, D.G., Rudolph, L., Schwiegelshohn, U. (eds.) *JSSPP 2004. LNCS*, vol. 3277, pp. 1–16. Springer, Heidelberg (2005)
10. Garg, S.K., Yeo, C.S., Anandasivam, A., Buyya, R.: Energy-Efficient Scheduling of HPC Applications in Cloud Computing Environments. *CoRR* abs/0909.1146 (2009)
11. von Laszewski, G., Wang, L., Younge, A.J., He, X.: Power-aware scheduling of virtual machines in dvfs-enabled clusters. In: *CLUSTER*, pp. 1–10 (2009)
12. Le, K., Bianchini, R., Zhang, J., Jaluria, Y., Meng, J., Nguyen, T.D.: Reducing electricity cost through virtual machine placement in high performance computing clouds. In: *SC*, p. 22 (2011)
13. Mämmelä, O., Majanen, M., Basmadjian, R., de Meer, H., Giesler, A., Homberg, W.: Energy-aware Job Scheduler for High-performance Computing (2012)
14. Mauch, V., Kunze, M., Hillenbrand, M.: High performance cloud computing. *Future Generation Comp. Syst.* 29(6), 1408–1416 (2013)
15. Panigrahy, R., Talwar, K., Uyeda, L., Wieder, U.: Heuristics for Vector Bin Packing. *Tech. rep.*, Microsoft Research (2011)
16. Pham, T.V., Jamjoom, H., Jordan, K.E., Shae, Z.Y.: A service composition framework for market-oriented high performance computing cloud. In: *HPDC*, pp. 284–287 (2010)
17. Quang-Hung, N., Thoai, N., Son, N.: Epobf: Energy efficient allocation of virtual machines in high performance computing. *J. Sci. Technol. Vietnamese Acad. Sci. Technol.*, Special on International Conference on Advanced Computing and Applications (ACOMP2013) 51(4B), 173–182 (2013)
18. Sotomayor, B.: Provisioning Computational Resources Using Virtual Machines and Leases. Ph.D. thesis, University of Chicago (2010)
19. Sotomayor, B., Keahey, K., Foster, I.T.: Combining batch execution and leasing using virtual machines. In: *HPDC*, pp. 87–96 (2008)
20. Takouna, I., Dawoud, W., Meinel, C.: Energy Efficient Scheduling of HPC-jobs on Virtualize Clusters using Host and VM Dynamic Configuration. *Operating Systems Review* 46(2), 19–27 (2012)
21. Viswanathan, H., Lee, E.K., Rodero, I., Pompili, D., Parashar, M., Gamell, M.: Energy-Aware Application-Centric VM Allocation for HPC Workloads. In: *IPDPS Workshops*, pp. 890–897 (2011)

Information-Flow Analysis of Hibernate Query Language

Agostino Cortesi¹ and Raju Halder²

¹ Università Ca' Foscari Venezia, Italy
cortesi@unive.it

² Indian Institute of Technology Patna, India
halder@iitp.ac.in

Abstract. Hibernate Query Language (HQL) provides a framework for mapping object-oriented domain models to traditional relational databases. In this context, existing information leakage analyses cannot be applied directly, due to the presence and interaction of high-level application variables and SQL database attributes. The paper extends the Abstract Interpretation framework to properly deal with this challenging applicative scenario, by using the symbolic domain of positive propositional formulae to capture variable dependences affecting (directly or indirectly) the propagation of confidential data.

Keywords: Hibernate Query Language, Information Leakage, Static Analysis, Abstract Interpretation.

1 Introduction

Hibernate Query Language (HQL) provides a framework for mapping object-oriented domain models to traditional relational databases [1, 2, 6]. Basically it is an ORM (Object Relational Mapping) which solves object-relational impedance mismatch problems, by replacing direct persistence-related database accesses with high-level object handling functions. Various methods in “*Session*” are used to propagate object’s states from memory to the database (or vice versa). Hibernate will detect any change made to an object in persistent state and synchronizes the state with the database when the unit of work completes. A HQL query is translated by Hibernate into a set of conventional SQL queries during run time which in turn performs actions on the database.

Preserving confidentiality of sensitive information in software systems always remains a thrust area for researchers. Sensitive data may be leaked maliciously or even accidentally through a bug in the program [14]. For example, any health information processing system may release patient’s data, or any on-line transaction system may release customer’s credit card information through covert channels while processing.

The following code fragments depict two different scenarios (explicit/direct flow and implicit/indirect flow) of information leakage:

Explicit Flow	Implicit flow
<code>l := h</code>	<code>if(h==0) l=5; else l=10;</code>

Assuming variables ‘h’ and ‘l’ are private and public respectively, it is clear from the code that confidential data in ‘h’ can be deduced by attackers observing ‘l’ on the output channel.

As traditional security measures (*e.g.* access control, encryption, etc.) do not fit to solve this when sensitive information is released from the source legitimately and it is propagated through the software during computations, various language-based information flow security analysis approaches are proposed [9, 10, 14, 15]. This is formalized by the non-interference principle that says “a variation of confidential data does not cause any variation to public data”. Works in this direction have been starting with the pioneering work of Dennings in the 1970s [5].

Most of the notable works [8–10, 13] which refer to imperative, object-oriented, functional, and structured query languages, can not be applied directly to the case of HQL due to the presence and interaction of high-level HQL variables and database attributes through `Session` methods. Moreover, analyzing object-oriented features of HQL does not meet our objectives neither.

In this paper, we extend the abstract interpretation-based framework in [16] to the case of HQL, focussing on `Session` methods which act as persistent manager. This allows us to perform leakage analysis of sensitive database information when is accessed through high-level HQL code.

The proposed approach is two-folded:

- Defining the concrete and an abstract transition semantics of HQL, by using symbolic domain of positive propositional formulae.
- Analyzing possible information leakage based on the abstract semantics, focussing on variable dependences of database attributes on high-level HQL variables.

The structure of the paper is as follows: Section 2 provides a motivational example. In Section 3, we formalize the concrete and an abstract transition semantics of HQL, by using the symbolic domain of positive propositional formulae. In Section 4, we perform information leakage analysis of programs based on the abstract semantics which captures possible leakage of confidential data. Section 5 concludes the paper.

2 Motivating Example

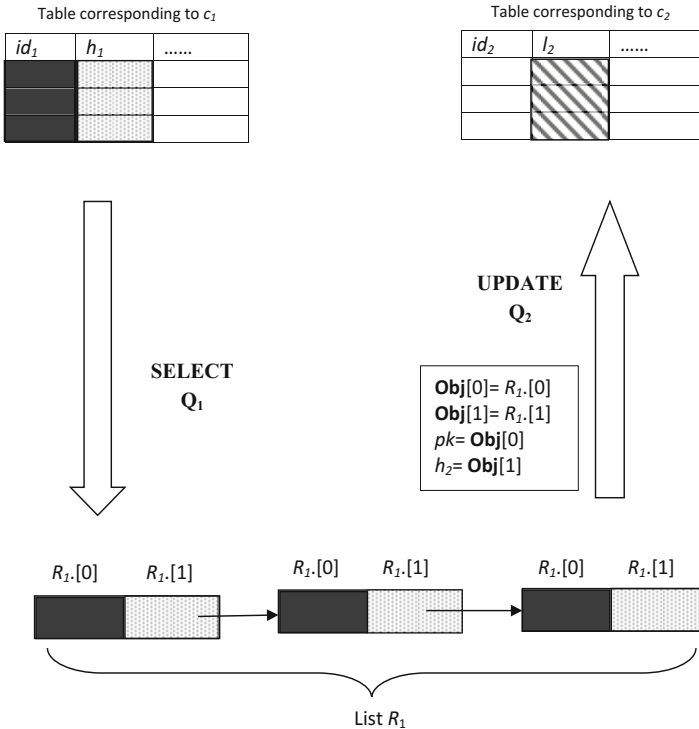
The language-based information flow security analysis has been applied in case of object-oriented languages, aiming at identifying possible information leakage to unauthorized users [9, 10, 12]. However, the conventional approaches do

```

1. Session session = getSessionFactory().openSession();
2. Transaction tx = session.beginTransaction();
3. Query Q1 = session.createQuery("SELECT id1, h1 FROM c1");
4. List R1 = Q1.list();
5. for(Object[] obj:R1{
6.     pk=(Int) obj[0];
7.     h2=(Int) obj[1];
8.     Query Q2 = session.createQuery("UPDATE c2 SET l2 = l2 + 1
                                     WHERE id2 = pk AND h2=1000");
9.     int result = Q2.executeUpdate();
10. tx.commit();
11. session.close();

```

(a) A HQL program *P*



(b) Execution view of *P*

Fig. 1. An example HQL program and its execution view

not fit to the case of HQL, when considering the sensitivity level of database information and influence on them through high-level HQL variables.

Consider, for instance, an example in Figure 1(a). Here, values of the table corresponding to the class c_1 are used to make a list, and for each element of the list an update is performed on the table corresponding to the class c_2 . Observe that there is an information-flow from confidential (denoted by h) to

public variables (denoted by l). In fact, the confidential database information h_1 which is extracted at statement 3, affects the public view of the database information l_1 at statement 8. This fact is depicted in Figure 1(b). The new challenge in this scenario *w.r.t.* state-of-the-art of information leakage detection is that we need to consider both application variables and SQL variables (corresponding to the database attributes).

3 Concrete and Abstract Semantics of HQL

We refer to the semantics of object-oriented programming language as defined in [11]. We just recall some basics of it. Then we formalize the concrete and abstract transition semantics of HQL, considering the Hibernate Session Objects, in order to identify possible information leakage.

3.1 Concrete Semantics

Object-Oriented Programming (OOP) language consists of a set of classes including a main class from where execution starts. Therefore, a program P in OOP is defined as $P = \langle c_{main}, L \rangle$ where Class denotes the set of classes, $c_{main} \in \text{Class}$ is the main class, $L \subset \text{Class}$ are the other classes present in P . A class $c \in \text{Class}$ is defined as a triplet $c = \langle \text{init}, F, M \rangle$ where init is the constructor, F is the set of fields, and M is the set of member methods in c .

Let Var , Val and Loc be the set of variables, the domain of values and the set of memory locations respectively. The set of environments is defined as $\text{Env} : \text{Var} \rightarrow \text{Loc}$. The set of stores is defined as $\text{Store} : \text{Loc} \rightarrow \text{Val}$.

The semantics of constructor and methods are defined below. Given a store s , the constructor maps its fields to fresh locations and then assigns values into those locations. Constructors never return output, but methods may return output.

Definition 1 (Constructor Semantics). *Given a store s . Let $\{a_{in}, a_{pc}\} \subseteq \text{Loc}$ be the free locations, $\text{Val}_{in} \subseteq \text{Val}$ be the semantic domain for input values. Let $v_{in} \in \text{Val}_{in}$ and pc_{exit} be the input value and the exit point of the constructor. The semantic of the class constructor init , $S[\text{init}] \in (\text{Store} \times \text{Val} \rightarrow \wp(\text{Env} \times \text{Store}))$, is defined by*

$$\begin{aligned} S[\text{init}](s, v_{in}) \\ = \left\{ (e_0, s_0) \mid (e_0 \triangleq V_{in} \rightarrow a_{in}, pc \rightarrow a_{pc}) \wedge (s_0 \triangleq s[a_{in} \rightarrow v_{in}, a_{pc} \rightarrow pc_{exit}]) \right\} \end{aligned}$$

Definition 2 (Method Semantics). *Let $\text{Val}_{in} \subseteq \text{Val}$ and $\text{Val}_{out} \subseteq \text{Val}$ be the semantic domains for the input values and the output values respectively. Let $v_{in} \in \text{Val}_{in}$ be the input values, a_{in} and a_{pc} be the fresh memory locations, and pc_{exit} be the exit point of the method m . The semantic of a method m , $S[m] \in (\text{Env} \times \text{Store} \times \text{Val}_{in} \rightarrow \wp(\text{Env} \times \text{Store} \times \text{Val}_{out}))$, is defined as*

$$\begin{aligned} S[m](e, s, v_{in}) = \left\{ (e', s', v_{out}) \mid (e' \triangleq e[V_{in} \rightarrow a_{in}, pc \rightarrow a_{pc}]) \wedge \right. \\ \left. (s' \triangleq s[a_{in} \rightarrow v_{in}, a_{pc} \rightarrow pc_{exit}]) \wedge v_{out} \in \text{Val}_{out} \right\} \end{aligned}$$

Object semantics in object-oriented languages is defined in terms of interaction history between the program-context and the object.

Set of Interaction States. The set of interaction states in object-oriented languages is defined by

$$\Sigma = \text{Env} \times \text{Store} \times \text{Val}_{\text{out}} \times \wp(\text{Loc})$$

where Env , Store , Val_{out} , and Loc are the set of application environments, the set of stores, the set of output values, and the set of addresses (escaped ones only) respectively.

Set of Initial Interaction States. The set of initial interaction states is defined by

$$\mathcal{I}_0 = \{ \langle e_0, s_0, \phi, \emptyset \rangle \mid S[\llbracket \text{init} \rrbracket](v_{in}, s) \ni \langle e_0, s_0 \rangle, v_{in} \in \text{Val}_{in} \}$$

Observe that ϕ denotes no output produced by the constructor and \emptyset represents the empty context with no escaped address.

Transition Relation. Let $\text{Lab} = (\mathbb{M} \times \text{Val}_{in}) \cup \{\text{upd}\}$ be a set of labels, where \mathbb{M} is the set of class-methods, Val_{in} is the set of input values and upd denotes an indirect update operation by the context.

The transition relation $\mathcal{T} : \text{Lab} \times \Sigma \rightarrow \wp(\Sigma)$ specifies which successor interaction states $\sigma' = \langle e', s', v', \text{Esc}' \rangle \in \Sigma$ can follow (i) when an object's methods $m \in \mathbb{M}$ with input $v_{in} \in \text{Val}_{in}$ is directly invoked on an interaction state $\sigma = \langle e, s, v, \text{Esc} \rangle$ (**direct interaction**), or (ii) the context indirectly updates an address escaped from an object's scope (**indirect interaction**).

Definition 3 (Direct Interaction \mathcal{T}_{dir}). Transition on Direct Interaction is defined below:

$$\mathcal{T}_{\text{dir}}[\llbracket m, v_{in} \rrbracket](\langle e, s, v, \text{Esc} \rangle) = \{ \langle e', s', v', \text{Esc}' \rangle \mid S[\llbracket m \rrbracket](\langle e, s, v_{in} \rangle) \ni \langle e', s', v' \rangle \wedge \text{Esc}' = \text{Esc} \cup \text{reach}(v', s') \}$$

where

$$\text{reach}(v, s) = \begin{cases} \text{if } v \in \text{Loc} \\ \quad \{v\} \cup \{ \text{reach}(e'(f), s) \mid \exists B. B = \{\text{init}, F, M\}, f \in F, \\ \quad s(v) \text{ is an instance of } B, s(s(v)) = e' \} \\ \text{else } \emptyset \end{cases}$$

Definition 4 (Indirect Interaction \mathcal{T}_{ind}). Transition on Indirect Interaction is defined below:

$$\mathcal{T}_{\text{ind}}[\llbracket \text{upd} \rrbracket](\langle e, s, v, \text{Esc} \rangle) = \{ \langle e, s', v, \text{Esc} \rangle \mid \exists a \in \text{Esc}. \text{Update}(a, s) \ni s' \}$$

where $\text{Update}(a, s) = \{s' \mid \exists v \in \text{Val}. s' = s[a \leftarrow v]\}$

Definition 5 (Transition relation \mathcal{T}). Let $\sigma \in \Sigma$ be an interaction state. The transition relation $\mathcal{T} : \text{Lab} \times \Sigma \rightarrow \wp(\Sigma)$ is defined as $\mathcal{T} = \mathcal{T}_{\text{dir}} \cup \mathcal{T}_{\text{ind}}$, where \mathcal{T}_{dir} and \mathcal{T}_{ind} represent direct and indirect transitions respectively.

Concrete Semantics of Session Objects. An attractive feature of HQL is the presence of `Hibernate Session` which provides a central interface between the application and `Hibernate` and acts as persistence manager. A transient object is converted into persistent state when associated with `Hibernate Session`, which has a representation in the underlying database. Various methods in `Hibernate Session` are used to propagate object's states from memory to the database (or vice versa).

We denote the abstract syntax of a `Session` method by a triplet $\langle C, \phi, \text{OP} \rangle$, where `OP` is the operation to be performed on the database tuples corresponding to a set of objects of classes $c \in C$ satisfying the condition ϕ . This is depicted in Table 1.

Following [7], the abstract syntax of any SQL statement Q is denoted by a tuple $\langle A, \phi \rangle$, meaning that Q first identifies an active data set from the database using a pre-condition ϕ that follows first-order logic, and then performs the appropriate operations A on the selected data set. For instance, the query “SELECT a_1, a_2 FROM t WHERE $a_3 \leq 30$ ” is denoted by $\langle A, \phi \rangle$ where A represents the action-part “SELECT a_1, a_2 FROM t ” and ϕ represents the conditional-part “ $a_3 \leq 30$ ”. The database environment ρ_d and the table environment ρ_t are defined as [7]:

Database Environment. We consider a database as a set of indexed tables $\{t_i \mid i \in I_x\}$ for a given set of indexes I_x . We define database environment by a function ρ_d whose domain is I_x , such that for $i \in I_x$, $\rho_d(i) = t_i$.

Table Environment. Given a database environment ρ_d and a table $t \in d$. We define $\text{attr}(t) = \{a_1, a_2, \dots, a_k\}$. So, $t \subseteq D_1 \times D_2 \times \dots \times D_k$ where, a_i is the attribute corresponding to the typed domain D_i . A table environment ρ_t for a table t is defined as a function such that for any attribute $a_i \in \text{attr}(t)$,

$$\rho_t(a_i) = \langle \pi_i(l_j) \mid l_j \in t \rangle$$

Where π is the projection operator, i.e. $\pi_i(l_j)$ is the i^{th} element of the l_j -th row. In other words, ρ_t maps a_i to the ordered set of values over the rows of the table t .

Given a HQL environment $e \in \text{Env}$, a HQL store $s \in \text{Store}$, and a database environment $\rho_d \in \mathbb{C}_d$. The concrete semantics of `Session` methods are defined by specifying how they are executed on (e, s, ρ_d) , resulting into new state (e', s', ρ_d') . These make the use of the semantics of database statements `SELECT`, `INSERT`, `UPDATE`, `DELETE` [7].

Fix-Point Semantics of HQL. We extend the notion of interaction states of OOP [11] to the case of HQL, considering the interaction of context with

Table 1. Abstract Syntax of Session Methods

Constants and Variables	
$n \in \mathbb{N}$	Set of Integers
$v \in \mathbb{V}$	Set of Variables
Arithmetic and Boolean Expressions	
$exp \in \mathbb{E}$	Set of Arithmetic Expressions
$exp ::= n \mid v \mid exp_1 \oplus exp_2$ where $\oplus \in \{+, -, *, /\}$	
$b \in \mathbb{B}$	Set of Boolean Expressions
$b ::= true \mid false \mid exp_1 \otimes exp_2 \mid \neg b \mid b_1 \circ b_2$ where $\otimes \in \{\leq, \geq, =, >, \neq, \dots\}$ and $\circ \in \{\vee, \wedge\}$	
Well-formed Formulas	
$\tau \in \mathbb{T}$	Set of Terms
$\tau ::= n \mid v \mid f_n(\tau_1, \tau_2, \dots, \tau_n)$ where f_n is an n-ary function.	
$a_f \in \mathbb{A}_f$	Set of Atomic Formulas
$a_f ::= R_n(\tau_1, \tau_2, \dots, \tau_n) \mid \tau_1 == \tau_2$ where $R_n(\tau_1, \tau_2, \dots, \tau_n) \in \{true, false\}$	
$\phi \in \mathbb{W}$	Set of Well-formed Formulas
$\phi ::= a_f \mid \neg \phi \mid \phi_1 \circ \phi_2$ where $\circ \in \{\vee, \wedge\}$	
HQL Functions	
$g(\vec{e}) ::= \text{GROUP BY}(e\vec{x}p) \mid id$ where $e\vec{x}p = \langle exp_1, \dots, exp_n \mid exp_i \in \mathbb{E} \rangle$	
$r ::= \text{DISTINCT} \mid \text{ALL}$	
$s ::= \text{AVG} \mid \text{SUM} \mid \text{MAX} \mid \text{MIN} \mid \text{COUNT}$	
$h(exp) ::= s \circ r(exp) \mid \text{DISTINCT}(exp) \mid id$	
$h(*) ::= \text{COUNT}(*)$ where * represents a list of database attributes denoted by \vec{v}_d	
$\vec{h}(\vec{x}) ::= \langle h_1(x_1), \dots, h_n(x_n) \rangle$ where $\vec{h} = \langle h_1, \dots, h_n \rangle$ and $\vec{x} = \langle x_1, \dots, x_n \mid x_i = exp \vee x_i = * \rangle$	
$f(e\vec{x}p) ::= \text{ORDER BY ASC}(e\vec{x}p) \mid \text{ORDER BY DESC}(e\vec{x}p) \mid id$	
Session Methods	
$c \in \text{Class}$	Set of Classes
$c ::= \langle \text{init}, F, M \rangle$ where init is the constructor, $F \subseteq \text{Var}$ is the set of fields, and M is the set of methods.	
$m_{ses} \in \mathbb{M}_{ses}$	Set of Session methods
$m_{ses} ::= \langle C, \phi, OP \rangle$ where $C \subseteq \text{Class}$	
$OP ::= \text{SEL}(f(e\vec{x}p'), r(\vec{h}(\vec{x})), \phi, g(e\vec{x}p))$ $\text{UPD}(\vec{v}, e\vec{x}p)$ $\text{SAVE}(\text{obj})$ $\text{DEL}()$ where ϕ represents 'HAVING' clause and obj denotes an instance of a class.	

Session objects. To this aim, we include database environment in the definition of HQL states. The set of interaction states of HQL is, thus, defined by

$$\Sigma = \text{Env} \times \text{Store} \times \mathcal{C}_d \times \text{Val}_{out} \times \wp(\text{Loc})$$

where Env , Store , \mathcal{C}_d , Val_{out} , and Loc are the set of application environments, the set of stores, the set of database environments, the set of output values, and the set of addresses respectively.

We now define the transition relation, by considering (i) the direct interaction, when a conventional method is directly invoked, (ii) the session interaction, when a `Session` method is invoked, and (iii) the indirect transition, when context updates any address escaped from the object's scope.

Definition 6 (Transition Relation \mathcal{T}). Let $\sigma \in \Sigma$ be an interaction state. The transition relation $\mathcal{T} : \text{Lab} \times \Sigma \rightarrow \wp(\Sigma)$ is defined as $\mathcal{T} = \mathcal{T}_{dir} \cup \mathcal{T}_{ind} \cup \mathcal{T}_{ses}$, where \mathcal{T}_{dir} , \mathcal{T}_{ind} and \mathcal{T}_{ses} represent direct, indirect, and session transitions respectively. Lab represents the set of labels which include `Session` methods \mathcal{M}_{ses} , conventional class methods \mathcal{M} , and an indirect update operation `Upd` by the context.

We denote a transition by $\sigma \xrightarrow{a} \sigma'$ when application of a label $a \in \text{Lab}$ on interaction state σ results into a new state σ' .

Let \mathcal{I}_0 be the set of initial interaction states. The fix-point trace semantics of HQL program P is defined as

$$\mathcal{T} \llbracket P \rrbracket (\mathcal{I}_0) = \text{lfp}_0^{\subseteq} \mathcal{F} (\mathcal{I}_0) = \bigcup_{i \leq \omega} \mathcal{F}^i (\mathcal{I}_0)$$

$$\text{where } \mathcal{F} (\mathcal{I}) = \lambda \mathcal{T}. \mathcal{I} \cup \left\{ \sigma_0 \xrightarrow{a_0} \dots \xrightarrow{a_{n-1}} \sigma_n \xrightarrow{a_n} \sigma_{n+1} \mid \sigma_0 \xrightarrow{a_0} \dots \xrightarrow{a_{n-1}} \sigma_n \in \mathcal{T} \right. \\ \left. \wedge \sigma_n \xrightarrow{a_n} \sigma_{n+1} \in \mathcal{T} \right\}$$

3.2 Abstract Semantics

Authors in [16, 17] used the Abstract Interpretation framework [3, 4] to define an abstract semantics of imperative languages using symbolic domain of positive propositional formulae in the form

$$\bigwedge_{0 \leq i \leq n, 0 \leq j \leq m} \{y_i \rightarrow z_j\}$$

which means that the values of variable z_j possibly depend on the values of variable y_i . Later, [8] extends this to the case of structured query languages. The computation of abstract semantics of a program in the propositional formulae domain provides a sound approximation of variable dependences, which allows to capture possible information flow in the program. The information leakage analysis is then performed by checking the satisfiability of formulae after assigning truth values to variables based on their sensitivity levels.

An abstract state $\sigma^\# \in \Sigma^\# \equiv \mathbb{L} \times \text{Pos}$ is a pair $\langle \ell, \psi \rangle$ where $\psi \in \text{Pos}$ represents the variables dependences, in the form of propositional formulae, among program variables up to the program label $\ell \in \mathbb{L}$.

Methods in HQL include a set of imperative statements¹. We assume, for the sake of the simplicity, that attackers are able to observe public variables inside of the main method only. Therefore, our analysis only aims at identifying variable dependences at input-output labels of methods.

The abstract transition semantics of constructors and conventional methods are defined below.

Definition 7 (Abstract Transition Semantics of Constructor). *Consider a class $c = \langle \text{init}, F, M \rangle$ where init is a default constructor. Let ℓ be the label of init . The abstract transition semantics of init is defined as*

$$\mathcal{T}^\# \llbracket \ell \text{init} \rrbracket = \{(\ell, \psi) \rightarrow (\text{Succ}(\ell \text{init}), \psi)\}$$

where $\text{Succ}(\ell \text{init})$ denotes the successor label of init . Observe that the default constructor is used to initialize the objects-fields only, and it does not add any new dependence.

The abstract transition semantics of parameterized constructors are defined in the same way as of class methods $m \in M$.

Definition 8 (Abstract Transition Semantics of Methods). *Let $U \in \wp(\text{Var})$ be the set of variables which are passed as actual parameters when invoked a method $m \in M$ on an abstract state (ℓ, ψ) at program label ℓ . Let $V \in \wp(\text{Var})$ be the formal arguments in the definition of m . We assume that $U \cap V = \emptyset$. Let a and b be the variable returned by m and the variable to which the value returned by m is assigned. The abstract transition semantics is defined as*

$$\mathcal{T}^\# \llbracket \ell m \rrbracket = \{(\ell, \psi) \rightarrow (\text{Succ}(\ell m), \psi')\}$$

where $\psi' = \{x_i \rightarrow y_i \mid x_i \in U, y_i \in V\} \cup \{a \rightarrow b\} \cup \psi$ and $\text{Succ}(\ell m)$ is the label of the successor of m .

We classify the high-level HQL variables into two distinct sets: Var_d and Var_a . The variables which have a correspondence with database attributes belong to the set Var_d . Otherwise, the variables are treated as usual variables and belong to Var_a . We denote variables in Var_d by the notation \bar{v} , in order to differentiate them from the variables in Var_a . This leads to four types of dependences which may arise in HQL programs: $x \rightarrow y$, $\bar{x} \rightarrow y$, $x \rightarrow \bar{y}$ and $\bar{x} \rightarrow \bar{y}$, where $x, y \in \text{Var}_a$ and $\bar{x}, \bar{y} \in \text{Var}_d$.

The abstract labeled transition semantics of various `Session` methods are defined in Table 2, where by $\text{Var}(\text{exp})$ and $\text{Field}(c)$ we denote the set of variables

¹ For a detailed abstract transition semantics of imperative statements, see [16].

in exp and the set of class-fields in the class c respectively. The function $\text{Map}(v)$ is defined as:

$$\text{Map}(v) = \begin{cases} \bar{v} & \text{if } v \text{ has correspondence with a database attribute,} \\ v & \text{otherwise.} \end{cases}$$

Notice that in Table 2 the notation \tilde{v} stands for either v or \bar{v} .

Table 2. Definition of Abstract Transition Function $\mathcal{F}^\#$ for Session methods

$\mathcal{F}^\# \llbracket \ell m_{save} \rrbracket$ $\stackrel{def}{=} \mathcal{F}^\# \llbracket \ell(C, \phi, \text{SAVE}(\text{obj})) \rrbracket$ $\stackrel{def}{=} \mathcal{F}^\# \llbracket \ell(\{c\}, \text{FALSE}, \text{SAVE}(\text{obj})) \rrbracket$ $\stackrel{def}{=} \{\langle \ell, \psi \rangle \xrightarrow{\text{SAVE}} \langle \text{Succ}(\ell m_{save}), \psi \rangle\}$
$\mathcal{F}^\# \llbracket \ell m_{upd} \rrbracket$ $\stackrel{def}{=} \mathcal{F}^\# \llbracket \ell(C, \phi, \text{UPD}(\vec{v}, e\vec{x}p)) \rrbracket$ $\stackrel{def}{=} \mathcal{F}^\# \llbracket \ell(\{c\}, \phi, \text{UPD}(\vec{v}, e\vec{x}p)) \rrbracket$ $\stackrel{def}{=} \{\langle \ell, \psi \rangle \xrightarrow{\text{UPD}} \langle \text{Succ}(\ell m_{upd}), \psi' \rangle\}$ where $\psi' = \bigwedge \{ \tilde{y} \rightarrow \tilde{z}_i \mid y \in \text{Var} \llbracket \phi \rrbracket, \tilde{y} = \text{Map}(y), \tilde{z}_i \in \vec{v} \} \cup$ $\bigwedge \{ \tilde{y}_i \rightarrow \tilde{z}_i \mid y_i \in \text{Var} \llbracket exp_i \rrbracket, exp_i \in e\vec{x}p, \tilde{y}_i = \text{Map}(y_i), \tilde{z}_i \in \vec{v} \} \cup \psi''$ and $\psi'' = \begin{cases} \psi \ominus (\tilde{a} \rightarrow \tilde{z}_i \mid \tilde{z}_i \in \vec{v} \wedge a \in \text{Var} \wedge \tilde{a} = \text{Map}(a)) & \text{if } \phi \text{ is TRUE by default.} \\ \psi & \text{otherwise} \end{cases}$
$\mathcal{F}^\# \llbracket \ell m_{del} \rrbracket$ $\stackrel{def}{=} \mathcal{F}^\# \llbracket \ell(C, \phi, \text{DEL}()) \rrbracket$ $\stackrel{def}{=} \mathcal{F}^\# \llbracket \ell(\{c\}, \phi, \text{DEL}()) \rrbracket$ $\stackrel{def}{=} \{\langle \ell, \psi \rangle \xrightarrow{\text{DEL}} \langle \text{Succ}(\ell m_{del}), \psi' \rangle\}$ where $\psi' = \bigwedge \{ \tilde{y} \rightarrow \tilde{z} \mid y \in \text{Var} \llbracket \phi \rrbracket, \tilde{y} = \text{Map}(y), \tilde{z} \in \text{Field}(c) \} \cup \psi''$ and $\psi'' = \begin{cases} \psi \ominus (\tilde{a} \rightarrow \tilde{z}_i \mid \tilde{z}_i \in \vec{v} \wedge a \in \text{Var} \wedge \tilde{a} = \text{Map}(a)) & \text{if } \phi \text{ is TRUE by default.} \\ \psi & \text{otherwise} \end{cases}$
$\mathcal{F}^\# \llbracket \ell m_{sel} \rrbracket$ $\stackrel{def}{=} \mathcal{F}^\# \llbracket \ell(C, \phi, \text{SEL}(f(e\vec{x}p'), r(\vec{h}(\vec{x})), \phi, g(e\vec{x}p))) \rrbracket$ $\stackrel{def}{=} \{\langle \ell, \psi \rangle \xrightarrow{\text{SEL}} \langle \text{Succ}(\ell m_{sel}), \psi' \rangle\}$ where $\psi' = \bigwedge \{ \tilde{y} \rightarrow \tilde{z} \mid y \in (\text{Var} \llbracket \phi \rrbracket \cup \text{Var} \llbracket \vec{e} \rrbracket \cup \text{Var} \llbracket \phi' \rrbracket \cup \text{Var} \llbracket \vec{e}' \rrbracket), z \in \text{Var} \llbracket \vec{x} \rrbracket, \tilde{y} = \text{Map}(y), \tilde{z} = \text{Map}(z) \} \cup \psi$

Let $\text{SF}(\psi)$ denotes the set of subformulas in ψ , and the operator \ominus is defined by $\psi_1 \ominus \psi_2 = \bigwedge (\text{SF}(\psi_1) \setminus \text{SF}(\psi_2))$.

4 Information Leakage Analysis

We are now in position to use the abstract semantics defined in the previous section to identify possible sensitive database information leakage through high-level HQL variables. After obtaining over-approximation of variable dependences at each program points, we assign truth values to each variable based on their sensitivity level, and we check the satisfiability of propositional formulae representing variable dependences [16].

Since our main objective is to identify the leakage of sensitive database information possibly due to the interaction of high-level variables, we assume, according to the policy, that different security classes are assigned to database attributes. Accordingly, we assign security levels to the variables in Var_d based on the correspondences. Similarly, we assign the security levels of the variables in Var_a based on their use in the program. For instance, the variables which are used on output channels, are considered as public variables. Observe that for the variables with unknown security class, it may be computed based on the dependence of it on the other application variables or database attributes of known security classes.

Let $\Gamma : \text{Var} \rightarrow \{L, H, N\}$ be a function that assigns to each of the variables a security class, either public (L) or private (H) or unknown (N).

After computing abstract semantics of HQL program P , the security class of variables with unknown level (N) in P are upgraded to either H or L , according to the following function:

$$\text{Upgrade}(v) = Z \text{ if } \exists (u \rightarrow v) \in \mathcal{S}^\# \llbracket P \rrbracket. \Gamma(u) = Z \wedge \Gamma(u) \neq N \wedge \Gamma(v) = N \quad (1)$$

We say that program P respects the confidentiality property of database information, if and only if there is no information flow from private to public attributes. To verify this property, a corresponding truth assignment function $\bar{\Gamma}$ is used:

$$\bar{\Gamma}(x) = \begin{cases} T & \text{if } \Gamma(x) = H \\ F & \text{if } \Gamma(x) = L \end{cases}$$

If $\bar{\Gamma}$ does not satisfy any propositional formula in ψ of an abstract state, the analysis will report a possible information leakage.

Let us illustrate this on the running example program P in section 2. According to the policy, let the database attribute corresponding to variable h_1 is private, whereas the attributes corresponding to id_1 , id_2 and l_2 are public. Therefore,

$$\Gamma(\bar{h}_1) = H \text{ and } \Gamma(\bar{id}_1) = \Gamma(\bar{id}_2) = \Gamma(\bar{l}_2) = L$$

For other variables in the program, the security levels are unknown. That is,

$$\Gamma(R_1.[0]) = \Gamma(R_1.[1]) = \Gamma(\text{obj}[0]) = \Gamma(\text{obj}[1]) = \Gamma(pk) = \Gamma(h_2) = N$$

Considering the domain of positive propositional formulae, the abstract semantics yields the following formulae at program point 9 in P :

$$\begin{array}{l} \bar{id}_1 \rightarrow R_1.[0]; \quad \bar{h}_1 \rightarrow R_1.[1]; \quad R_1.[0] \rightarrow \text{obj}[0]; \quad R_1.[1] \rightarrow \text{obj}[1]; \\ \text{obj}[0] \rightarrow pk; \quad \text{obj}[1] \rightarrow h_2; \quad pk \rightarrow \bar{l}_2; \quad \bar{id}_2 \rightarrow \bar{l}_2; \quad h_2 \rightarrow \bar{l}_2; \end{array}$$

According to equation 1, the security levels of the variables with unknown security level N are upgraded as below:

$$\begin{array}{l} \Gamma(R_1.[0]) = L, \Gamma(R_1.[1]) = H, \Gamma(\text{obj}[0]) = L, \Gamma(\text{obj}[1]) = H \\ \Gamma(pk) = L, \quad \Gamma(h_2) = H \end{array}$$

Finally, we apply the truth assignment function $\bar{\Gamma}$ which does not satisfy the formula $h_2 \rightarrow \bar{l}_2$, as $\bar{\Gamma}(h_2) = T$ and $\bar{\Gamma}(\bar{l}_2) = F$ and $T \rightarrow F$ is false.

Therefore, the analysis reports that the example program P is vulnerable to leakage of confidential database data.

5 Conclusions

Our approach can capture information leakage on “permanent” data stored in a database when a HQL program manipulates them. This may also lead to a refinement of the non-interference definition that focusses on confidentiality of the data instead of variables. We are now investigating a possible enhancement of the analysis integrating with other abstract domains.

Acknowledgments. Work partially supported by PRIN “Security Horizons” project.

References

1. Bauer, C., King, G.: *Hibernate in Action*. Manning Publications Co. (2004)
2. Bauer, C., King, G.: *Java Persistence with Hibernate*. Manning Publications Co. (2006)
3. Cousot, P., Cousot, R.: Abstract interpretation: a unified lattice model for static analysis of programs by construction or approximation of fixpoints. In: *Proceedings of the 4th ACM SIGACT-SIGPLAN Symposium on Principles of Programming Languages*, pp. 238–252. ACM Press, Los Angeles (1977)
4. Cousot, P., Cousot, R.: Systematic design of program analysis frameworks. In: *Proceedings of the 6th ACM SIGACT-SIGPLAN Symposium on Principles of Programming Languages*, pp. 269–282. ACM Press, San Antonio (1979)
5. Denning, D.E.: A lattice model of secure information flow. *Communications of the ACM* 19, 236–243 (1976)
6. Elliott, J., O’Brien, T., Fowler, R.: *Harnessing Hibernate*, 1st edn. O’Reilly (2008)
7. Halder, R., Cortesi, A.: Abstract interpretation of database query languages. *Computer Languages, Systems & Structures* 38, 123–157 (2012)
8. Halder, R., Zanioli, M., Cortesi, A.: Information leakage analysis of database query languages. In: *Proceedings of the 29th Annual ACM Symposium on Applied Computing (SAC 2014)*, March 24–28, pp. 813–820. ACM Press, Gyeongju (2014)

9. Hammer, C., Snelling, G.: Flow-sensitive, context-sensitive, and object-sensitive information flow control based on program dependence graphs. *International Journal of Information Security* 8, 399–422 (2009)
10. Li, B.: Analyzing information-flow in java program based on slicing technique. *SIGSOFT Software Engineering Notes* 27, 98–103 (2002)
11. Logozzo, F.: Class invariants as abstract interpretation of trace semantics. *Computer Languages, Systems & Structures* 35, 100–142 (2009)
12. Myers, A.C.: Jflow: Practical mostly-static information flow control. In: *Proceedings of the 26th ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*, pp. 228–241. ACM Press, San Antonio (1999)
13. Pottier, F., Simonet, V.: Information flow inference for ml. *ACM Transactions on Programming Languages and Systems* 25, 117–158 (2003)
14. Sabelfeld, A., Myers, A.C.: Language-based information-flow security. *IEEE Journal on Selected Areas in Communications* 21, 5–19 (2003)
15. Smith, S.F., Thober, M.: Refactoring programs to secure information flows. In: *Proceedings of the Workshop on Programming Languages and Analysis for Security*, pp. 75–84. ACM Press, Canada (2006)
16. Zanioli, M., Cortesi, A.: Information leakage analysis by abstract interpretation. In: Černá, I., Gyimóthy, T., Hromkovič, J., Jefferey, K., Královič, R., Vukolić, M., Wolf, S. (eds.) *SOFSEM 2011. LNCS*, vol. 6543, pp. 545–557. Springer, Heidelberg (2011)
17. Zanioli, M., Ferrara, P., Cortesi, A.: Sails: Static analysis of information leakage with sample. In: *Proceedings of the 27th Annual ACM Symposium on Applied Computing (SAC 2012)*, pp. 1308–1313. ACM Press, Trento (2012)

Investigation of Regularization Theory for Four-Class Classification in Brain-Computer Interface

Le Quoc Thang and Chivalai Temiyasathit

International Colledge, King Mongkuts Institute of Technology Ladkrabang,
Chalongkrung Rd., Ladkrabang, Bangkok 10520, Thailand
{s6610016,ktchival}@kmitl.ac.th

Abstract. Common spatial patterns (CSP) is one of the most prevalent feature extraction approaches that has been used in Brain-computer interfaces (BCI) due to its simplicity and efficiency. Nevertheless, CSP suffers from the problems of sensitivity to noise and overfitting. To overcome these issues, the regularized CSP (RCSP) has been proposed recently. In addition, CSP was originally designed for two-class classification. However, a practical BCI usually needs four-class commands to be able to operate. Thus, there is a high demand for increasing the performance of multi-class BCI. In this paper, we provide a complete study of classification accuracy in multi-class BCI using regularization theory, and compare it with the standard CSP to determine the suitable method for feature extraction in BCI learning. Besides CSP, linear discriminant analysis (LDA) has shown its robust and widespread use for machine learning in BCI. LDA estimates covariance matrices from extracted features. But for high-dimensional features with only a small amount of training data given, the estimation may become imprecise. In the attempt of clarifying the regularizing effects in BCI, this paper also provides the classification results of the regularized LDA (RLDA). The performance evaluation of this work was taken on data from 9 subjects, from BCI competition datasets. Results show that the combination of standard CSP and LDA has a slightly better accuracy than the regularizing methods.

Keywords: Brain-computer interfaces (BCI), common spatial patterns (CSP), Linear discriminant analysis (LDA), regularization, electroencephalography (EEG).

1 Introduction

Brain-computer Interfaces (BCIs) are systems that provide a novel communication channel, which enables users to control external devices using brain signals [16]. Most paralyzed patients lose muscle movement, but their minds are unaffected. Thus, translating human thoughts directly to the outside world can address the problem of communication for the patients. This process is generally measured by electroencephalography (EEG) due to its good quality and low-cost recording devices [1,13,14].

An overview of BCI system is shown in Figure 1. Typically, BCI system consists of different components i.e. signal acquisition, preprocessing, feature extraction and classification or usually called pattern recognition. This part in BCI is designed according to the feature extraction from EEG signals, i.e., using a classifier to separate the user's intentions from such EEG features [9]. One of the most popular feature extraction methods is the common spatial patterns (CSPs) that can extract the topographic patterns of brain oscillations by maximizing the discriminability of two classes [11,15]. The algorithm was extended to multiple class paradigm and significantly increased the performance in BCI application [4,5]. Despite the popularity and efficiency of CSP algorithm, it still has some drawbacks which are highly sensitive to noise or artifacts, data overfitting especially when the training sets are small. Moreover, it has been a challenging issue on the number of electrodes used for the recording data due to its effects on patient mentality and the effort to set up equipments [3]. Another problem in BCI is it requires long period of calibration sessions since the EEG signal has a huge inter-subject variability [1].

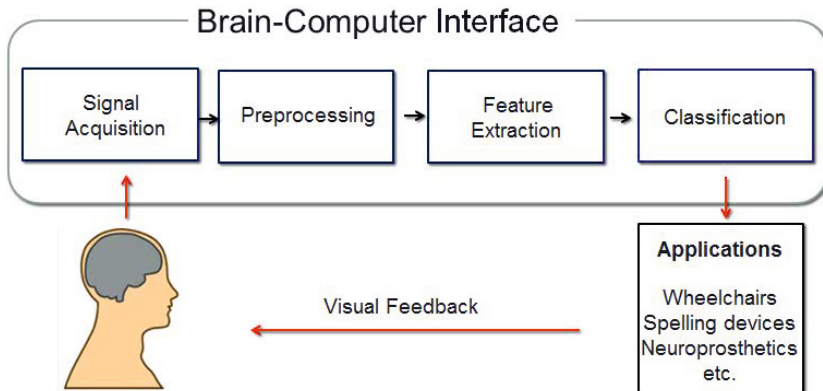


Fig. 1. Overview of the Brain-computer Interface system

To address these shortcomings, recent studies have added prior information into the CSP method which is the form of regularization [6,10]. F. Lotte et al. provided a comprehensive study about CSP regularization approaches. The authors compared classification accuracies of multiple CSP variants across three different datasets [10]. Results have shown that these regularized CSP outperformed the original CSP. However, the experiment only estimated the discrimination between two classes of EEG data. Moreover, in the calculation of CSP matrices the author chose fixed parameters which are the time interval and the pairs of spatial filters to extract features for all subjects, thus made it difficult to access the peak performances in case of multiple class problem. The use of BCI systems is severely limited because of the low bit-transfer rate. Hence, researchers have tried to extend the classification in BCIs into multi-class

paradigms to increase the bit transfer rate [4]. The performance of the system usually decreases as the number of classes increases, and the duration that user has to produce a continuous signal seems to be longer [3].

In this paper, we investigate the performance of BCI when applying regularization theory for four-class classification. Specifically, we select the CSP with Tikhonov Regularization (TRCSP), which is one of the best approaches in [10], for the feature extraction process. It should be mentioned that we use our preliminary study of increasing performance for four-class BCI to improve the accuracy [7] for each subject in the dataset in terms of selecting suitable parameters. Moreover, regularization has shown its effects on overcoming the imprecision in the calculation of covariance matrices [2]. Thus, it has been an unexplored aspect on whether or not should the regularization be implemented in both feature extraction (e.g. CSP) and machine learning (e.g. classification using linear approaches). To clarify this issue, we also study the regularizing approach for classification mechanism with linear discriminant analysis (LDA) or RLDA for short, and compare with the accuracies of the standard CSP and LDA. As a result, four different types of combinations are measured in our study, i.e., CSP/LDA, RCSP/LDA, CSP/RLDA and RCSP/RLDA. The experiment was done with the dataset IIa in BCI competition 2008.

2 Methodology

2.1 CSP Algorithm

The purpose of CSP is to learn spatial filters that maximize the variance of one class data, and simultaneously minimize the variance for the other class [11,15]. Since the variance of the band-pass filtered EEG signals in a given frequency band is equal to band-power [3], CSP aims at obtain an optimal discrimination of user's intentions. A solution for the following maximization problem is the spatial filters $w \in \mathbb{R}^n$ (n is the number of data channels) that are used in CSP technique:

$$\underset{w \in \mathbb{R}^n}{\text{maximize}} \quad J(w) = \frac{w^T C_1 w}{w^T C_2 w} \quad (1)$$

where C_i denotes the spatial covariance matrix from class i , calculated by $C_i = X_i^T X_i$ ($i \in \{1,2\}$), T denotes the transpose and X_i is the bandpass filtered EEG data matrix for class i . One way to solve this optimization problem is using the Lagrange multiplier method, maximizing $J(w)$ is equivalent to maximizing the following function:

$$L(\lambda, w) = w^T C_1 w - \lambda(w^T C_2 w - 1) \quad (2)$$

The filters w can be obtained by solving the derivative of L

$$\begin{aligned} \frac{\partial L}{\partial w} &= 2w^T C_1 - 2\lambda w^T C_2 = 0 \\ &\iff C_1 w = \lambda C_2 w \\ &\iff C_2^{-1} C_1 w = \lambda w \end{aligned}$$

The result matrix w that extremizing (1) is the eigenvectors of $M = C_2^{-1}C_1$, which corresponds to its lowest and largest eigenvalues.

2.2 CSP with Tikhonov Regularization

The regularization technique is used for overcoming the sensitivity of CSP to outliers and overfitting. In [10], Dr. F. Lotte et al. reviewed two means of regularization in CSP, which are the covariance matrix estimation level and the regularizing at the CSP objective function (1). The authors proposed the CSP with Tikhonov Regularization (TRCSP) that is based on the regularization of the CSP objective function. Generally, such method aims at penalize the spatial filters by adding a regularization term to the CSP objective function. The objective function formally becomes

$$J_{P_1}(w) = \frac{w^T C_1 w}{w^T C_2 w + \alpha P(w)} \quad (3)$$

where $P(w)$ is the penalty function and α is a user-defined regularization parameter ($\alpha > 0$). The more the spatial filters w satisfy a given prior, the lower $P(w)$. Hence, maximizing $J_{P_1}(w)$ is equivalent to minimizing $P(w)$. This regularization is expected to guide the optimization process toward good spatial filters, particularly in the limited or noisy training EEG data.

Tikhonov regularization was initially introduced for regression problems [10], which consists the penalty term $P(w) = \|w\|^2 = w^T w = w^T I w$ (where I is the identity matrix). This method is expected to mitigate the influence of noises and artifacts since it constrains the solution to filters with a small norm. With a quadratic penalty term, (3) becomes

$$J_{P_1}(w) = \frac{w^T C_1 w}{w^T C_2 w + \alpha w^T I w} = \frac{w^T C_1 w}{w^T (C_2 + \alpha I) w}$$

The corresponding Lagrange is

$$L_{P_1}(\lambda, w) = w^T C_1 w - \lambda(w^T (C_2 + \alpha I) w - 1) \quad (4)$$

By applying the same approach in the CSP algorithm section, the eigenvalue problem is as the following:

$$\begin{aligned} \frac{\partial L}{\partial w} &= 2w^T C_1 - 2\lambda w^T (C_2 + \alpha I) = 0 \\ &\iff w C_1 = \lambda w (C_2 + \alpha I) \\ &\iff (C_2 + \alpha I)^{-1} C_1 w = \lambda w \end{aligned}$$

Then, the result spatial filters w that satisfy the penalty function $P(w)$ are the eigenvectors corresponding to the largest eigenvalues of $M_1 = (C_2 + \alpha I)^{-1} C_1$. In normal CSP, the spatial filters are only the eigenvectors corresponding to both the largest and smallest eigenvalue of $M = C_2^{-1} C_1$. However, for RCSP in

general, or TRCSP specifically, the spatial filters are obtained by maximizing another objective function:

$$J_{P_2}(w) = \frac{w^T C_2 w}{w^T C_1 w + \alpha P(w)} \quad (5)$$

The eigenvectors correspond to the largest eigenvalues of $M_2 = (C_1 + \alpha I)^{-1} C_2$ are also used in the filters w for TRCSP. The reason for this is because the eigenvectors correspond to the lowest eigenvalues of M_1 minimize (3), and then maximize the penalty term. Thus, maximizing the function (5) help to obtain the filters that maximize C_2 while minimizing C_1 . In short, TRCSP uses the spatial filters that are eigenvectors corresponding to the largest eigenvalues of M_1 and to the largest eigenvalues of M_2 .

2.3 Linear Discriminant Analysis (LDA)

LDA is one of the most popular approaches in machine learning for BCI [1,9,12] due to its efficiency and simplicity. LDA separates two classes data with a hyperplane by minimizing the risk of misclassification [1]. It assumes that data follows the normal distribution with equal covariance for both classes. The function for classifying two classes data is given by the following :

$$w_0 + w^T x = y \quad (6)$$

with w_0 and w^T are calculated by:

$$\begin{aligned} \Sigma &= (\Sigma_1 + \Sigma_2)/2 \\ w_0 &= -1/2(\mu_1 + \mu_2)\Sigma^{-1}(\mu_1 - \mu_2)^T \\ w &= \Sigma^{-1}(\mu_1 - \mu_2)^T \end{aligned}$$

where Σ_j denotes the common covariance matrix and μ_j ($j = 1,2$) denotes the class mean. The separating function is obtained by calculating the projection that minimizes the distance of the interclass variance and maximizes it between two classes means [9]. For a two-class problem, the sign of y (e.g. $y > 0$ or $y < 0$) in (6) determines the class of a feature vector. To solve a multiclass problem, several separating functions are employed. The strategy that is generally used in four-class BCI is the one-versus-the-rest, which composes of discriminations for each class from all the others.

2.4 Regularized Linear Discriminant Analysis (RLDA)

The LDA needs to calculate the covariance matrix for features of each class. However, if the training data set is noisy or small, the estimation may become imprecise. Thus, it degrades the classification performance of LDA. It is appropriate to add prior information to these estimates by using regularization terms. One form of regularizing covariance matrix in literature is the method of

Ledoit and Wolf [8], which consists in shrinking the covariance matrix toward the identity matrix. Based on the method, it can be performed as follows:

$$\tilde{\Sigma}_j = (1 - \gamma)\Sigma_j + \gamma I \quad (7)$$

where $\tilde{\Sigma}_j$ is the regularized estimate, I is the identity matrix, γ is a parameter that can be automatically identified using then method in [8]. The aim of γ is to shrink the covariance matrix toward the identity matrix, to neutralize a possible estimation bias in small training set. The regularized covariance matrix $\tilde{\Sigma}_j$ is then used for calculating the function (6) above.

3 Experiments

3.1 EEG Datasets

In order to access and compare the performance of regularization theory in the four-class classification BCI, we used EEG data from 9 subjects, from the publicly available datasets BCI competition IV - data set IIa [12]. The dataset contains motor imagery (MI) EEG signals, i.e., data recorded while subjects were performing left hand, right hand, foot and tongue MI. The EEG signals were measured using 25 electrodes which composed of 22 EEG signals from brain activities and 3 Electrooculography (EOG) signals from eye movements. Only signals corresponding to brain oscillations were used for this experiment to investigate the performance of CSP and TRCSP in the noisy environment. Each subject had done the recording in two sessions on different days, 288 trials in each session which contains 72 trials for each MI task. The first data session of a subject was used for training, the testing phase was performed on the other data session.

3.2 Data Processing

In this paper, we design a paradigm that used for four-class BCI training as in Figure 2. At the beginning, data was divided into separate trials. To eliminate the noise from the initial period, the first 0.5s of each trial after the cue instructed the subject to perform MI task is omitted in this study. In the preprocessing process, the bandpass filter was applied for the training data in the frequency range of 7-30Hz, as recommended in [3], in order to cover the mu (7-12Hz) and beta (18-26Hz) bands of brain oscillations in which the MI tasks are characterized.

As mentioned before, for the feature extraction process, the CSP and TRCSP were used to measure the effects of regularization theory. In literature, CSP for multi-class problem is designed, as in [4], by combining two or more spatial filters, then can reduce the multi-class classification into several binary decisions. In the paradigm, four CSP filters (one spatial filter for each MI task) in one-versus-the-rest schema were used. There are some parameters that need to be considered when estimating the spatial filters. The first one is the time

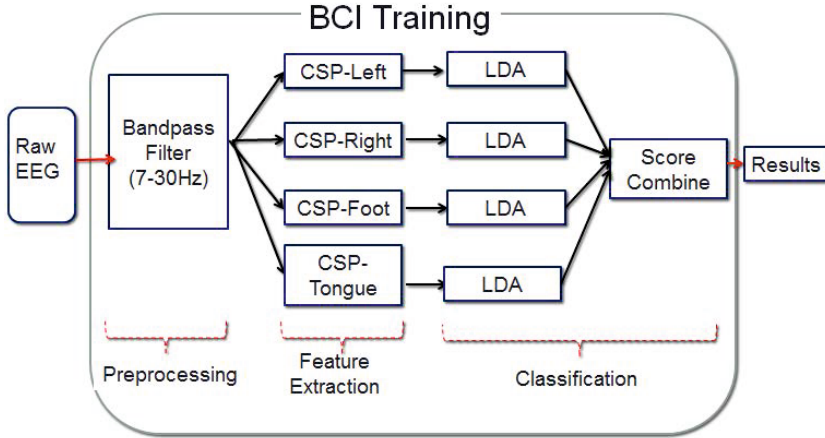


Fig. 2. Four-class BCI training paradigm

interval in order to calculate covariance matrices i.e. C_1 , C_2 . Most of other experiments, they employed a fixed time interval for all subjects [14,10]. After obtaining the spatial filter w , features of each trial are extracted by the function $\log(w^T X X^T w)$. It reveals the second parameter which is the number of components extracted from preprocessed EEG data, and it depends on the number of columns of matrix w .

Studies have shown that the classifier's accuracy depends on both the two parameters i.e. the number of columns in the spatial filters and the time interval used in the feature extraction process [3,12,17]. Thus, we divide the time range in each trial into multiple sub-intervals whose length varies from 1s - 3.5s to find an optimal time range for each subject. Similarly, we use three, four and five pair of filters, as the result of our preliminary work, in [7], to extract features from the EEG signals in order to obtain the peak performance for each person i.e. classifiers run through all combinations of parameters to learn the optimal parameters.

Besides the normal parameters of CSP, the TRCSP algorithm has one more regularization parameter which is α , whose value must be defined by the user. As in [10], we selected values among the set $[10^{-10}, 10^{-9}, \dots, 10^{-1}]$ for α in the training phase.

3.3 Classification and Validation

The log-variances of the spatially filtered EEG data were then used as input data to the corresponding LDA classifier. In order to classify four different subsets of features from four CSP filters, four LDA classifiers were employed to discriminate data in each feature subset. The output of these four classifiers contains the class labels of each binary problem (e.g. left hand imagery versus rest), and the probability of that class label. Final result for one trial training is based

on the combination of the four LDA classifier by applying the method of majority voting strategy, i.e., the class label is assigned to the classifier that has the largest probability value. After evaluating the standard procedure, which employs the standard CSP for feature extraction and LDA for classification, the other two approaches (e.g. RCSP and RLDA) then replace the methods in extracting features and classification mechanism respectively. As a result, the performances of four combinations (i.e., CSP/LDA, RCSP/LDA, CSP/RLDA and RCSP/RLDA) were estimated in the experiment.

To prevent over-fitting in the calibration process, the 9-fold cross validation procedure was performed, meaning that the data were partitioned into 9 equal parts, in which one part was used as test set and the other 8 parts were employed in model training. This process was repeated nine times with alternation of the testing part. The models with the best average classification results were then applied in the testing with unseen dataset.

3.4 Results and Discussion

For the purpose of comparison, each subject in the dataset was tested in both standard CSP and the TRCSP algorithms for the features extraction, and then took the output features for the classification using LDA and RLDA classifiers. The first data session in the data set was used for training data models and the testing accuracies were measured on the other session. Table 1 reports the classification performance in the training phase obtained from the data sets.

Table 1. Classification accuracies in the training phase for each subject

Subject	CSP/LDA	RCSP/LDA	CSP/RLDA	RCSP/RLDA
S1	79.51	77.43	65.28	65.63
S2	71.18	69.09	64.93	65.28
S3	85.42	86.81	82.64	81.25
S4	65.94	56.94	54.86	56.25
S5	45.14	45.14	45.49	45.48
S6	51.74	51.74	47.92	52.08
S7	79.17	80.91	71.53	72.57
S8	82.64	85.76	81.59	82.29
S9	83.33	83.68	85.07	85.07
Average	70.56	70.83	66.59	67.32
Std	15.3	15.9	14.19	14.1

Results show that, the standard CSP/LDA and RCSP/LDA perform slightly better than the other combinations for this data set in terms of average results of classification accuracy in training. The highest mean, which belongs to the RCSP/LDA method is at 70.83%, with CSP/LDA is very close behind at 70.56%. The other methods obtained little lower average results, it decreases approximately 3% - 4%.

The training results in some subjects are decreased from 2% - 9% when using the regularization theory (e.g. subjects S1, S2, S4), which means their highest training results lie in the combination CSP/LDA. There are a slight increase of subjects S3, S6, S7, S8, S9 around 1%-3% in the regularization schemas i.e. using RCSP or RLDA in the training for four-class BCI. Subject S5 nearly obtains the same accuracy in each combination. The models with the best results, listed in Table 1, are tested with the unseen data sessions to determine the performance of each model.

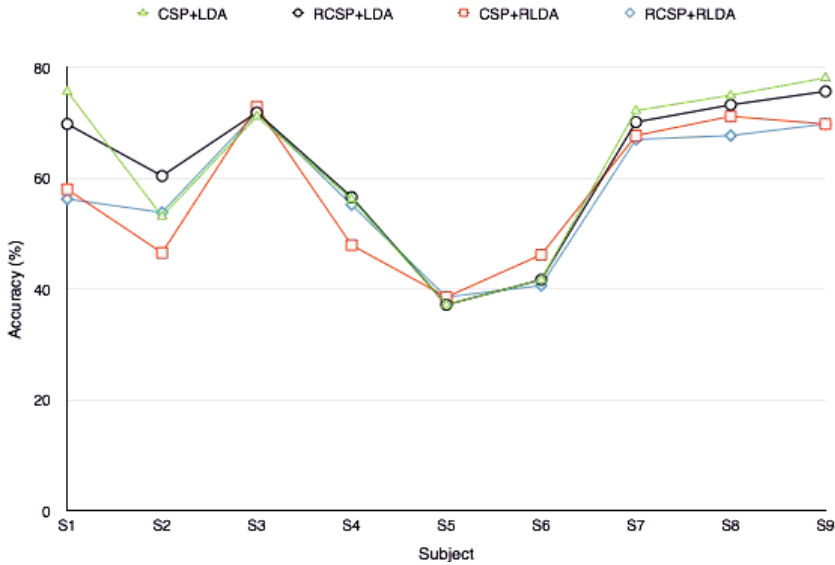


Fig. 3. Testing accuracies of four types of combinations for 9 subjects

Figure 3 illustrates the classification accuracies in the testing phase of 9 subjects for four different types of combinations, i.e., CSP/LDA, RCSP/LDA, CSP/RLDA and RCSP/RLDA. CSP/LDA observed a better performance in most subjects among four methods. This, together with the result for each subject in the training phase, suggest that the RCSP is not effective for subjects with good initial performance (e.g., S1, S3, S7, S8, S9). In fact, only subject S2, S4 were observed an increase in its testing performance when using the RCSP, the others performed roughly the same in both standard CSP and RCSP.

Testing results from the RLDA method reveal that it is not efficient to apply the regularization theory in LDA in this case since the dimension of extracted features had been optimized by exhaustive search, which means that the dimension of features is relatively small in comparison with the training samples. Precisely, most subjects in the RLDA scheme acquired 1% - 9% lower testing accuracy than it does in the LDA method. Only subjects S3, S5 and S6 showed the increase in CSP/RLDA combination. However, two subjects S5 and S6 reached very low testing accuracy (< 50%), which is far from acceptable, then its result

is rather be random. Interestingly, subject S3 possessed nearly the same testing performance in different methods (about 72%). It shows that the regularizing technique need to be accessed more in other cases, in which the large training sets are not available, to see the truth effects of the regularization.

Our experiment result for four-class classification is agreed well with the result in [10] for two-class problem, it also showed a slightly improvement for the dataset when applying the TRCSP. Base on this, we concluded that the standard CSP/LDA algorithm is more robust for already good and clean data. In addition, the RCSP needs to turn the (α) parameter which consumes more time when training BCI.

4 Conclusion

In this paper, we provide a complete study of regularization theory in multiple classes BCI. We also give a performance comparison between the standard CSP/LDA and RCSP/RLDA. The study evaluated the algorithms on EEG data from 9 subjects, from BCI competition dataset. Results show that the standard CSP/LDA performed slightly better the other combinations (i.e., RCSP/LDA, CSP/RLDA and RCSP/RLDA). It leads to the conclusion of using regularizing technique for BCI, i.e., it is not necessary to apply the regularization for subjects whose performance are good in the training period. We, therefore, would recommend that in training BCI we could detect the performance of a subject in the first training and apply the regularization in case the produced signals are not in good quality.

Our future work will be possibly related to measuring performances of the regularization algorithms with very small data sets, since the methods aim at dealing with noisy and limited data. It would also be interesting to build an online feedback BCI system, as in [1], to study the performance of the proposed approach for multiclass BCI.

Acknowledgments. This work is supported by King Mongkuts Institute of Technology Ladkrabang [KREF125608] and AUN/SEED-Net.

References

1. Blankertz, B., Dornhege, G., Krauledat, M., Müller, K.R., Curio, G.: The non-invasive Berlin Brain-Computer Interface: Fast acquisition of effective performance in untrained subjects. *NeuroImage* 37(2), 539–550 (2007)
2. Blankertz, B., Lemm, S., Treder, M., Haufe, S., Müller, K.R.: Single-trial analysis and classification of ERP components—a tutorial. *NeuroImage* 56(2), 814–825 (2011), <http://www.ncbi.nlm.nih.gov/pubmed/20600976>
3. Blankertz, B., Tomioka, R., Lemm, S., Kawanabe, M., Muller, K.R.: Optimizing spatial filters for robust EEG single-trial analysis. *IEEE Signal Processing Magazine*, 41–56 (January 2008)

4. Donoghue, J., Blankertz, B., Curio, G., Müller, K.: Boosting bit rates in non-invasive EEG single trial classification by feature combination and multi class paradigm. *IEEE Trans. Biomed. Eng.* (2004)
5. Dornhege, G., Blankertz, B., Curio, G., Müller, K.: Increase Information Transfer Rates in BCI by CSP Extension to Multi-class. In: *NIPS* (2003)
6. Kang, H., Nam, Y., Choi, S.: Composite Common Spatial Pattern for Subject-to-Subject Transfer. *IEEE Signal Processing Letters* 16(8), 683–686 (2009)
7. Le, T., Temiyasathit, C.: Increase performance of four-class classification for Motor-Imagery based Brain-Computer interface. In: *2014 International Conference on Computer, Information and Telecommunication Systems (CITS 2014)*, Jeju, Korea (July 2014)
8. Ledoit, O., Wolf, M.: A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis* 88(2), 365–411 (2004), <http://linkinghub.elsevier.com/retrieve/pii/S0047259X03000964>
9. Lotte, F., Congedo, M., Lécuyer, A., Lamarche, F., Arnaldi, B.: A review of classification algorithms for EEG-based brain-computer interfaces. *Journal of Neural Engineering* 4(2), R1–R13 (2007), <http://www.ncbi.nlm.nih.gov/pubmed/17409472>
10. Lotte, F., Guan, C.: Regularizing common spatial patterns to improve BCI designs: Unified theory and new algorithms. *IEEE Transactions on Bio-medical Engineering* 58(2), 355–362 (2011)
11. Mu, J., Pfurtscheller, G., Flyvbjerg, H.: Designing optimal spatial filters for single-trial EEG classification in a movement task. *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology* 110 (1999)
12. Naeem, M., Brunner, C., Leeb, R., Graimann, B., Pfurtscheller, G.: Separability of four-class motor imagery data using independent components analysis. *Journal of Neural Engineering* 3(3), 208–216 (2006)
13. Nijholt, A., Tan, D.: Brain-Computer Interfacing for Intelligent Systems. *IEEE Intelligent Systems* 23(3), 72–79 (2008)
14. Novi, Q., Guan, C., Dat, T.H., Xue, P.: Sub-band Common Spatial Pattern (SBCSP) for Brain-Computer Interface. In: *2007 3rd International IEEE/EMBS Conference on Neural Engineering*, pp. 204–207 (May 2007)
15. Ramoser, H., Müller-Gerking, J., Pfurtscheller, G.: Optimal spatial filtering of single trial EEG during imagined hand movement. *IEEE Transactions on Rehabilitation Engineering: A Publication of the IEEE Engineering in Medicine and Biology Society* 8(4), 441–446 (2000)
16. Wolpaw, J.R., Birbaumer, N., McFarland, D.J., Pfurtscheller, G., Vaughan, T.M.: Brain-computer interfaces for communication and control. *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology* 113(6), 767–791 (2002), <http://www.ncbi.nlm.nih.gov/pubmed/12048038>
17. Yijun, W., Shangkai, G., Xiaorong, G.: Common Spatial Pattern Method for Channel Selection in Motor Imagery Based Brain-computer Interface, pp. 5392–5395 (2005)

Enhancing Genetic Algorithm with Cumulative Probabilities to Derive Critical Test Scenarios from Use-Cases

An T. Tran, Tho T. Quan, and Thuan D. Le

Faculty of Computer Science and Engineering
Ho Chi Minh City University of Technology
Ho Chi Minh, Vietnam
tranthienan.90@gmail.com,
{qttho,thuanle}@cse.hcmut.edu.vn

Abstract. Derivation of test scenarios from requirement presented by use-cases is an emerging methodology in software testing. Path analysis is commonly used to derive the most effective test-cases in white-box testing technique. However, when dealing with sophisticated and complex use-cases, this approach is unfeasible due to the extremely large number of possibilities. Genetic algorithm (GA) technique has been proposed to prioritize test scenarios, allowing the most critical ones to be taken into account first. In this paper, we further enhance this technique by employing cumulative probabilities to improve the convergence speed and fitness quality of the generated test scenarios. Our experiments have proven the better effectiveness achieved compared to other previous works in this approach.

Keywords: Genetic algorithm, cumulative probability, test scenarios generation.

1 Introduction

Testing a software or a large system is a complex and time-consuming task, not to mention it also bears a substantial cost [2]. Hence, researchers have conducted much investigation into implementing the testing phase automatically to alleviate the cost in the software development process. There are many approaches reported when it comes to dealing with this issue. White-box testing technique, which calls for internal design of the system to derive test cases [2], is one of them. Accordingly, it is hard as well as expensive to perform and usually done manually.

Use case is the *de facto* artifact for testers to derive test cases in professional industry nowadays since it helps to specify the system behaviors from user's perspective. Since Unified Modeling Language (UML) has become a widely accepted standard for system specification, strategies for automatic acquisition of test cases from UML-based diagrams have been remarkably suggested [8]. However, in reality, use-cases are often expressed in natural language first, instead of being directly composed in UML forms.

Evolutionary Testing is an emerging methodology for automatically producing high quality test data [8]. Genetic algorithm (GA) is a notable form of the evolutionary algorithms, and it is suitable for problem whose search space is large. Some previous researches in testing have applied it for white-box testing approaches, such as the work discussed in 1.

In this paper, we proposed a new approach to deal with the problem of test case derivation from requirement presented in use-case. First, we developed a GUI-based program to assist users in designing use-cases, actors involving in the system and the relation between them in a use-case diagram. It also allows users to input detailed specification for each use-case. After that, the test case scenarios can be extracted from activity diagrams based on the use-case specification. However, in a sophisticated and complex use-case, the act of generating all those scenarios is unfeasible due to the extremely large number of possibilities. Therefore, based on the idea of [1], a GA-based approach is applied in order to prioritize these possible scenarios. Especially, we enhance the GA algorithm with cumulative probabilities, inspired from the work reported in [4]. The usage of cumulative probabilities allows higher chances for the GA algorithm to explore other execution path once the currently investigated flow seems to get stuck. Therefore, our enhanced algorithm is able to produce test scenarios in faster speed and better fitness quality, as proven by experiment. It is also our major contribution in this paper.

The rest of this paper is organized as follows: in Section 2, we will touch upon test scenarios derivation from activity diagram. Section 3 gives the overview of test scenarios derivation using GA. In Section 4, we will go into details on our proposed GA algorithm. Then, Section 5 presents our experimental results. Finally, Section 6 concludes the paper.

2 Derivation of Test Scenarios from Activity Diagram

Activity diagrams are commonly used to describe business processes and data flow in a system. They also come in handy when modeling the logic captured from a use case. The data flow or activities are described in a sequential or concurrent fashion. Moreover, the main advantage of this diagram is its simplicity and ease of understanding the flow of logic of the system. Then, at the test generation phase, the activity diagram's flows are extracted so as to derive test scenarios.

Based on the work presented in [3], we formalize an activity diagram and a test scenario as follows:

Definition 1 (Activity diagram): An activity diagram is a tuple $D = (A, T, F, C, a_i, fr)$, where

- $A = \{a_1, a_2, \dots, a_m\}$ is a finite set of activity states.
- $T = \{t_1, t_2, \dots, t_n\}$ is a finite set of completion transitions.
- $C = \{c_1, c_2, \dots, c_n\}$ is a finite set of guard conditions, and C_i is in the corresponding transition t_i .
- $F \subseteq (A \times T \times C) \cup (T \times C \times A)$ is a flow relation between activities and transitions.

$a_i \in A$ is the initial activity state, $a_f \in A$ is the final activity state there is only one transition t such that $(a_i, t) \in F$, and $(t', a_i) \notin F$ and $(a_f, t') \notin F$ for any t' .

Definition 2 (Test Scenario): Let $D = (A, T, F, C, a_i, a_f)$ be an activity diagram, TS be the set of test scenarios of D ; CS be the current state. Furthermore:

- $t, t \cdot$ denote pre-set and post-set of t respectively.
- $enabled(CS)$ is all of transitions started from CS .
- $fired(CS)$ denotes the only fireable transition from CS at certain moment, then $fired(CS) = \{t | t \in enabled(CS)\}$ and $(CS \dashv t) \cap t \cdot = \emptyset$ and after t was fired the new current state $CS' = (CS \dashv t) \cup t \cdot$ if there are more than one transition satisfies the condition, we can randomly choose one which is still unfired.
- $t_s \in TS$ is a sequence of activity states and conditioned transition, then $t_s = CS_0 \xrightarrow{[c_0]t_0} CS_1 \xrightarrow{[c_1]t_1} \dots \xrightarrow{[c_{n-1}]t_{n-1}} CS_n$ where $CS_0 = \{a_i\}$, $CS_n = \{a_f\}$, CS_i is current state, and $t_i = fired(CS_i)$, C_i is the guard condition on $t_i, i \geq 0, CS_i = CS_{i-1} \dashv t_{i-1} \cup t_{i-1} \cdot, i \geq 1$.

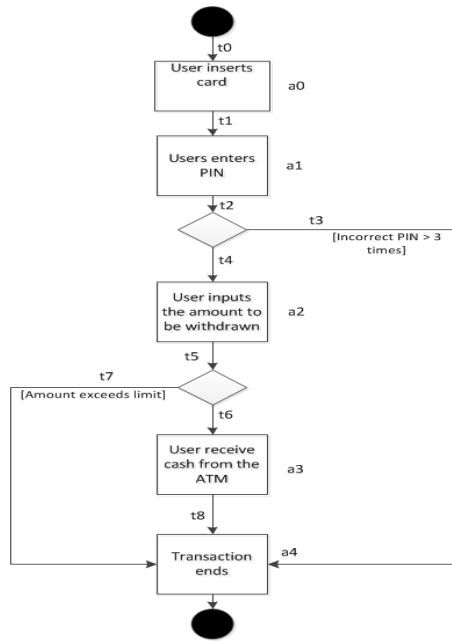


Fig. 1. An activity diagram

There have been some researches in automatically generating test scenario from activity diagrams. In [2], an approach is presented in which the activity diagram is

constructed from use-case. Then, a graph can be generated based on the states and transitions. Finally, graph traversal is applied to derive all scenarios from the activity graph. Let us illustrate this approach with a simple example depicted in Figure 1. It could be easily seen that on the diagram, there are 3 scenarios corresponding to 3 covering paths, namely: (i) $a_0 \rightarrow a_1 \rightarrow a_2 \rightarrow a_3 \rightarrow a_4 \rightarrow a_5$; (ii) $a_0 \rightarrow a_1 \rightarrow a_4$; and (iii) $a_0 \rightarrow a_1 \rightarrow a_2 \rightarrow a_4$. From that, there are three test scenarios as presented in Table 1.

Table 1. Test scenarios obtained from the activity graph in Figure 1

ID	Action	Condition
Test Scenario 1	User inserts card User enters PIN User enters amount to be withdrawn User receives cash from the ATM End	Correct PIN Amount < limit
Test scenario 2	User inserts card User enters PIN End	Incorrect PIN input > 3 times
Test scenario 3	User inserts card User enters PIN User enters amount to be withdrawn End	Amount > limit

However, this approach is only feasible for small and medium activity diagrams in which the number of possible paths can be obtained by traversing. Supposing we have a complicated activity diagram whose steps are extremely large, it will be a cumbersome and time-consuming task. More importantly, the diagram can contain considerable loops, especially backward and inner ones, which leads to significant path explosion. Additionally, in some cases, the act of deriving all possible is not only costly but also redundant because we just have to test a subset of it in order to make sure of path coverage.

Therefore, an emerging approach when dealing with this issue is to extract “best” path from a given activity diagram. However, as finding all of possible test scenarios is a well-known NP-hard problem (equivalent to the SAT problem), the act of finding the best test scenario is thus also a NP-hard problem as well. This is when we should fall back on alternative optimization methods, among which a technique approach recently emerging is using genetic algorithms.

3 Derivation of the Best Test Scenarios Using Genetic Algorithm

3.1 Genetic Algorithm

Genetic Algorithms (GAs) are a part of evolutionary computing which is a subfield of artificial intelligence. GAs have manifested itself to be effective for exploring NP-hard

or complex non-linear search spaces, in comparison with intensive exhaustive search. The strength of GAs has been shown in their ability to exploit in a highly efficient manner in a large number of individuals. GAs are commonly used to solve a variety of problems and are becoming an important tool in machine learning and function optimization.

GAs are search algorithms inspired by natural selection in which fittest individuals will survive after a series of selection. That individual in computer science field is represented as a string of binary digits called *chromosome*. Each bit composed. A chromosome consists of *DNA* which is a single digit and a sequence of DNA constitutes *gene*.

The pseudo code of a GA is as follows:

```

Initialize (population)
Evaluate (population)
While (stopping condition not satisfied)
{
    Selection (population)
    Crossover (population)
    Mutate (population)
    Evaluate (population)
}

```

The main task in each step is as follows:

- *Population initialization*: The initial population can be randomly created when the chromosome's length has been identified.
- *Population evaluation*: The population is evaluated based on the predefined fitness function.
- *Selection*: The purpose of this step in GAs is to select out potential individuals for mating in order to produce new offspring. In [1], the selection phase solely relies on fitness value in which individuals with highest fitness values are best candidates for the mating step.
- *Crossover*: After selection, the crossover operation is applied to the selected chromosomes. They will be swapped their genes with each other.
- *Mutation*: The DNA or single bit in the chromosome will be changed in this step. This operation helps to make the population more diverse.

3.2 Previous Work

Next, we will introduce the approach of [1] on applying GAs in activity diagram so as to prioritize test scenarios. In the next section, we will introduce our proposed approach by enhancing this work by using cumulative probabilities.

Initializing the Population

In order to create the initial population, we have to identify the length of each chromosome first. A chromosome is used to represent a possible path in the activity diagram. We will use binary digits to represent chromosomes and their lengths depend on numbers of decision nodes and numbers of out-going paths of each decision node in the activity diagram. For instance, given an activity diagram with 3 decision nodes and each decision node has 2 outgoing paths, so we need 1 binary digit to describe each decision node. Therefore, it requires a string whose length is 3 binary digits to represent all its possible paths.

After the chromosome length is identified, we can randomly create the initial population.

Evaluating the Population

The fitness function is used to evaluate each individual in the population. The fitness value of each path is comprised of the value of each node belonging to that path. The value of each node is calculated based on 2 pieces of information: *stack-based-weight value* and *information flow (IF)* metric, which are described as follows.

- Stack-based-weight value is the number of operation to access an element in a stack which is formed by depth-first-search traversal.
- IF metric is calculated by using the following equation:

$$IF(A) = FANIN(A) \times FANOUT(A)$$

Where $FANIN(A)$ is the count of the number of other nodes that can call, or pass control to node A and $FANOUT(A)$ is the number of nodes that are called by node A .

Therefore, the sum of the weight of a node by stack-based-weight assignment approach and IF complexity contributes to the total weight of a node in an activity diagram.

$$w_A = IF(A) + STACKBASEDWEIGHT(A)$$

Hence, the fitness value for each path in the activity diagram equals the total of its belonging nodes.

$$F = \sum_{i=1}^n w_i$$

Selection

In this phase, two paths whose fitness values are the highest will be selected as parents for the mating process.

Algorithm

Algorithm 1. Original genetic algorithm for prioritization of test case scenarios [1]

- 1: Convert the activity diagram into Control Flow Graph (CFG)
 - 2: Use the decision nodes to generate the test data or chromosome population randomly
 - 3: **foreach** test data $i = 1 \dots n$ **do**
 - a) Traverse the CFG by applying Depth first search (DFS) as well as Breadth first search (BFS) and identify the path.
 - i. Find the neighbor node for the current node having next higher depth d_i by applying DFS
 - ii. For each concurrent node, c_i traverse the next neighbor node having next higher breadth value, b_i
 - iii. Update the top pointer, size s and k of the Stack
 - iv. Assign weight, w to each node by applying weight assignment algorithm
 - b) Calculate the fitness value of each test data by using (2) and stack based weight assignment approach.
 - c) For sub activity of each node, calculate the fitness value using 80-20 rule and by using (2).
 - d) Select initial test data by ranking the fitness of chromosomes
 - e) If initial population is not enough then randomly generate them
 - if** $r < 0.8$ **then** perform crossover
 - else if** $r < 0.2$ **then** perform mutation
 - 4: **if** test data for all the paths have not been covered **then**
Repeat the GA process
else
End system
-

Example

Let us illustrate this approach with the activity diagram presented in Figure 2.

There are 6 decision nodes in this activity diagram. Each of them has 3 outgoing paths, which requires 2 bits to represent each node. Therefore, each path need $2 * 6 = 12$ bits to describe itself. In other words, the essential length for each chromosome is 12. We have the following information for the initial population:

Table 2. Initial population information

S No.	X	Fitness value (F)
1	010101010001	631
2	000101010100	679
3	000101000101	670
4	010100000101	648

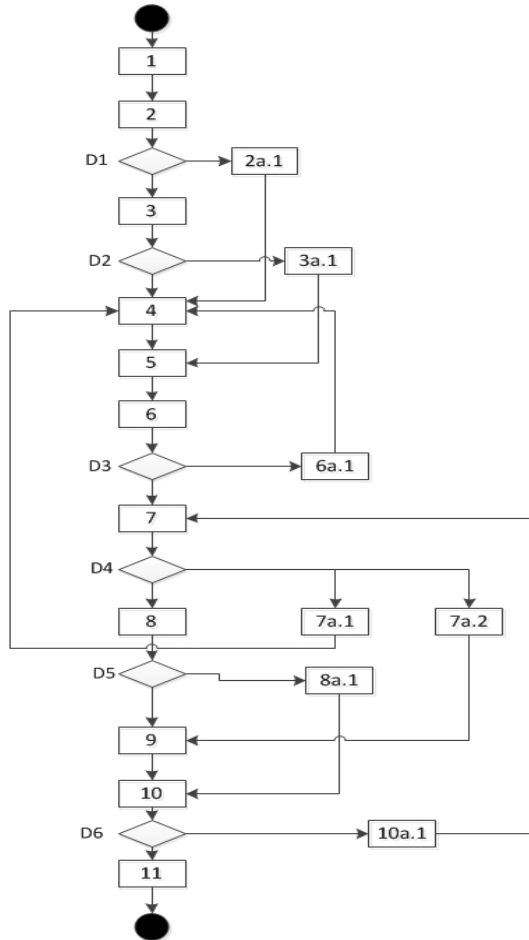


Fig. 2. Illustrative example for algorithm in [1]

Then, we have the following stack-based-weight assignment to all nodes in the illustrative graph.

The pop-operation-based complexity is obtained from the stack constructed by graph traversal. It is calculated by relying on the number of occurrences of a node in the stack.

Let us take the first individual (010101010001) as an example to calculate the fitness value. Each 2 bits represent a decision node. In this example, “00” means it will follow the first branch and “01”, “10” and “11” continue with remaining branches including basic path. Hence, the corresponding path of the individual is 1 → 2 → D1 → 3 → D2 → 4 → 5 → 6 → D3 → 7 → D4 → 7a.2 → 9 → 10 → D6 → 11 → End .

Therefore, the fitness value for this path is the total complexity of each node present in it: $96 + 91 + 86 + 64 + 57 + 51 + 47 + 34 + 31 + 13 + 12 + 10 + 8 + 3 + 14 + 14 = 631$.

Table 3. Complexity of nodes of activity diagram in figure 2

Node	Complexity based on pop operation (I)	IF = Fan In (A) * Fan Out (A) (II)	Total Complexity = (I) + (II)
1	96	0	96
2	90	1	91
D1	84	2	86
2a.1	65	1	66
3	13	1	14
D2	12	2	14
3a.1	11	1	12
4	60	4	64
5	55	2	57
6	50	1	51
D3	45	2	47
6a.1	8	1	9
7	32	2	34
D4	28	3	31
7a.1	6	1	7
7a.2	12	1	13
8	6	1	7
D5	5	2	7
8a.1	4	1	5
9	10	2	12
10	8	2	10
D6	6	2	8
10a.1	2	1	3
11	2	1	3

Similarly, we compute the fitness value for remaining chromosomes. Two individuals with highest fitness value will be selected as parents, which are X_2 and X_3 . After 16th iteration, the stopping condition is satisfied. So the algorithm stopped with the best chromosome 000101010000 and its fitness value is 679. It means that the “best” test scenario for this activity diagram is $1 \rightarrow 2 \rightarrow D1 \rightarrow 2a.1 \rightarrow 4 \rightarrow 5 \rightarrow 6 \rightarrow D3 \rightarrow 7 \rightarrow D4 \rightarrow 7a.2 \rightarrow 9 \rightarrow 10 \rightarrow D6 \rightarrow 10a.1 \rightarrow 7 \rightarrow D4 \rightarrow 8 \rightarrow D5 \rightarrow 8a.1 \rightarrow 10 \rightarrow D6 \rightarrow 11 \rightarrow \text{End}$. In other words, if the program has bug, it is most likely that the bug will occur on this path.

4 Enhancing the Genetic Algorithm with Cumulative Probability

The major problem with aforementioned approach is that it only focuses on individual with highest fitness value, which results in the lack of diversity in the population. In other hand, the overriding factor to make the population more diversified is the mutation. However, the probability for this operation is low, which is less than 0.2. So, it takes a long iteration until all paths have been covered.

This leads us to use cumulative probability to get around the mentioned issue [5]. By applying this technique, we can allow an individual with low fitness value to have a chance to participate in the mating process. Consequently, the population will increase its diversity and the iteration time therefore will be significantly decreased.

After all fitness values are calculated for each individual, the probability of selection p_j for each path j is determined as follows:

$$p_j = \frac{F_j}{\sum_{j=1}^n F_j}$$

where n is the number of individuals in the population and F_j is the fitness value for the individual j .

Then cumulative probability c_k is calculated for each path k with following equation:

$$c_k = \sum_{j=1}^k p_j$$

Hence, we have modified the algorithm in [1] with our enhanced one.

Algorithm 2. Genetic algorithm with cumulative probability

```

1: Set BackInit_flag = false
2: repeat
3:   Initializing population
4:   Evaluating population
5:   while Fitness(new_population) > Fitness(old_population)
6:     if random  $r_1 < 0.8$  then
7:       select parents
8:       Calculate  $p_j$  for each individual as in (4)
9:       Calculate  $c_k$  as in (5)
10:      Choosing candidates for mating based on  $c_k$  value and
corresponding random number
11:      if mating pool contains 1 candidate then
12:         $BackInit\_flag = true$ 
13:        break
14:      Performs crossover
15:      foreach individual in population do
16:        if random  $r_2 < 0.2$  then
17:          performs mutation
18:      until  $BackInit\ flag = true$ 

```

Let us illustrate our enhanced algorithm on the activity diagram in Figure 2 again.

Table 4. Population information for enhanced algorithm

Chromosome (X)	Fitness value	Probability (p)	Cumulative probability (c)	Random (r)
010101010001	631	0.2453	0.2453	0.4998
000101010100	679	0.2616	0.5069	0.9669
000101000101	670	0.2612	0.7681	0.0105
010100000101	648	0.2319	1	0.4484

Unlike the original algorithm, we do not choose X_2 and X_3 as parents. We have not decided parents when computing fitness value yet. Actually, we continue to compute the probability value for each individual. Then, cumulative probability is calculated as in (5), for instant, $c_1 = p_1 = 0.2453$; $c_2 = p_1 + p_2 = 0.5069$ and so on.

Next, we decide which individual will be placed in mating pool based on cumulative probability value and the corresponding number for each chromosome. Let us start with X_1 first whose random number r_1 is $0.4998 > c_1$ (0.2453). We advance to the next cumulative probability until it exceeded r_1 . Hence, we stopped at c_2 since its value is larger than c_1 . So, X_2 will be selected and placed in mating pool. Similarly, X_4 , X_1 and X_2 will be selected too in the next 3 iterations.

After the selection phase, we have 2 X_2 , 1 X_1 and 1 X_4 in the mating pool. Among themselves, X_1 and X_2 have the largest fitness value, so they are chosen as parents for mating process. We continue the algorithm until the stopping condition is satisfied which is after 3th iteration. The survival individual is 000101010000, which is the same as the original algorithm but the number of iteration is far smaller.

5 Experimental Results

We conducted the experiment on various test cases collected from a practical project developed at LARION Computing¹, a professional software company. Each test case was run about 20 times and we carried out on nearly 50 test cases with diverse kinds: diagrams with no backward loop, diagram with backward loops, and diagram with outer and inner loop. We also tested on different decision nodes with various outgoing paths.

In addition, we would also like to demonstrate the efficiency of our algorithm in comparison with the exhaustive search approach and the original approach as in [1]. They are tested with the same initial population and fitness function. We also set the threshold for those (maximum number of iteration is 50000) in order to prevent endless running time.

Furthermore, we also conducted the experiment with two stopping conditions:

- *First stopping condition*: similar to that of in [1] – when all individuals have been covered and the new population is better than the old one.
- *Second stopping condition*: following the standard stopping condition in GA – when the improvement between the new population and the old one is not that much under a given threshold.

Experimental results have shown that the outcome of our enhanced algorithm is better than the original one at about 65% under first stopping condition. Whereas under second stopping condition, our enhanced algorithm produced better outcome in comparison with the original one at about 80%.

We would like to take a result of a test case out of all test cases as a representative. In doing so, we will have a clear observation so as to compare effectiveness. We also obtain similar results with remaining test cases under this stopping condition.

Table 5 presents the performance comparison of these 3 approaches in terms of execution time (which is number of iteration) and the value of the survival individual in 15 consecutive runs.

Table 5. Performance comparison between 3 approaches under first stopping condition

#	Brute force		Original algorithm		Enhanced algorithm	
	<i>fitness</i>	<i>execution time</i>	<i>Fitness</i>	<i>execution time</i>	<i>fitness</i>	<i>execution time</i>
1	2156	9866	1711	12	1750	6
2	2156	9866	1750	2	1912	3
3	2156	9866	1977	9	1982	11
4	2156	9866	2156	10	2140	6
5	2156	9866	1918	50000	1918	26
6	2156	9866	1711	9	1984	13
7	2156	9866	1987	21	1988	3
8	2156	9866	2141	50000	1750	41
9	2156	9866	2145	34	2156	50
10	2156	9866	2155	83	2150	20
11	2156	9866	2156	76	1983	71
12	2156	9866	2141	5	2151	5
13	2156	9866	1711	13	1711	5
14	2156	9866	1750	3	2140	16
15	2156	9866	2130	40	2156	28

There are 10 out of 15 cases in which the enhanced algorithm produces better result than the original one. There are 2 times when the execution time of the original algorithm reached the climax. Additionally, when we increased the size of the test case, the exhaustive search becomes unfeasible, since it always reached the given threshold.

The above result is illustrated by the following line graph:

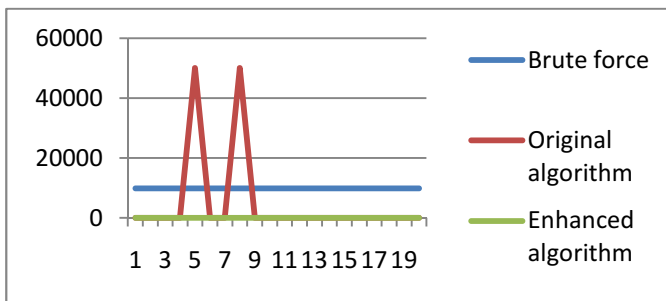


Fig. 3. Performance comparison of execution time under first stopping condition

Table 6 presents the result of a representative under the second stopping condition.

There are 12 out of 15 cases in which enhanced algorithm produces better result than the original one. It can be seen clearly in the table, the original algorithm incurs a large number of iterations when compared with our algorithm.

Table 6. Performance comparison between 3 approaches under second stopping condition

#	Brute force		Original algorithm		Enhanced algorithm	
	<i>fitness</i>	<i>execution time</i>	<i>fitness</i>	<i>execution time</i>	<i>fitness</i>	<i>execution time</i>
1	1430	12134	1252	1464	1353	98
2	1430	12134	1236	4724	1236	32
3	1430	12134	1305	5	1330	12
4	1430	12134	1332	26	1354	8
5	1430	12134	1338	5	1332	13
6	1430	12134	1353	50000	1228	120
7	1430	12134	1306	36	1354	9
8	1430	12134	1330	1173	1354	74
9	1430	12134	1332	26	1355	8
10	1430	12134	1338	7	1332	13
11	1430	12134	1340	71	1430	119
12	1430	12134	1309	17	1354	2304
13	1430	12134	1233	23529	1428	71
14	1430	12134	1407	5	1407	21
15	1430	12134	1354	11	1410	37

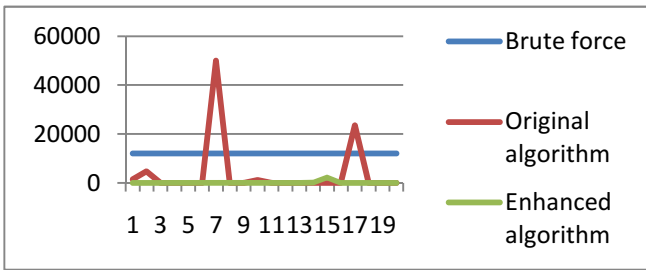


Fig. 4. Performance comparison of execution time under second stopping condition

As aforesaid elucidation, the original GA allows only individuals with highest fitness value to participate in the mating process, which results in the lack of diversity in the population. Consequently, it usually calls for long-winded iteration so as to terminate the algorithm.

6 Conclusions and Future Work

This paper has proposed a GA-based approach with cumulative probability enhancement for deriving critical test scenario from use-cases. We have tested our improved algorithm in comparison with the brute force approach and other previous GA-based work [1]. Initial experimental results have shown that our proposed approach mostly produce better outcome in terms of execution time and survival individual.

In view of supporting tool, we have developed a tool allowing the user to input the use-cases and actors via GUI-based interaction. Our future work involves taking account of related use-cases in the Extend and Include relationship when analyzing a use-case.

References

1. Sabharwal, S., Sibal, R., Sharma, C.: Applying Genetic Algorithm for Prioritization of Test Case Scenarios Derived from UML Diagrams. *IJCSI International Journal of Computer Science* 8(3(2)) (May 2011)
2. Heumann, J.: Generating Test Cases from Use Cases. *Journal of Software Testing Professionals* (2002)
3. Chen, M., Qiu, X., Li, X.: Automatic test case generation for UML activity diagrams. In: *Factivity Diagram (AST 2006)*, pp. 2–8. ACM, New York (2006)
4. Srivastava, P.R., Kim, T.-H.: Application of Genetic Algorithm in Software Testing. *International Journal of Software Engineering and Its Applications* 3(4) (October 2009)
5. Hetzel, B.: *The Complete Guide to Software Testing*, 2nd edn. (1993)
6. Hutcheson, M.L.: *Software Testing Fundamentals-Methods and Metrics*. Wiley Publishing (2003)
7. Zielczynski, P.: Traceability from Use Cases to Test Cases. IBM research report (2006), <http://www.ibm.com/developerworks/rational/library/04/r-3217/>
8. Kundu, D., Samanta, D.: A Novel Approach to Generate Test Cases from UML Activity Diagrams. *Journal of Object Technology* 8(3) (May-June 2009)
9. Ribeiro, J.C.B., Rela, M.Z., de Vega, F.F.: A Strategy for Evaluating Feasible and Unfeasible Test Cases for the Evolutionary Testing of Object-oriented Software. *ACM* (2008)

Towards a Semantic Linked Data Retrieval Model

Van Bich Nguyen and Dang Tuan Nguyen

Faculty of Computer Science
University of Information Technology, VNU-HCM
Ho Chi Minh City, Vietnam
{vannb, dangnt}@uit.edu.vn

Abstract. This research aims to introduce our Semantic Linked Data Retrieval Model (SLDRM) which is conceived to perform the traditional Ad-hoc Retrieval task of INEX 2013 Linked Data Track. SLDRM was built based on combining some novel semantic representation models for processing English retrieval queries, and proposed techniques for creating and optimizing queries on Indri search engine to effectively retrieve structured text data. Based on SLDRM, a Semantic Linked Data Retrieval System (SLDRS) was built to assess the performance of SLDRM. The evaluation methods of the INEX 2013 Linked Data Track, based on using standard Ad-hoc search task's testing topics and two standard scores MRR and TREC MAiP, were used to evaluate the accuracy of SLDRS. The experiments show that MAiP score of our SLDRS is the highest in comparison with all of submitted runs of INEX 2013 Linked Data Track.

Keywords: Structured Text Retrieval, Linked Data, Search Engine, Query Language, Semantic Model.

1 Introduction

In this paper we present an approach to perform the traditional Ad-hoc Retrieval task of INEX 2013 Linked Data Track [1]. The Ad-hoc Retrieval task of INEX 2013 Linked Data Track [1] aims to study effective methods to retrieve standard structured text data, which compose the data of Wikipedia [2], a partial data of DBpedia [1] (see [3]), and YAGO2 [1] (see [4]). Our research focused on proposing a semantic linked data retrieval model which can enhance our experimental system in comparison with all other systems submitted to INEX 2013 Linked Data Track [1].

To build a semantic linked data retrieval model, which can answer testing "topics" (this terminology is suggested by INEX 2013 Linked Data Track [1] to refer to English "query"), our approach bases on following steps:

- Step 1: Use the Stanford Parser [5] to parse the topic and receive its dependency relations of the Stanford Dependencies [6].
- Step 2: Based on the dependency relations of the topic, analyze these semantic relations with our specific semantic representation models to construct the semantic representation of the topic.
- Step 3: Use IndriBuildIndex [7] tool of Indri [9] search engine to index the semantic linked data of Wikipedia [2].

- Step 4: Use the semantic representation of the topic in step 2 to create its search engine query following the syntax of Indri Query Language [8].
- Step 5: Optimize the Indri [9] search engine's query created in step 4.
- Step 6: Execute the optimized query in step 5 on Indri [9] search engine.

2 Architecture of Semantic Linked Data Retrieval Model

The architecture of our Semantic Linked Data Retrieval Model (SLDRM) is designed to perform six operational steps mentioned above. Its components are:

- Syntactic Parser: parses the dependency syntax of the retrieval query. This parser performs the step 1 of the model.
- Semantic Constructor: constructs the semantic representation of the retrieval query. This component performs the step 2 of the model.
- Linked Data Indexer: indexes the linked data which will be retrieve on Indri search engine. As mentioned above, we use IndriBuildIndex [7] tool of Indri [9] search engine to index the linked data.
- Indri Query Creator: creates query on Indri Query Language [8]. This component performs the step 4 of the model.
- Query Optimizer: optimizes the query on Indri Query Language [8]. This component performs the step 5 of the model.
- Retrieving and Ranking: retrieves query on Indri [9] search engine and ranking the found data. This component performs the step 6 of the model. (In this paper we do not discuss this component.)

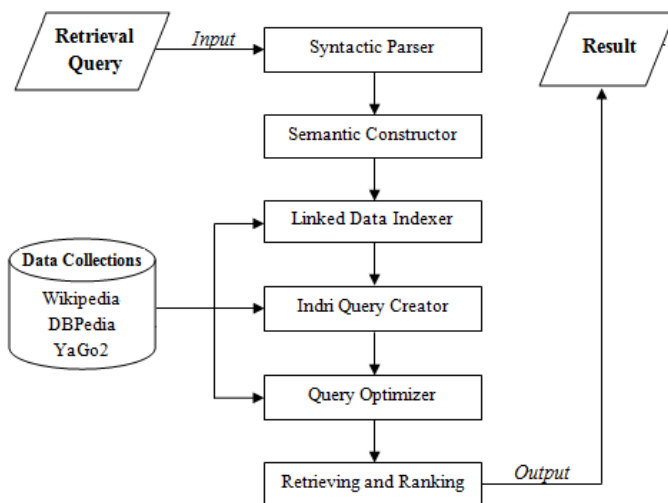


Fig. 1. Architecture of SLDRM

3 Syntactic Parser

The syntactic parsing performs the first stage in the process of analyzing English retrieval query to determine their grammatical relations. SLDRM uses Stanford Parser [5] to parse the English retrieval query and to return its Stanford Dependencies [6] relationships. A Stanford Dependencies [6] relationship consists of following parts: 1) *relation* (an abbreviation of relationship name), 2) *gov* (abbreviation of “Governor”), and 3) *dep* (abbreviation of “Dependency”). In addition, each *Gov* and *Dep* part has a position index number attached after it.

Example 1: “*famous computer scientists disappeared at sea*”

Stanford Parser parses the query in example 1 into dependency relations as presented in Table 1.

Table 1. Parsed query of example 1

Relation	Gov	Dep
amod	scientists-3	famous-1
nn	scientists-3	computer-2
nsubj	disappeared-4	scientists-3
prep	disappeared-4	at-5
pobj	at-5	sea-6

4 Semantic Constructor

The Semantic Constructor is built based on definition of semantic components, and semantic models which use these semantic components to form semantic representations of English retrieval query (cf. [18], [19]).

4.1 Components of Semantic Models

The proposed semantic models for English retrieval queries in SLDRM use the following semantic components:

- 1) *SLDRM_Main_Object*: represents object, thing, or event to be retrieved in an English question.
- 2) *SLDRM_Identify_Object*: provides supplementary information for *SLDRM_Main_Object* component. This semantic component includes two sub-components: *SLDRM_Circumstances* and *SLDRM_Active_Modifier*.
 - *SLDRM_Circumstances*: describes the context.
 - *SLDRM_Active_Modifier*: includes the following elements:
 - *SLDRM_Verb*: describes the action.
 - *SLDRM_Time*: describes the time.

- *SLDRM_Poss*: describes the possession.
- *SLDRM_Modifier*: describes the modification.

Example 2: “*probabilistic models in information retrieval*”

→ *SLDRM_Main_Object* = “*probabilistic models*”, *SLDRM_Circumstances* = “*information retrieval*”.

Example 3: “*July, 1850 president died Millard Fillmore sworn following day*”

→ *SLDRM_Verb* = {*died, sworn*}, *SLDRM_Time* = {*July, 1850*}.

4.2 Semantic Models of English Retrieval Queries

SLDRM has three semantic models which are used to represent the semantics of retrieval queries.

- Semantic model of type 1: [*SLDRM_Main_Object*]

Example 4: “*yoga exercise*”, “*IBM computer*”, “*Paul Auster*”, “*Einstein Relativity theory*”, “*best movie*”, “*greatest guitarist*”, “*French revolution*”,

- Semantic model of type 2:

[*SLDRM_Main_Object*] + [*SLDRM_Circumstances*]

Example 5: “*Singer in Britain's Got Talent*”, “*probabilistic models in information retrieval*”.

→ *SLDRM_Main_Object* = “*Singer*”, *SLDRM_Circumstances* = “*Britain's Got Talent*”

→ *SLDRM_Main_Object* = “*probabilistic models*”, *SLDRM_Circumstances* = “*information retrieval*”

- Semantic model of type 3:

[*SLDRM_Main_Object*] + [*SLDRM_Active_Modifier*]

SLDRM_Active_Modifier = {*SLDRM_Verb, SLDRM_Time, SLDRM_Poss, SLDRM_Modifier*}

Example 6: “*1994 short story collection Alice Munro is Open*”.

→ *SLDRM_Main_Object* = “*short story collection*”,

→ *SLDRM_Active_Modifier* = {*1994, Alice Munro, is Open*}

5 Construction of Query on Indri Search Engine

For each semantic model of English query, SLDRM has a respective query structure form which is defined to automatically build queries in Indri Query Language [8].

- Indri query of form 1: [*SLDRM_Main_Object*]

The form of query in Indri Query Language [8]:

```
#weight[passage800:500] (3.0 #3 (#3 (title
SLDRM_Main_Object) {tobe}) 3.0
```

```
#3(SLDRM_Main_Object).paragraph 3.0
#3({SLDRM_Main_Object}).template 3.0
#3({SLDRM_Main_Object}).tag
```

Example 7: “*yoga exercise*”

The generated query of example 7 in Indri Query Language [8]:

```
#weight[passage800:500](3.0 #3(#3(title yoga exer-
cise) {is are was were}) 3.0 #3(yoga exer-
cise).paragraph 3.0 #3({yoga exercise}).template
3.0 #3({yoga exercise}).tag)
```

- Indri query of form 2: [SLDRM_Main_Object] + [SLDRM_Circumstances]

The form of query in Indri Query Language [8]:

```
#weight[passage800:500](3.0 #3(SLDRM_Main_Object
SLDRM_Circumstances).paragraph 3.0 #3(title
SLDRM_Main_Object SLDRM_Circumstances) 3.0 #3(title
SLDRM_Circumstances) 3.0
#3(SLDRM_Circumstances).arg 3.0
#3({SLDRM_Main_Object SLDRM_Circumstances}).tag 3.0
#3({SLDRM_Main_Object SLDRM_Circumstances}).arg)
```

Example 8: “*electricity source in France*”

The generated query of example 8 in Indri Query Language [8]:

```
#weight[passage800:500](3.0 #3(electricity source
in France).paragraph 3.0 #3(title electricity
source in France) 3.0 #3(title France) 3.0
#3(France).arg 3.0 #3({electricity source in
France}).tag 3.0 #3({electricity source in
France}).arg)
```

- Indri query of form 3: [SLDRM_Main_Object] + [SLDRM_Active_Modifier]

The form of query in Indri Query Language [8]:

```
#weight[passage800:500](3.0 #3({SLDRM_Main_Object
SLDRM_Active_Modifier}).paragraph 3.0
#3(SLDRM_Verb).paragraph 3.0
#3(SLDRM_Time).Paragraph 3.0
#3(SLDRM_Poss).paragraph 3.0
#3(SLDRM_Modifier).paragraph 3.0
#3({SLDRM_Main_Object SLDRM_Active_Modifier}).tag
3.0 #3({SLDRM_Main_Object
SLDRM_Active_Modifier}).arg)
```

Example 9: “*countries make up Central America*”

The generated query in Indri Query Language [8]:

```
#weight[passage800:500](3.0 #3({countries make up
Central America}).paragraph 3.0
#3(countries).paragraph 3.0 #3({make up}).paragraph
3.0 #3(Central America).paragraph 3.0 #3({countries
make up Central America}).tag 3.0 #3({countries
make up Central America}).arg)
```

6 Optimizing Indri Search Engine's Queries

In the dataset of INEX 2013 Linked Data Track [2], each Wikipedia article is linked to DBpedia and YAGO2 datasets. Our improving method is to exploit the information of these links, and to modify the structure and parameters of queries on Indri [9] search engine.

6.1 Optimizing Technique 1

The data DBpedia_Page_IDs of DBpedia [12] contains the titles of Wikipedia's articles. The idea of our optimizing technique is to check the the title of a Wikipedia's article in the searching information. If so, add "3.0 # 3 (title Retrieval_Query)" in the query. Otherwise, add "3.0 # 3 ({Retrieval_Query}).paragraph" in the query.

Example 10: "John Turturro 1991 Coen Brothers film"

The original query in Indri Query Language [8]:

```
#weight[passage800:500] (3.0 #3({John Turturro 1991
Coen Brothers film }).paragraph 3.0 #3(John
Turturro Coen Brothers film).paragraph 3.0
#3(1991).paragraph 3.0 #3({John Turturro 1991 Coen
Brothers film }).tag 3.0 #3({John Turturro 1991
Coen Brothers film}).arg)
```

Use optimizing technique 1:

"John Turturro" and "Coen Brothers" are titles of Wikipedia articles, then we should add "3.0 # 3 (title John Turturro)" and "3.0 # 3 (title Coen Brothers)" in the query.

The optimized query in Indri Query Language [8]:

```
#weight[passage800:500] (3.0 #3(title John
Turturro) 3.0 #3(title Coen Brothers) 3.0 #3({John
Turturro 1991 Coen Brothers film }).paragraph 3.0
#3(John Turturro Coen Brothers film).paragraph 3.0
#3(1991).paragraph 3.0 #3({John Turturro 1991 Coen
Brothers film }).tag 3.0 #3({John Turturro 1991
Coen Brothers film}).arg)
```

6.2 Optimizing Technique 2

The knowledge data of YAGO2 [2] is organized in several groups having the same meaning. Therefore, our model will checks YagoFacts [13], YagoLiteralFacts [14] and YagoTypes [15] of YAGO2 [2] data. If they contains the information that links with SLDRM_Main_Object then we add "3.0 # 3 (Information_Yago {})." Paragraph" in the query.

Example 11: "Joseph Luns"

Data in YagoFacts [13]:

```
<Joseph_Luns> <wasBornIn> <Rotterdam> .
```

Data in YagoLiteralFacts [14]:

```
<Joseph_Luns> <diedOnDate> "2002-07-17"^^xsd:date.
```

Data in YagoTypes [15]:

```
<wikicategory_Dutch_people_of_World_War_II>.
<wikicategory_Ministers_of_Foreign_Affairs_of_the_Netherlands>.
<wikicategory_Secretaries_General_of_NATO>.
<wikicategory_Dutch_civil_servants>.
<wikicategory_Roman-Catholic_State_Party_politicians>.
<wikicategory_Alumni_of_the_London_School_of_Economics>.
<wikicategory_People_from_Rotterdam>.
<wikicategory_Dutch_diplomats>.
<wikicategory_Catholic_People's_Party_politicians>.
<wordnet_officeholder_110371450>.
<wikicategory_Leiden_University_alumni>.
```

In YAGO2 [2] data, “*Joseph Luns*” will contains information of events “*{wasBornIn, diedOnDate}*” and links to information of events: “*{Dutch people of World War II, Ministers of Foreign Affairs of the Netherlands, Secretaries General of NATO, Dutch civil servants, Roman-Catholic State Party politicians, Alumni of the London School of Economics, People from Rotterdam, Dutch diplomats, Catholic People’s Party politicians, Leiden University alumni}*”.

7 Experiments

Based on our SLDRM, we tried to build a Semantic Linked Data Retrieval System (SLDRS) by using IndriCsharp.dll library and several supported indexing files which are published and provided by Indri [9] search engine: IndriBuildIndex.exe, IndexFile.xml, StructuresWiki.xml for indexing linked data and querying indexed data.

7.1 Data Collections and Testing Topics

SLDRS is conceived and designed to work on the standard dataset of INEX Linked Data Track [2]. It includes data of Wikipedia (version 2.0) [2], and a subset of Dbpedia (version 3.8) [16], YAGO2s [17].

For testing, SLDRS uses the “144 Ad-hoc search task topics” [11] provided by INEX 2013 Linked Data Track [1]. These testing topics are related to several domains.

7.2 Indexing and Querying Semantic Linked Data

To index the data, SLDRS use the following supported tools and files of the Indri [9] search engine: IndriBuildIndex.exe, IndexFile.xml, and StructuresWiki.xml. SLDRS uses IndriCsharp.dll library supported by Indri [9] search engine to query the indexed data.

7.3 Evaluation

We use the evaluation methods of the INEX 2013 Linked Data Track [1], which are based on using “144 Ad-hoc search task topics” [11] and two principal standard scores MRR and TREC MAiP, to assess the performance of the system. All of the P@5, P@10, P@20, P@30 are secondary scores used to consider the performance of the system. Precision-at-k (P@k) is the portion of the relevant documents in the first k ranks. Table 2 presents the experimental results of our SLDRS.

Table 2. The experimental results of SLDRS

MRR	MAiP	P@5	P@10	P@20	P@30
0.7814	0.43269	0.454	0.3459	0.259	0.19827

Table 3 presents the evaluation results of all the 8 submitted runs (systems) of the INEX 2013 Linked Data Track [1].

Table 3. Results of the INEX 2013 Linked Data Track (Source: S. Gurajada et al. [10])

Score	Run_1	Run_2	Run_3	Run_4	Run_5	Run_6	Run_7	Run_8
MRR	0.8772	0.7957	0.8861	0.7922	0.7449	0.8684	0.8888	0.8786
MAiP	0.3733	0.2408	0.388	0.2577	0.2112	0.1489	0.1677	0.1629
P@5	0.7028	0.5958	0.725	0.5986	0.5458	0.4444	0.4689	0.4689
P@10	0.6424	0.5229	0.6674	0.5403	0.509	0.3204	0.3459	0.3443
P@20	0.5979	0.4549	0.6174	0.4903	0.4618	0.2407	0.2615	0.2607
P@30	0.5544	0.4134	0.5646	0.4426	0.4127	0.171	0.1891	0.1896

Notes:

- Table 3 composes the evaluation results published by S. Gurajada et al. [10] in INEX 2013 Linked Data Track [1], but they are re-organized for our purpose of presentation.
- Name of runs: ruc-all-2200 (Run_1), ruc-all-2200-rerank (Run_2), ruc-all-2200-paragraph-80 (Run_3), ruc-all-2200-paragraph-80-rerank (Run_4), OaucLD1 (Run_5), MPISupremacy (Run_6), MPIUltimatum_Phrases (Run_7), MPIUltimatum_NoPhrase (Run_8).
- In total, 4 Ad-hoc search runs (Run_1, Run_2, Run_3, Run_4) were submitted by Renmin University of China (RUC), 1 Ad-hoc search run (Run_5) was submitted by Oslo and Akershus University College of Applied Sciences (OAUC), 3 Ad-hoc search runs (Run_6, Run_7, Run_8) were submitted by the Max-Planck Institute for Informatics (MPI).

Comparing MAiP scores of our SLDRS in Table 2 with all MAiP scores of submitted runs of INEX 2013 Linked Data Track [1] in Table 3, we observe that our MAiP score 0.43269 is the highest.

8 Conclusion

This paper presents the architecture and the building method of the Semantic Linked Data Retrieval Model (SLDRM) which was conceived to enhance the retrieval performance of our experiments. The approach of SLDRM is founded on two main methods: 1) proposing semantic representation models to process the meaning of English topics and, 2) optimizing the query on Indri Query Language [8].

Based on SLDRM, we built an experimental Semantic Linked Data Retrieval System (SLDRS) to assess the proposed retrieval model on semantic linked data. The experiments demonstrate that the MAiP score of our SLDRS is the highest in comparison with all MAiP scores of submitted runs of INEX 2013 Linked Data Track [1].

However, in future works, we will study to improve the MRR score of SLDRS.

References

1. INEX 2013 Linked Data Track (2013),
<https://inex.mmci.uni-saarland.de/tracks/lod/2013/>
2. INEX Wikipedia LOD Collection, Version 2.0,
<http://inex-lod.mpi-inf.mpg.de/2013/>
3. DBpedia, <http://dbpedia.org/About>
4. YAGO2s, A.: High-Quality Knowledge Base,
<http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/>
5. The Stanford Parser: A statistical parser,
<http://nlp.stanford.edu/software/lex-parser.shtml>
6. http://nlp.stanford.edu/software/dependencies_manual.pdf
7. <http://www.lemurproject.org/lemur/indexing.php#IndriBuildIndex>
8. Indri Query Language Quick Reference,
<http://www.lemurproject.org/lemur/IndriQueryLanguage.php>
9. INDRI: Language modeling meets inference networks,
<http://www.lemurproject.org/indri/>
10. Gurajada, S., Kamps, J., Mishra, A., Schenkel, R., Theobald, M., Wang, Q.: Overview of the INEX 2013 Linked Data Track. In: CLEF 2013 Evaluation Labs and Workshop, Online Working Notes (2013)
11. <https://inex.mmci.uni-saarland.de/protected/dc/2013-ld-adhoc-topics.xml>
12. http://downloads.dbpedia.org/3.8/en/page_ids_en.nt.bz2
13. <http://resources.mpi-inf.mpg.de/yago-naga/yago/download/yago/yagoFacts.ttl.7z>
14. <http://resources.mpi-inf.mpg.de/yago-naga/yago/download/yago/yagoLiteralFacts.ttl.7z>
15. <http://resources.mpi-inf.mpg.de/yago-naga/yago/download/yago/yagoTypes.ttl.7z>
16. <http://downloads.dbpedia.org/3.8/en/>

17. YAGO2s: Downloads,
<http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/downloads/>
18. Nguyen, D.T., Tran, P.N., Nguyen, V.B.: Transforming Semantic Models of WHAT-Questions into Indri Query Language on IndriBasedQA System. In: The 2010 3rd International Conference on Computer and Electrical Engineering (ICCEE 2010), Chengdu, China, vol. 1, pp. 455–458 (2010)
19. Nguyen, D.T., Nguyen, V.B., Tran, P.N.: Semantic Representation of WHO-Questions on IndriBasedQA System. In: Proceedings of the 2010 2nd International Conference on Software Technology and Engineering (ICSTE 2010), San Juan, Puerto Rico, USA, vol. 2, pp. 175–179 (2010)

Author Index

- Antunes, Pedro 136, 164
van Banerveld, Maarten 29
Cortesi, Agostino 262
Dang, Tran Khanh 44, 86, 204
Dang, Tran Tri 86
Dien, Nguyen Kim 58
Do, Nghi T. 16
Duong, Anh-Duc 177, 220
Halder, Raju 262
Hieu, Tran Trung 58
Ichise, Ryutaro 151
Johnstone, David 136, 164
Kechadi, M-Tahar 29
Kim, Howon 191
Kosorus, Hilda 1
Küng, Josef 1, 44
Le, Bac 151
Le, Duy-Khanh 248
Le, Thuan D. 286
Le, Tuan Dinh 97
Le, Van T. 16
Le-Khac, Nhien-An 29
Mai, Khang Trong 97
Manan, Jamalul-Lail Ab 122
Misra, Sanjay 234
Mubarak, Mohd Faizal 122
Nguyen, Dang Tuan 300
Nguyen, Ha Van 191
Nguyen, Minh Tan 204
Nguyen, Thanh D. 234
Nguyen, Thanh T.T. 234
Nguyen, Toan Tan 97
Nguyen, Van Bich 300
Nguyen, Vu Thanh 97
Ong, Hoang 107
Pham, Khang N. 16
Phan, Duong-Tien 177
Phan, Trong Nhan 44
Quan, Tho T. 286
Quang-Hung, Nguyen 248
Regner, Peter 1
Seo, Hwajeong 191
Shao, Jianhua 107
Son, Nguyen Thanh 248
Tang, Yi 71
Temiyasathit, Chivalai 275
Thang, Le Quoc 275
Thi, Danh Bui 151
Thinh, Tran Ngoc 58
Thoai, Nam 248
Thuan, Nguyen Hoang 136, 164
Tran, An T. 286
Tran, Minh-Triet 177, 220
Tran-Nguyen, Thu M. 16
Truong, Minh Nhat Quang 164
Truong, Quang Hai 204
Truong, Toan-Thinh 177, 220
Yahya, Saadiah 122
Zhang, Ji 71
Zhang, Xiaolei 71