Hsinchun Chen Michael Chau
Shu-hsing Li Shalini Urs
Srinath Srinivasa G. Alan Wang (Eds.)

LNCS 6122

Intelligence and Security Informatics

Pacific Asia Workshop, PAISI 2010
Hyderabad, India, June 2010
Proceedings

Springer

Lecture Notes in Computer Science 6122

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

Alfred Kobsa
University of California, Irvine, CA, USA

Friedemann Mattern
ETH Zurich, Switzerland

John C. Mitchell
Stanford University, CA, USA

Moni Naor
Weizmann Institute of Science, Rehovot, Israel

Oscar Nierstrasz
University of Bern, Switzerland

C. Pandu Rangan
Indian Institute of Technology, Madras, India

Bernhard Steffen
TU Dortmund University, Germany

Madhu Sudan
Microsoft Research, Cambridge, MA, USA

Demetri Terzopoulos
University of California, Los Angeles, CA, USA

Doug Tygar
University of California, Berkeley, CA, USA

Gerhard Weikum
Max-Planck Institute of Computer Science, Saarbruecken, Germany

Hsinchun Chen Michael Chau
Shu-hsing Li Shalini Urs
Srinath Srinivasa G. Alan Wang (Eds.)

Intelligence and Security Informatics

Pacific Asia Workshop, PAISI 2010
Hyderabad, India, June 21, 2010
Proceedings

Springer

Volume Editors

Hsinchun Chen
The University of Arizona, Tucson, AZ, USA
E-mail: hchen@eller.arizona.edu

Michael Chau
The University of Hong Kong, Hong Kong, China
E-mail: mchau@business.hku.hk

Shu-hsing Li
National Taiwan University, Taipei, Taiwan, R.O.C.
E-mail: shli@management.ntu.edu.tw

Shalini Urs
University of Mysore, Mysore, India
E-mail: shalini@isim.ac.in

Srinath Srinivasa
International Institute of Information Technology
Bangalore, India
E-mail: sri@iiitb.ac.in

G. Alan Wang
Virginia Tech, Blacksburg, VA, USA
E-mail: alanwang@vt.edu

Library of Congress Control Number: 2010928273

CR Subject Classification (1998): K.6.5, D.4.6, H.4, H.2.8, C.2, J.1

LNCS Sublibrary: SL 4 – Security and Cryptology

ISSN 0302-9743
ISBN-10 3-642-13600-1 Springer Berlin Heidelberg New York
ISBN-13 978-3-642-13600-9 Springer Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

springer.com

© Springer-Verlag Berlin Heidelberg 2010
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper 06/3180

Preface

Intelligence and security informatics (ISI) is concerned with the study of the development and use of advanced information technologies and systems for national, international, and societal security-related applications. The annual IEEE International Conference series on ISI (http://www.isiconference.org/) was started in 2003. In 2006, the Workshop on ISI (http://isi.se.cuhk.edu.hk/2006/) was held in Singapore in conjunction with the Pacific Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2006), with over 100 contributors and participants from all over the world. This would become the start of a new series of ISI meetings in the Pacific Asia region. PAISI 2007 (http://isi.se.cuhk.edu.hk/2007/) was then held in Chengdu, China. PAISI 2008 (http://isi.se.cuhk.edu.hk/2008/) was held in Taipei, Taiwan, in conjunction with IEEE ISI 2008. PAISI 2009 (http://www.business.hku.hk/paisi/2009/) was held in Bangkok, Thailand, in conjunction with PAKDD 2009. These past ISI conferences and workshops brought together academic researchers, law enforcement and intelligence experts, information technology consultants and practitioners to discuss their research and practice related to various ISI topics. These topics include ISI data management, data and text mining for ISI applications, terrorism informatics, deception and intent detection, terrorist and criminal social network analysis, public health and bio-security, crime analysis, cyber-infrastructure protection, transportation infrastructure security, policy studies and evaluation, information assurance, enterprise risk management, information systems security, among others. We continued this series of ISI workshops in this region by organizing the 2010 Pacific Asia Workshop on ISI (PAISI 2010) in conjunction with the Pacific Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2010) in Hyderabad, India. PAISI 2010 was hosted by the University of Arizona, the University of Hong Kong, and the University of Mysore, India. The one-day program included a keynote speech by Professor Hsinchun Chen of the University of Arizona, presentations of 9 long papers, 7 short papers, and 2 posters. We hope PAISI can continue to provide a stimulating forum for ISI researchers in Pacific Asia and other regions of the world to exchange ideas and report research progress. We wish to express our gratitude to all workshop Program Committee members, who provided valuable and constructive review comments.

June 2010

Hsinchun Chen
Michael Chau
Shu-hsing Li
Shalini R. Urs
Srinath Srinivasa
G. Alan Wang

Organization

Workshop Co-chairs

Hsinchun Chen	The University of Arizona, USA
Michael Chau	The University of Hong Kong
Shu-hsing Li	National Taiwan University
Shalini Urs	University of Mysore, India

Program Co-chairs

Srinath Srinivasa	International Institute of Information Tech. – Bangalore, India
G. Alan Wang	Virginia Tech, USA

Program Committee

Indranil Bose	The University of Hong Kong
Robert Chang	Central Police University, Taiwan
Kuo-Tay Chen	National Taiwan University
Patrick Chen	Tatung University, Taiwan
Tsai-Jyh Chen	National Chengchi University, Taiwan
Reynold Cheng	The University of Hong Kong
Uwe Glasser	Simon Fraser University, Canada
Raymond Hsieh	California University of Pennsylvania, USA
Eul Gyu Im	Hanyang University, Republic of Korea
Da-Yu Kao	Central Police University, Taiwan
Siddharth Kaza	Towson University, USA
Paul W.H. Kwan	The University of New England, Australia
Kai Pui Lam	The Chinese University of Hong Kong
Wai Lam	The Chinese University of Hong Kong
Sheau-Dong Lang	University of Central Florida, USA
Ickjai Lee	James Cook University, Australia
You-lu Liao	Central Police University, Taiwan
Ee-peng Lim	Nanyang Technological University, Singapore
Hongyan Liu	Tsinghua University, China
Hsin-min Lu	National Taiwan University
Anirban Majumdar	SAP Research, Germany
Byron Marshall	Oregon State University, USA
Robert Moskovitch	Ben Gurion University, Israel
Dorbin Ng	The Chinese University of Hong Kong
Shaojie Qiao	Southwest Jiaotong University, China
Jialun Qin	The University of Massachusetts Lowell, USA

Muttukrishnan Rajarajan	City University London, UK
Aixin Sun	Nanyang Technological University, Singapore
Shrisha Rao	International Institute of Information Technology – Bangalore, India
Dmitri Roussinov	University of Strathclyde, UK
Paul Thompson	Dartmouth College, USA
Jau-Hwang Wang	National Central Police University, Taiwan
Philips Wang	Caritas Institute of Higher Education, Hong Kong
Shiuh-Jeng Wang	Central Police University, Taiwan
Jennifer Xu	Bentley University, USA
Wei Yan	Huawei Symantec Technologies, China
Wei T. Yue	City University of Hong Kong
Yilu Zhou	George Washington University, USA
William Zhu	The University of Auckland, New Zealand

Table of Contents

Terrorism Informatics and Crime Analysis

Combined Detection Model for Criminal Network Detection 1
Fatih Ozgul, Zeki Erdem, Chris Bowerman, and Julian Bondy

An Association Model for Implicit Crime Link Analysis 15
Jau-Hwang Wang and Chien-Lung Lin

Design and Deployment of a National Detecting Stolen Vehicles Network System 22
Weiping Chang and Chingwei Su

Fighting Cybercrime: A KM Perspective 28
Weiping Chang

Transportation Infrastructure Security

Trajectory Similarity of Network Constrained Moving Objects and Applications to Traffic Security 31
Sajimon Abraham and Paulose Sojan Lal

An Emergent Traffic Messaging Service Using Wireless Technology 44
Patrick S. Chen, Yong-Kuei Liu, and Chia-Shih Hsu

Information Access and Security

A Model for Detecting “Global Footprint Anomalies” in a Grid Environment 52
Pramod S. Pawar and Srinath Srinivasa

Secure Anonymous Routing for MANETs Using Distributed Dynamic Random Path Selection 65
Vakul Mohanty, Dhaval Moliya, Chittaranjan Hota, and Muttukrishnan Rajarajan

MAD-IDS: Novel Intrusion Detection System Using Mobile Agents and Data Mining Approaches 73
Imen Brahmi, Sadok Ben Yahia, and Pascal Poncelet

Data Encryption

Fuzzy Biometric Signcryption Scheme with Bilinear Pairings in the Standard Model 77
Mingwu Zhang, Bo Yang, Tsuyoshi Takagi, Yanping Shen, and Wenzheng Zhang

Key Independent Decryption of Graphically Encrypted Images 88
Ram Ratan

Towards Confidentiality of ID-Based Signcryption Schemes under Without Random Oracle Model 98
Mingwu Zhang, Pengcheng Li, Bo Yang, Hao Wang, and Tsuyoshi Takagi

JPEG Steganalysis Using HBCL Statistics and FR Index 105
Veena H. Bhat, Krishna S., P. Deepa Shenoy, Venugopal K.R., and L.M. Patnaik

Data and Text Mining

Text Mining Technique for Chinese Written Judgment of Criminal Case 113
Shihchieh Chou and Tai-Ping Hsing

Exploration of Document Relation Quality with Consideration of Term Representation Basis, Term Weighting and Association Measure 126
Nichnan Kittiphattanabawon, Thanaruk Theeramunkong, and Ekawit Nantajeewarawat

Identifying Controversial Issues and Their Sub-topics in News Articles 140
Yoonjung Choi, Yuchul Jung, and Sung-Hyon Myaeng

Efficient Privacy Preserving K-Means Clustering 154
Maneesh Upmanyu, Anoop M. Namboodiri, Kannan Srinathan, and C.V. Jawahar

A Fuzzy Threshold Based Modified Clustering Algorithm for Natural Data Exploration 167
Binu Thomas and G. Raju

Author Index 173

Combined Detection Model for Criminal Network Detection

Fatih Ozgul[1], Zeki Erdem[2], Chris Bowerman[3], and Julian Bondy[4]

[1,3] Faculty of Computing, Engineering &Technology,
University of Sunderland, SR6 0DD, Sunderland, United Kingdom
[2] TUBITAK- UEKAE, Information Technologies Institute,
41470 Gebze, Kocaeli, Turkey
[4] School of Global Studies, Social Science & Planning,
RMIT University, Melbourne, Australia
fatih.ozgul@istanbul.com, chris.bowerman@sunderland.ac.uk,
zeki.erdem@bte.tubitak.gov.tr, bondy@rmit.edu.au

Abstract. Detecting criminal networks from arrest data and offender demographics data made possible with our previous models such as GDM, OGDM, and SoDM and each of them proved successful on different types of criminal networks. To benefit from all features of police arrest data and offender demographics, a new combined model is developed and called as combined detection model (ComDM). ComDM uses crime location, date and modus operandi similarity as well as surname and hometown similarity to detect criminal networks in crime data. ComDM is tested on two datasets and performed better than other models.

Keywords: Criminal networks, crime data mining, clustering, group detection, police arrest data, offender demographics.

Criminal networks have been an interesting domain for computer scientists. Analyzing criminal networks, using various methods to find out relationships between criminals are investigated by social network analysis scientists as well. What is the lacking of research for criminal networks is how to extract possible relationships between criminals and using these relationship links to assume the existence of previously unseen hidden criminal networks. To look for these links, it is best to look for similarities in crime data. Traditionally there are two types of crime data are kept by the police; the first is police arrest records, and the second is offender demographics information.

1 Introduction

Current intelligence security informatics is mainly focused on using social network analysis (SNA) and machine learning techniques for structural and positional analysis [1, 2] of criminal networks [3, 4, 5] where mostly required information is provided

H. Chen et al. (Eds.): PAISI 2010, LNCS 6122, pp. 1–14, 2010.
© Springer-Verlag Berlin Heidelberg 2010

from non-crime data. More research therefore should be looking for using crime data provided by the police to detect criminal networks. This is because most of criminals have previous crime records and they have similarities for features of their crimes and their demographics information. GDM and OGDM [6, 7] is tested for detecting criminal networks previously and they were successful to extract co-offending knowledge, similarities of crime location, date preferences and modus operandi. Another model Socio cultural Detection Model (SoDM), which is based on criminals' surname and hometown similarity, is tested on two datasets where GDM and OGDM are also tested on. SoDM performed less well than GDM and OGDM, however. Since combining GDM, OGDM, and SoDM into one model can produce better results, a new model (ComDM) is offered to get maximum performance.

In this study, we test Combined Detection Model (ComDM) on two datasets which contain terrorist, drug, mafia, violence, and theft networks from two cities in Turkey; Bursa(including 85 criminal networks) and Diyarbakir (including 40 criminal networks). Detection results are compared against previously known and available criminal networks in datasets; findings of ComDM are measured with precision, recall and f-values. Feedbacks from domain experts in Bursa and Diyarbakir are also added. This paper shows:

- Criminological approach to criminal networks (Section 2).
- A brief literature review for criminal networks (Section 2).
- Formal definition of ComDM (Section 3).
- Results of datasets for ComDM (Section 4)
- Evaluation results for ComDM (Section 5)
- Domain experts' feedback on ComDM (Section5).

2 Criminal Network Detection

Criminal networks can be gangs, drug dealing networks, mafia type violence and theft groups, street pickpockets, hooligan groups, or terrorist cells. As social networks, criminal networks consist of two sorts of elements: actors (such as criminals, suppliers) and relations between these actors. In addition to actors' personal characteristics and skills, criminal network has also special characteristics as an entity; so each network can be holding specific behavioural heuristics, crime committing habits (e.g. modus operandi), and associations with other criminals in order to provide specific tools and skills.

There are criminological approaches for the reason why criminals choose to work together. If we can find those reasons, it is easy to select suitable features in data for using these features in criminal network detection. Some previous research has shown that the most of criminal networks are not dominated by centrally controlled organisations with a clear hierarchy and strict division of tasks where they rather operate in an informal and flexible way [8]. In many ways criminal groups are either similar to each other because of being its members' friend-of-a friend position, referral chain or might be in need of assistance and special expertise from other criminals. The expertise of knowledge is required to operate for their team to function properly. In terms of stability [8] in criminal networks, there are two types of actors (e.g.criminals); *life* *course persistent criminals* begin offending in early ages, engage in an array of offences and persist over the life course. On the other hand *adolescent limited criminals* begin offending later, engage in group based delinquent acts that are indicative of teenage rebellion, but then *adolescent limited criminals* tend to be in general decline of criminality. Canter [9] suggests that two dominant trends are identified for criminal networks; firstly the size of the network, secondly the product of the centrality of leadership. He states that using these criteria, three types of criminal organisation to be specified; *ad-hoc groups, oligarchies,* and *organised criminals*. *Ad-hoc groups* are with relatively little structure, sometimes with just the presence of key central figures. *Oligarchies* are the kind of networks where their communications appear to be controlled by a small group of people. *Organised criminals* are the closest to being an illegal organisation with most of differentiation, indicating a management hierarchy. *Ad-hoc groups* are the smallest sized groups whereas *organised criminals* are the largest sized ones. For instance, he suggests that hooligan groups are less structured whereas drug dealing networks are the most structured were found for all types of criminal activity.

In computer science literature, there are few social computing models developed for detection of criminal networks. The nearest similar research area is online social networks. For instance, Goldbeck [10] worked on online social networks, where users maintain lists of friends and express their preferences for items like movies, music, or books, are very popular. They say that for online social network systems to be effective, it is important to understand *the relationship between social and personal preferences*. Similar to that, Crandall [11] and his friends say that a fundamental open question in the analysis of social networks is to understand the interplay between *similarity and social ties*. He says people are similar to their neighbors in a social network for two distinct reasons: first, they grow to resemble their current friends due to social influence; and second, they tend to form new links to others who are already like them. This phenomenon is called as *selection* by sociologists. People tend to have attributes similar to those of their friends. There are two underlying reasons for this. First, the process of social influence leads people to adopt behaviors exhibited by those they interact with; this effect is at work in many settings where new ideas diffuse by word-of-mouth or imitation through a network of people. Second, distinct reason is that people tend to form relationships with others who are already similar to them.

One of the pioneering projects about criminal networks was COPLINK Connect, Detect and CrimeNet Explorer [12] in Arizona which did entity extraction text mining from police narrative reports, then created links between entities from same documents and created possible criminal networks, plotted these networks using multi dimensional scaling techniques. Xu et al. [13, 14] defined a framework for automated network analysis and visualization. Using COPLINK Connect and COPLINK Detect [12] structure to obtain link data from text, CrimeNet Explorer used a Reciprocal Nearest Neighbour (RNN) based clustering algorithm to find out links between offenders, as well as discovery of previously unknown groups. CrimeNet Explorer used concept space approach for network creation, RNN-based hierarchical clustering algorithm for group detection; social network analysis based structural analysis and Multi Dimensional Scaling (MDS) for network visualization.

FLINTS project [15] used soft behavioural and hard forensic (fingerprints, DNA) to give analysts the ability to build a graphical image of relations between crimes and criminals. FinCEN project [16] also aimed to reveal money laundering networks by comparing financial transactions. Oatley et al. [15] did some link analysis work on burglary cases in the OVER project. Skillicorn [17] did similar work on detection of the clusters within clusters to filter the surplus of information on possible terrorist networks and present the police a viable subset of suspects to work on. Another remarkable work is done by Adderly and Mushgrove [18, 19] applied clustering techniques and Self Organising Maps to model the behaviour of sex offenders.

TMODS, which is developed by 21st Century Technologies [20, 21, 22, 23, 24, 25] automates the tasks of searching for and analyzing instances of particular threatening activity patterns. With TMODS, the analyst can define an attributed relational graph to represent the pattern of threatening activity he or she is looking for. TMODS then automates the search for that threat pattern through an input graph representing the large volume of observed data. TMODS pinpoints the subset of data that match the threat pattern defined by the analyst thereby transforming a manual search into an efficient automated graph matching tool. User defined threatening activity or pattern graph can be produced with possible terrorist network ontology and this can be matched against observed activity graph. At the end, human analyst views matches that are highlighted against the input graph.

3 ComDM

Combined Group Detection Model (ComDM) is developed in order to benefit from maximum use of similarities in criminal behaviour (e.g. choice of crime location, time and modus operandi) between criminals and use of demographic similarities such as family bonds, relative bonds, and coming from the same hometown circumstances. ComDM is a combined model of OGDM [7] and SoDM. Police arrest records and offender demographic data are the input source of this model. Namely; crime location, crime date, crime modus operandi from arrest data, offender surnames and hometown information from offender demographics data are used as input in ComDM. Requirements of applying ComDM are two; the first is availability of crime location, date and modus operandi information in police arrest data and offender surname and place of birth or hometown information in offender demographics data. The second requirement is availability of these data in relational table format.

Six steps are taken in ComDM as presented in figure 1. In general, spatial, temporal, modus operandi, surname, and hometown links are created and then a clustering approach is applied for criminal network detection. The first step is linking similarly behaving criminals (operating in the same location, on the same dates, using similar modus operandi) who also come from the same family and hometown with SQL inner join queries. Using three fields of arrest table are crime features which are spatial, temporal and modus operandi fields for inner join queries; spatial, temporal, and modus operandi links are created. Similar to this operation, using two fields of demographics table are criminal features, which are surname and hometown fields; surname and hometown links are created.

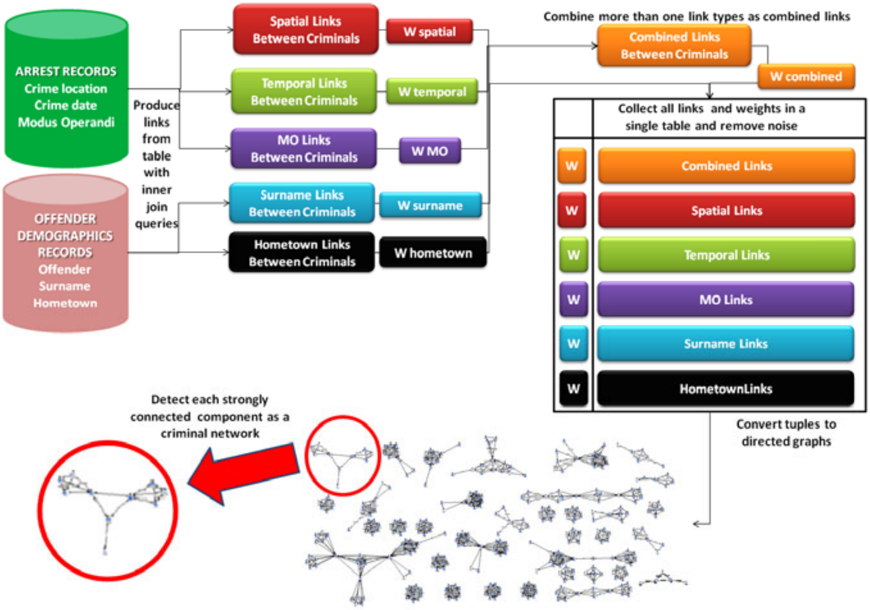

ARREST RECORDS
Crime location
Crime date
Modus Operandi
OFFENDER DEMOGRAPHICS RECORDS
Offender
Surname
Hometown
Produce links from table with inner join queries
Spatial Links Between Criminals
W spatial
Temporal Links Between Criminals
W temporal
MO Links Between Criminals
W MO
Surname Links Between Criminals
W surname
Hometown Links Between Criminals
W hometown
Combine more than one link types as combined links
Combined Links Between Criminals
W combined
Collect all links and weights in a single table and remove noise
W
Combined Links
W
Spatial Links
W
Temporal Links
W
MO Links
W
Surname Links
W
HometownLinks
Convert tuples to directed graphs
Detect each strongly connected component as a criminal network

Fig. 1. Combined Detection Model (ComDM)

For each two offenders (e.g. Offender A, and Offender B) two offenders are linked with spatial link (A, B), temporal link (A, B), MO link (A, B), combined link (A, B), surname link (A, B), hometown link (A, B) as well as spatial link (B,A), temporal link (B,A), MO link (B,A), combined link (B,A), surname link (B,A), hometown link (B,A).

$$\begin{gathered}
spatial\ Link\ (A,B) = Offender\ A \rightarrow Offender\ B \\
spatial\ Link\ (B,A) = Offender\ B \rightarrow Offender\ A \\
temporal\ Link\ (A,B) = Offender\ A \rightarrow Offender\ B \\
temporal\ Link\ (B,A) = Offender\ B \rightarrow Offender\ A \\
MO\ Link\ (A,B) = Offender\ A \rightarrow Offender\ B \\
MO\ Link\ (B,A) = Offender\ B \rightarrow Offender\ A \\
Combined\ Link\ (A,B) = Offender\ A \rightarrow Offender\ B \\
Combined\ Link\ (B,A) = Offender\ B \rightarrow Offender\ A \\
Surname\ Link\ (A,B) = Offender\ A \rightarrow\ Offender\ B \\
Surname\ Link\ (B,A) = Offender\ B \rightarrow Offender\ A \\
Hometown\ Link\ (A,B) = Offender\ A \rightarrow Offender\ B \\
Hometown\ Link\ (B,A) = Offender\ B \rightarrow Offender\ A \\
where\ Surname\ Link(A,B) = Surname\ Link\ (B,A) \\
and\ Hometown\ Link(A,B) = Hometown\ Link\ (B,A)
\end{gathered}$$

The second step is giving spatial, temporal, modus operandi, surname, and hometown link weights according to distribution of links for Offender A. Offender A has five link weights compared against all similar offenders as,

$$W_{spatial\ Link(A,B)} = \frac{\sum number\ of\ spatial\ links\ between\ (A,B)}{\sum number\ of\ spatial\ links\ for\ A} \tag{1}$$

$$W_{spatial\ Link(B,A)} = \frac{\sum number\ of\ spatial\ links\ between\ (A,B)}{\sum number\ of\ spatial\ links\ for\ B} \tag{2}$$

$$W_{temporal\ Link(A,B)} = \frac{\sum number\ of\ temporal\ links\ between\ (A,B)}{\sum number\ of\ temporal\ links\ for\ A} \tag{3}$$

$$W_{temporal\ Link(B,A)} = \frac{\sum number\ of\ temporal\ links\ between\ (A,B)}{\sum number\ of\ temporal\ links\ for\ B} \tag{4}$$

$$W_{MO\ Link(A,B)} = \frac{\sum number\ of\ modus\ operandi\ links\ between\ (A,B)}{\sum number\ of\ modus\ operandi\ links\ for\ A} \tag{5}$$

$$W_{MO\ Link(B,A)} = \frac{\sum number\ of\ modus\ operandi\ links\ between\ (A,B)}{\sum number\ of\ modus\ operandi\ links\ for\ B} \tag{6}$$

$$W_{Surname\ Link(A,B)} = \frac{1}{\sum number\ of\ Surname\ links\ for\ A} \tag{7}$$

$$W_{Surname\ Link(B,A)} = \frac{1}{\sum number\ of\ Surname\ links\ for\ B} \tag{8}$$

$$W_{Hometown\ Link(A,B)} = \frac{1}{\sum number\ of\ Hometown\ links\ for\ A} \tag{9}$$

$$W_{Hometown\ Link(B,A)} = \frac{1}{\sum number\ of\ Hometown\ links\ for\ B} \tag{10}$$

The third step is identifying combined links between two offenders where there are more than one type of link exist (e.g. there are links between two offenders as spatial link, temporal link, modus operandi link, surname link, and hometown link). In case of more than two link types between Offender A and Offender B;

$$\begin{gathered} spatial\ Link\ (A,B) = Offender\ A \rightarrow Offender\ B \\ temporal\ Link\ (A,B) = Offender\ A \rightarrow Offender\ B \\ MO\ Link\ (A,B) = Offender\ A \rightarrow Offender\ B \\ Surname\ Link\ (A,B) = Offender\ A \rightarrow\ Offender\ B \\ Hometown\ Link\ (A,B) = Offender\ A \rightarrow Offender\ B \end{gathered}$$

Then these links are reduced to one single link as;

$$Combined\ Link\ (A,B) = Combined\ Link\ (B,A) = Offender\ A \rightarrow Offender\ B \tag{11}$$

and link weight for combined link is calculated as geometric mean of spatial, temporal, modus operandi, surname and hometown link weights (or just any two of them) as follows;

$$W_{Combined\ Link(A,B)} = W_{Combined\ Link(B,A)} = \sqrt{\begin{matrix} W_{spatial\ Link\ (A,B)}2 \\ + W_{temporal\ Link\ (A,B)}2 \\ + W_{MO\ Link\ (A,B)}2 \\ + W_{Surname\ Link\ (A,B)}2 \\ + W_{Hometown\ Link\ (A,B)}2 \end{matrix}} \tag{12}$$

The fourth step is removing weaker links which hold link weighting below a threshold level. Threshold level is 0.1 and links below this threshold is removed as noise. This is represented as,

$$\begin{gathered} W_{spatial\ Link(A,B)}\ and\ W_{spatial\ Link(B,A)} \geq 0.1, \\ W_{temporal\ Link(A,B)}\ and\ W_{temporal\ Link(B,A)} \geq 0.1\ , \\ W_{Mo\ Link(A,B)}\ and\ W_{Mo\ Link(B,A)} \geq 0.1, \\ W_{Surname\ Link(A,B)}\ and\ W_{Surname\ Link(B,A)} \geq 0.1, \\ W_{Hometown\ Link(A,B)}\ and\ W_{Hometown\ Link(B,A)} \geq 0.1, \\ W_{Combined\ Link(A,B)}\ and\ W_{Combined\ Link(B,A)} \geq 0.1 \end{gathered}$$

The fifth step is gathering all combined links and other remaining links which are holding more than 0.1 link weights (e.g. threshold value) in graph format. Resulting graph contains criminal networks which have at least three offenders linked to each other with different type of link weights above the selected threshold (e.g. 0.1). Such as;

$$\begin{aligned} Graph_{criminal\ network} = A, B, C\ and\ spatial\ (A,B), spatial(B,A), surname(A,C), \\ surname\ (C,A), combined\ (B,C), combined(C,B) \end{aligned} \tag{13}$$

Finally the sixth step is detecting individual criminal networks using strongly connected components (SCC) algorithm [26]. We finally get a criminal network in graph format. When a detected graph contains two (or more) criminal networks and they are equal number of criminals in two subgroups, then feature selection method [27] is applied when deciding which criminal network is prevailing. It is based on comparison of means and variances of total link weights. Using selection scores exhibited in the following equations for subgroups, the higher scoring subgroup is decided as detected major criminal network within this graph.

$$Test\ Score\ (1st\ \ subgroup) = \sqrt{\frac{var\ \left(W_{links\ of\ 1st\ subgroup}\right)}{number\ of\ subgraps\ for\ 1st\ subgroup}} \tag{14}$$

$$Test\ Score\ (2nd\ \ subgroup) = \sqrt{\frac{var\ \left(W_{links\ of\ 2nd\ subgroup}\right)}{number\ of\ subgraps\ for\ 2nd\ subgroup}} \tag{15}$$

$$\text{let } mean\left(W_{links\ of\ 1st\ subgroup}\right) > mean\left(W_{links\ of\ 2nd\ subgroup}\right)$$

$$Selection\ Score\ (1st\ subgroup) = \frac{|mean\ (W_{links\ of\ 1st\ subgroup}) - mean\ (W_{links\ of\ 2nd\ subgroup})|}{Test\ Score\ (1st\ subgroup)} \quad (16)$$

$$Selection\ Score\ (2nd\ subgroup) = \frac{|mean\ (W_{links\ of\ 1st\ subgroup}) - mean\ (W_{links\ of\ 2nd\ subgroup})|}{Test\ Score\ (2nd\ subgroup)} \quad (17)$$

Confusion matrix for ComDM is constructed as similar to metrics offered by Kaza et al.[28];

Table 1. Confusion matrix for ComDM

	Offenders considered as Criminal Network Members	**Offenders considered as not Criminal Network Members**
Offenders considered refer to the same Criminal Network	**TP**- True Positive	**FP** - False Positive
Offenders considered refer to other Criminal Networks, not to the same Criminal Network	**FN** - False Negative	**TN** - True Negative

In a detected strongly connected component, when detected criminals are all within detected graph, it is accepted as true positive (TP). In case of detected criminals are mostly in the same graph but there are also other unrelated individual criminals, unrelated individual criminals are accepted as false negatives (FN), rightly detected criminals are accepted are still true positive (TP). In case of detected criminals are mostly in strongly connected component but they are belong to more than one criminal networks, such as two or three, criminals which are in the second (or third) unexpected criminal network are accepted as false positive (FP). Precision, recall and f-measure are given in equations below. Precision means that retrieved group is relevant. Recall means that relevant group is retrieved. F-measure is harmonic mean of precision and recall values.

$$precision = \frac{TP}{TP + FP} \quad (18)$$

$$recall = \frac{TP}{TP + FN} \quad (19)$$

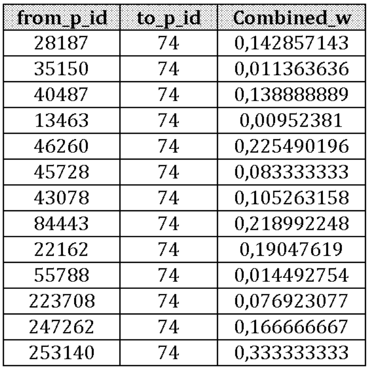
$$F - measure = 2.\frac{precision\,.recall}{precision + recall} \quad (20)$$

4 ComDM Testbeds

To test whether ComDM performs well for detecting criminal networks, two datasets are used as testbeds. They are Bursa Criminal Networks (BCN) and Diyarbakir Drug Networks (DDN) and each has different characteristics. BCN includes 85 criminal network of various types; mostly theft and violence networks, including 8 terrorist networks. DDN includes 40 drug dealing and -mafia type- organised crime networks. Both of the datasets are extracted from massive police databases and they also include some unrelated crimes and criminals. BCN and DDN both are gathered from two sources of crime data; police arrest records, and offender demographics records.

4.1 ComDM Testbed: Bursa Dataset

In this task, all types of links for Bursa Criminal Networks, which are obtained in the previous tasks, are collected. Those are namely spatial, temporal, modus operandi links, and surname and hometown links. All these links are merged altogether, and if there is more than one link between two criminals are available, and then geometric mean of those links weighting values are calculated and given as combined link weight between those two criminals. According to this definition, links and link weights are calculated and they are partly presented in the figure 2. In general, with ComDM, the number of links is increased and there are many more relations between criminals in ComDM.



from_p_id	to_p_id	Combined_w
28187	74	0,142857143
35150	74	0,011363636
40487	74	0,138888889
13463	74	0,00952381
46260	74	0,225490196
45728	74	0,083333333
43078	74	0,105263158
84443	74	0,218992248
22162	74	0,19047619
55788	74	0,014492754
223708	74	0,076923077
247262	74	0,166666667
253140	74	0,333333333

Fig. 2. Combined links and link weights in Bursa Dataset

When these links are collected for building graphs, some members are observed (see Figure 3) to have “brokerage” roles. Detection findings for ComDM are better than GDM, OGDM, and SoDM for Bursa dataset. As presented in figure 4, just one criminal network (e.g. BCN#85) is not detected. Most of the networks are detected with high precision. Another success is about high accuracy of recall score. The only problem is about merging many networks into a single big network, thus detections are typical subgraph detections and some network members are highly connected to some “outer nodes”, which causes them to be “brokers”.

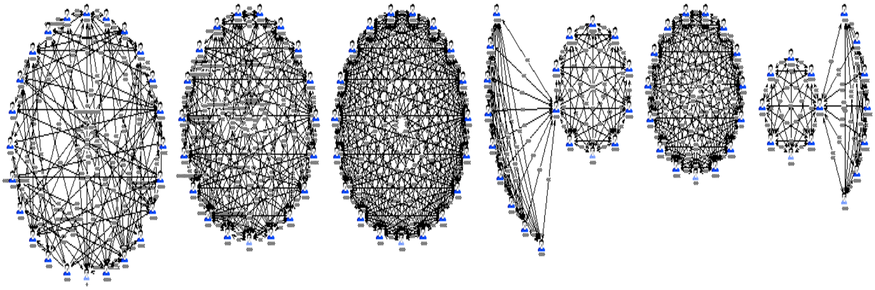
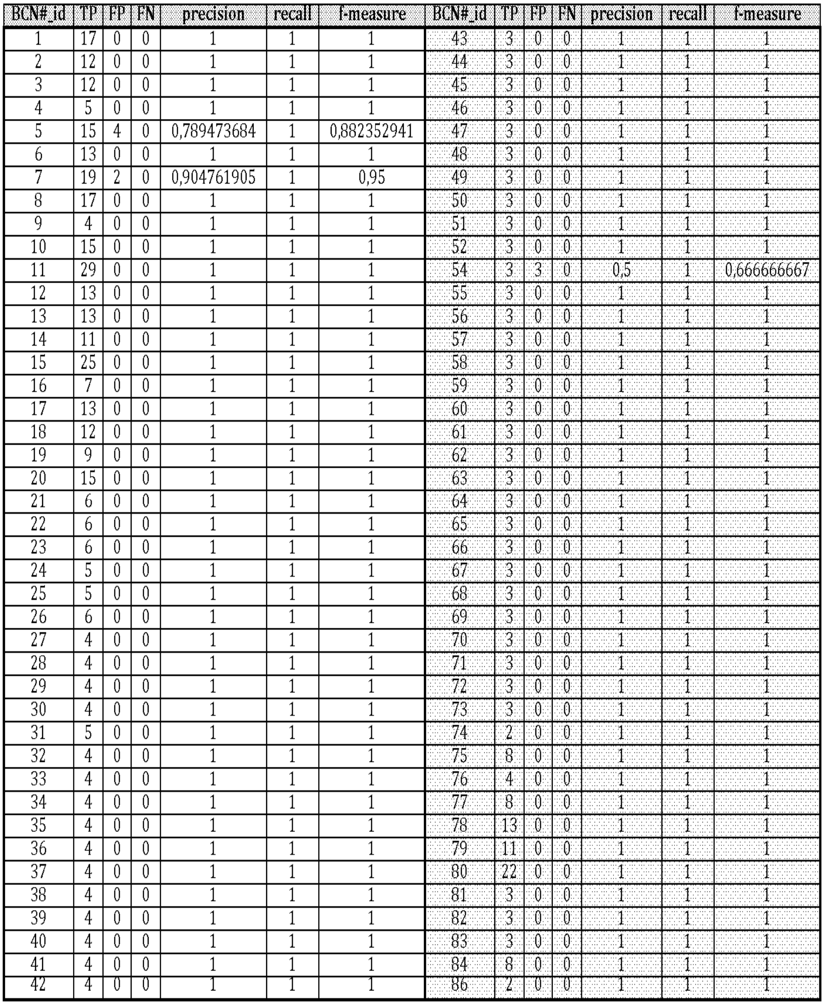


Fig. 3. Example detected criminal networks using ComDM in Bursa Dataset

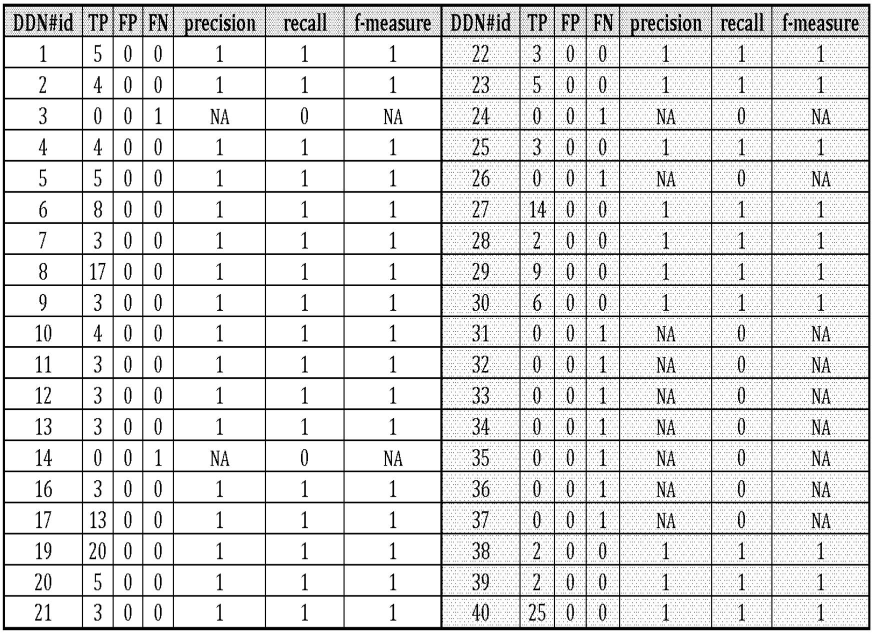


BCN#_id	TP	FP	FN	precision	recall	f-measure	BCN#_id	TP	FP	FN	precision	recall	f-measure
1	17	0	0	1	1	1	43	3	0	0	1	1	1
2	12	0	0	1	1	1	44	3	0	0	1	1	1
3	12	0	0	1	1	1	45	3	0	0	1	1	1
4	5	0	0	1	1	1	46	3	0	0	1	1	1
5	15	4	0	0,789473684	1	0,882352941	47	3	0	0	1	1	1
6	13	0	0	1	1	1	48	3	0	0	1	1	1
7	19	2	0	0,904761905	1	0,95	49	3	0	0	1	1	1
8	17	0	0	1	1	1	50	3	0	0	1	1	1
9	4	0	0	1	1	1	51	3	0	0	1	1	1
10	15	0	0	1	1	1	52	3	0	0	1	1	1
11	29	0	0	1	1	1	54	3	3	0	0,5	1	0,666666667
12	13	0	0	1	1	1	55	3	0	0	1	1	1
13	13	0	0	1	1	1	56	3	0	0	1	1	1
14	11	0	0	1	1	1	57	3	0	0	1	1	1
15	25	0	0	1	1	1	58	3	0	0	1	1	1
16	7	0	0	1	1	1	59	3	0	0	1	1	1
17	13	0	0	1	1	1	60	3	0	0	1	1	1
18	12	0	0	1	1	1	61	3	0	0	1	1	1
19	9	0	0	1	1	1	62	3	0	0	1	1	1
20	15	0	0	1	1	1	63	3	0	0	1	1	1
21	6	0	0	1	1	1	64	3	0	0	1	1	1
22	6	0	0	1	1	1	65	3	0	0	1	1	1
23	6	0	0	1	1	1	66	3	0	0	1	1	1
24	5	0	0	1	1	1	67	3	0	0	1	1	1
25	5	0	0	1	1	1	68	3	0	0	1	1	1
26	6	0	0	1	1	1	69	3	0	0	1	1	1
27	4	0	0	1	1	1	70	3	0	0	1	1	1
28	4	0	0	1	1	1	71	3	0	0	1	1	1
29	4	0	0	1	1	1	72	3	0	0	1	1	1
30	4	0	0	1	1	1	73	3	0	0	1	1	1
31	5	0	0	1	1	1	74	2	0	0	1	1	1
32	4	0	0	1	1	1	75	8	0	0	1	1	1
33	4	0	0	1	1	1	76	4	0	0	1	1	1
34	4	0	0	1	1	1	77	8	0	0	1	1	1
35	4	0	0	1	1	1	78	13	0	0	1	1	1
36	4	0	0	1	1	1	79	11	0	0	1	1	1
37	4	0	0	1	1	1	80	22	0	0	1	1	1
38	4	0	0	1	1	1	81	3	0	0	1	1	1
39	4	0	0	1	1	1	82	3	0	0	1	1	1
40	4	0	0	1	1	1	83	3	0	0	1	1	1
41	4	0	0	1	1	1	84	8	0	0	1	1	1
42	4	0	0	1	1	1	86	2	0	0	1	1	1

Fig. 4. Results, evaluation of ComDM clustering in Bursa Dataset

4.2 ComDM Testbed :Diyarbakir Dataset

In this task, all types of links for Diyarbakir Drug Networks' are collected, merged, and links that are more than one pertaining to the same two nodes are treated by combined links and combined weighting score is used. Spatial, temporal, modus operandi, surname, hometown and combined links are merged altogether. Resulting links and link weights are calculated. Just like in Bursa dataset, the number of links is increased and there are many more relations between criminals than single link types any more. After completion of graph building, disconnected components are identified and treated as individual criminal networks. Results and evaluation metrics for the results are presented in figure 5. Contrary to Bursa dataset, there are twelve networks which cannot be totally detected. DDN#18 is entirely undetected at all. Other eleven criminal networks (DDN#3, DDN#14, DDN#24, DDN#26, DDN#31, DDN#32, DDN#33, DDN#34, DDN#35, DDN#36, DDN#37) are wrongly detected, and just one member is offered which is not member.



DDN#id	TP	FP	FN	precision	recall	f-measure	DDN#id	TP	FP	FN	precision	recall	f-measure
1	5	0	0	1	1	1	22	3	0	0	1	1	1
2	4	0	0	1	1	1	23	5	0	0	1	1	1
3	0	0	1	NA	0	NA	24	0	0	1	NA	0	NA
4	4	0	0	1	1	1	25	3	0	0	1	1	1
5	5	0	0	1	1	1	26	0	0	1	NA	0	NA
6	8	0	0	1	1	1	27	14	0	0	1	1	1
7	3	0	0	1	1	1	28	2	0	0	1	1	1
8	17	0	0	1	1	1	29	9	0	0	1	1	1
9	3	0	0	1	1	1	30	6	0	0	1	1	1
10	4	0	0	1	1	1	31	0	0	1	NA	0	NA
11	3	0	0	1	1	1	32	0	0	1	NA	0	NA
12	3	0	0	1	1	1	33	0	0	1	NA	0	NA
13	3	0	0	1	1	1	34	0	0	1	NA	0	NA
14	0	0	1	NA	0	NA	35	0	0	1	NA	0	NA
16	3	0	0	1	1	1	36	0	0	1	NA	0	NA
17	13	0	0	1	1	1	37	0	0	1	NA	0	NA
19	20	0	0	1	1	1	38	2	0	0	1	1	1
20	5	0	0	1	1	1	39	2	0	0	1	1	1
21	3	0	0	1	1	1	40	25	0	0	1	1	1

Fig. 5. Results, evaluation of ComDM clustering in Diyarbakir Dataset

General detection view is similar to those of co-offending and demographics clustering view (figure 6). The only advantage for ComDM detection is its high accuracy for precision and recall values. In general, ComDM results are better than SoDM results. But GDM and OGDM detection results are far better than ComDM detection results for Diyarbakir drug networks.

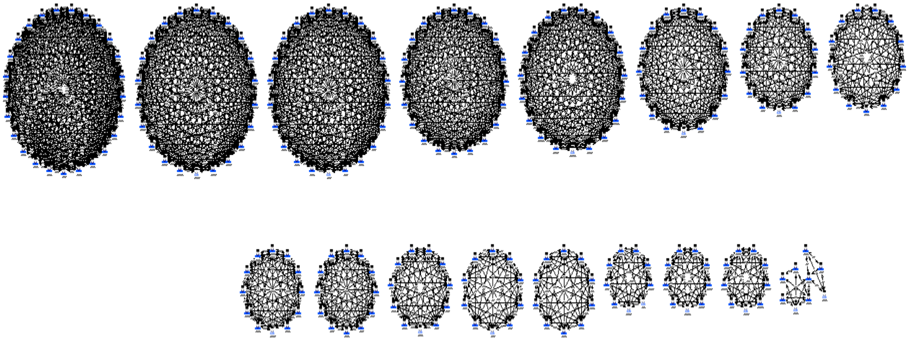


Fig. 6. Example detected criminal networks using ComDM in Diyarbakir Dataset

5 Evaluations and Domain Expert Feedback

Precision, recall and f-test values for Bursa Criminal Networks are 0.99, 0.99, 0.99, for Diyarbakir Drug Networks they are 0.71, 0.71, 0.71. In average, precision, recall and f-test scores for ComDM are 0.85, 0.85, 0.85. ComDM is accepted as successful according to these results. ComDM performs better on BCN to compare against DDN.

Diyarbakir domain experts viewed ComDM as the best detection model compared to others (e.g. GDM, OGDM, and SoDM). Diyarbakir domain experts always recommended using location, modus operandi, surname, and hometown features as potentially the most lucrative features when detecting criminal networks. Contrary to Diyarbakir domain experts, Bursa domain experts offered co-offending, location, and modus operandi as the most valuable features of crime for detecting criminal networks. Bursa domain experts found ComDM as the second most successful model after OGDM. But they said using too much features such as in ComDM might cause bulky results, which is undesirable.

6 Conclusion

We used crime features and offender demographics for detection of criminal networks. ComDM, which uses crime location, date, modus operandi, surname and hometown of criminals to look for similarities between criminals, is designed for better performance on many features of crime and offender demographics. These similarities are processed to obtain criminal networks. Experimental tests on two datasets showed that ComDM performs well on terrorist, theft, violence networks, where it performs less well on drug and mafia networks.

References

1. Brandes, U.: A Faster Algorithm for betweenness centrality. Journal of Mathematical Sociology 25(2), 163–177 (2001)
2. Coffman, T.R., Marcus, S.E.: Pattern Classification in Social Network Analysis: A case study. In: 2004 IEEE Aerospace Conference, March 6-13 (2004)

3. Hunter, A.: Leninist Cell Data Analysis. 21st Century Technologies Inc., Austin (2002)
4. Smith, M.N., King, P.J.H.: Incrementally Visualising Criminal Networks. In: Sixth International Conference on Information Visualisation (IV'02), IEEE, Los Alamitos (2002)
5. Wang, G.A., Xu, J.J., Chen, H.: Using Social Contextual Information to Match Criminal Identities. In: Proceedings of the 39th Annual Hawaii International Conference on System Sciences, vol. 04, pp. 81.2. IEEE Computer Society, Los Alamitos (2006)
6. Ozgul, F., Bondy, J., Aksoy, H.: Mining for offender group detection and story of a police operation. In: Sixth Australasian Data Mining Conference (AusDM 2007), Australian Computer Society Conferences in Research and Practice in Information Technology (CRPIT), Gold Coast, Australia (2007)
7. Ozgul, F., Erdem, Z., Aksoy, H.: Comparing Two Models for Terrorist Group Detection: GDM or OGDM? In: Yang, C.C., Chen, H., Chau, M., Chang, K., Lang, S.-D., Chen, P.S., Hsieh, R., Zeng, D., Wang, F.-Y., Carley, K.M., Mao, W., Zhan, J. (eds.) ISI Workshops 2008. LNCS, vol. 5075, pp. 149–160. Springer, Heidelberg (2008)
8. Mcglorin, J.M., Sullivan, C.J., Piquero, A.R., Bacon, S.: Investigating the stability of co-offenders among a sample of youthful offenders. Criminology 46(1), 155–187 (2008)
9. Canter, D.: A partial order scalogram analysis of criminal network structures. Behaviormetrika 31(2), 131–152 (2004)
10. Golbeck, J.: Trust and nuanced profile similarity in online social networks. ACM Trans. Web 3, 4, Article 12, 33 pages(2009)
11. Crandall, D., Cosley, D., Huttenlocher, D., Kleinberg, J., Suri, S.: Feedback Effects between Similarity and Social Influence in Online Communities. In: KDD'08, Las Vegas, Nevada, USA, August 24-27. ACM, New York (2008)
12. Chau, M., Xu, J., Chen, H.: Extracting meaningful entities from police narrative reports. In: National Conference on Digital Government Research (2001)
13. Xu, J., Chen, H.C.: Fighting Organised Crimes: using shortest-path algorithms to identify associations in criminal networks. Decision Support Systems 38(3), 473–487 (2003)
14. Xu, J., Chen, H.C.: CrimeNet Explorer: A Framework for Criminal Network Knowledge Discovery. ACM Transactions on Information Systems 23(2), 201–226 (2005)
15. Oatley, G.C., Zeleznikov, J., Ewart, B.W.: Matching and predicting crimes. In: AI 2004-The 24th SGAI International Conference on Knowledge Based Systems and Applications of Artificial Intelligence (2004)
16. Goldberg, H.G., Wong, R.W.H.: Restructuring transactional data for link analysis in FinCEN AI System. In: AAAI Fall Symposium (1998)
17. Skillicorn, D.B.: Clusters within clusters: SVD and counterterrorism. In: Workshop on Data Mining for Counterterrorism and Security (2003)
18. Adderley, R., Badii, A., Wu, C.: The automatic identification and prioritization of criminal networks from police crime data. In: Ortiz-Arroyo, D., Larsen, H.L., Zeng, D.D., Hicks, D., Wagner, G. (eds.) EuroIsI 2008. LNCS, vol. 5376, pp. 5–14. Springer, Heidelberg (2008)
19. Adderley, R., Mushgrove, P.B.: Data mining case study: Modeling the behavior of offenders who commit sexual assaults. In: ACM SIGKDD 2001 International Conference on Knowledge Discovery and Data Mining, New York, pp. 215–220 (2001)
20. Coffman, T., Greenblatt, S., Marcus, S.: Graph-based technologies for intelligence analysis. Communication of ACM 47(3), 45–47 (2004)
21. Marcus, S., Coffman, T.: Terrorist Modus Operandi Discovery System 1.0: Functionality, Examples, and Value. 21st Century Technologies, Austin (2002)
22. Marcus, S.E., Moy, M., Coffman, T.: Social Network Analysis. In: Cook, D.J., Holder, L.B. (eds.) Mining Graph Data. John Wiley & Sons, Inc., Hoboken (2007)

23. Moy, M.: Using TMODS to run best friends group detection algorithm. 21st Century Technologies, Austin (2005)
24. Coffman, T.R., Marcus, S.E.: Dynamic Classification of Suspicious Groups using social network analysis and HMMs. In: 2004 IEEE Aerospace Conference, March 6-13 (2004)
25. Greenblatt, S., Coffman, T., Marcus, S.: Emerging Information Technologies and enabling policies for counter terrorism. In: Behaivoural Network Analysis for Terrorist Detection. Wiley-IEEE Press, Hoboken (2005)
26. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to Algorithms, 2nd edn. (2001)
27. Kantardzic, M.: Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons, New York (2003)
28. Kaza, S., Hu, D., Atabakhsh, H., Chen, H.: Predicting criminal relationships using multivariate survival analysis. In: Proceedings of the 8th annual international conference on Digital government research: bridging disciplines & domains, pp. 290–291. Digital Government Society of North America, Philadelphia (2007)

An Association Model for Implicit Crime Link Analysis

Jau-Hwang Wang and Chien-Lung Lin

Department of Information management
Central Police University, Taiwan
56 Shu-Ren Road, Ta-Kang, Kwei-Shan
Tao-Yuan, Taiwan, 333
jwang@mail.cpu.edu.tw

Abstract. Link analysis has been an important tool in crime investigation. Explicit social links, such as kinship, financial exchange, and telephone connection, are often used to construct links between criminals. However implicit links, such as modus operandi, times of day, and geographic relationship, are seldom used to establish relationships between crime entities. This paper proposes an association model based on modus operandi and geographic relationship to establish links among crime cases and criminals. A data collection from local police department is used for experiment to evaluate the performance of the proposed approach.

Keywords: Link Analysis, Implicit Crime Link, Crime Association Model, Modus operandi, Geographic Relationship.

1 Introduction

One of the major goals of data mining is to discover implicit knowledge or patterns hidden in data set [5]. The concept of implicit link is defined as the association established from mining database contents. Explicit social links [14], such as friendship, kinship, financial exchange, sexual relationships and telephone connection, are often used to establish social network among suspects and victims in crime investigation. Although implicit crime links, such as geographic relationship and modus operandi, are valuable and often used in crime profiling, the effectiveness of their usage is mainly depended on the interpretation and expertise of an investigator. While link analysis of solved cases could be based solely on explicit links, associations among unsolved ones can only be constructed using implicit links from crime database mining.

In Taiwan, the concluding step of crime investigation procedures is to file a report for each case. Besides essential case data, such as case category, location, time of a day, suspect's and victim's names, the report also includes five modus operandi variables, namely crime cause(CC), crime habit(CH), preparation action(PA), crime method(CM), and crime tool(CT). Each modus operandi variable has a set of predefined values for selection in filing report for the case.

Crime is the result of complex changes in the whole community, such as economic, social and cultural factors [4]. For examples, localities that are more stable both residentially and socially usually have lower crime rates, violent and property offences

H. Chen et al. (Eds.): PAISI 2010, LNCS 6122, pp. 15–21, 2010.
© Springer-Verlag Berlin Heidelberg 2010

may increase faster in regional areas, and illicit drug usage levels in metropolitan area may be higher than that of suburban areas. It is often observed that the similarity or strength of link between crime cases is a inverse function of the distance between their crime locations [3].

This paper proposes an association model based on modus operandi and geographic relationship of crime to compute implicit links among crime cases and criminals. The organization of this paper is as follows: section 2 describes related literatures, section 3 defines the proposed model, section 4 presents the experimental design and experiment results, and section 5 gives conclusions of this paper.

2 Related Literature

2.1 Link Analysis

Link analysis is based on a field called graph theory, which is a branch of mathematics. Link analysis is also a major application of social network analysis and can be used to solve real world problems [1, 2]. For example:

1. Telecommunication industry often collects time and frequency of telephone calls to analyze customer's telephone call patterns using link analysis tools.
2. Referral patterns among physicians are usually able to reveal through referral link analysis.
3. Crime intelligences from different sources could be combined together to derive crime leads using link analysis techniques.

Social network, once mainly based on explicit links, such as kinship, marriage, financial exchange, web links, blog interactions, and mentorship, has evolved to include also implicit links, such as values, visions, friends, dislike, co-occurrence relationship etc [5, 9, 11, 14]. For example, implicit links, such as co-occurrence, have been used widely in information retrieval and association rule mining.

2.2 Criminal Profiling

According Locard's exchange principle: every contact of the perpetrators of a crime scene leaves a trace, i.e." the perpetrators will both bring something into the scene and leave with something from the scene."[15]. Furthermore, the type 、quantity 、position 、and status of evidences left at a crime scene often provides clues for reconstructing crime [7]. By analyzing evidences left at a crime scene, a criminal profiling specialist can derive the information needed to reconstruct a crime and narrowing down suspects [12]. Criminal profiling is mainly based on the 'method of operations' or modus operandi, such as identity and common characteristics of victim(s), weapon(s) used, hostility, presence or lack of any torture and/or sexual molestation etc. The information derived from profiling may include perpetrator's personality, sex, age, background, and possible physical features such as height and weight. Furthermore, after a criminal gets used to a certain method of operation, he/she will use the same modus operandi again in committing other cases [8]. Therefore, modus operandi information not only can be used in identifying the relationship between suspects and crime cases, but also can be used to discover the association among unsolved cases [13].

3 Proposed Model

3.1 Modeling Variable Weight Using Information Entropy

In Shannon information theory, entropy is the measure of information content of messages [10]. Information entropy is often used as attribute selection measure in decision tree induction algorithms, such as ID3 and C4.5 [5]. The entropy E of a discrete random variable X with possible values { $x_1, x_2, x_3 ..., x_n$ } is defined as:

$$E(X) \equiv -\sum_{i=1}^{n} p(x_i)\log_2 p(x_i)$$

Where p(x_i) is the likelihood of x_i . The information entropy represents the importance or distinguishing power of a random variable. As stated, the modus operandi consists of five variables. Each variable has its domain and probability distribution. The information entropy of each variable can be calculated and then used as its weight.

3.2 Modeling Value Weight Using Frequency and Inverse Case Frequency

According to research results in criminology, a small fraction of criminals are chronic offenders, however they commit a large portion of crime cases [6]. The statistic of our data collection also shows that the average number of cases committed by chronic offenders is a lot more than that by occasional offenders. Further analysis on the data collection reveals that domains of modus operandi variables of a chronic offender are much smaller than those of modus operandi variables in data collection. Furthermore, domains of modus operandi variables of a chronic offender have their own probability distribution and thus the weight of each value has to be modeled accordingly. We adopt the concept similar to the term frequency and inverse document frequency used in information retrieval [11]. The weight of each value V on a variable F of a chronic criminal P, is computed by:

$$\log \frac{p_{V_{PF}}}{p_{V_F}}$$

Where $p_{V_{PF}}$ is the probability of value V of variable F in the cases committed by chronic offender P, and p_{V_F} is the probability of value V of variable F in data set. In other words, the weight of a value is log proportional to the probability of its appearance in those cases committed by a chronic offender and inverse log proportional to the probability of its appearance in data set.

3.3 Association Modeling for Modus Operandi

Assuming that the modus operandi variables are independent, the association between a known criminal P and a crime case Q can then be calculated by the following formula:

$$\sum_{F\in\{CC,CH,PA,CM,CT\}} E(F) \times match(P_F, Q_F)$$

Where $match(P_F, Q_F) = \log \frac{p_{V_{PF}}}{p_{V_F}}$ if $P_F = Q_F$ and $\log \frac{p_{V_{PF}}}{p_{V_F}} > 0$, otherwise $match(P_F, Q_F) = 0$, etc.

3.4 Association Modeling for Crime Geographic Relationship

As stated the similarity or strength of link between crime cases is a inverse function of the distance between their crime locations and according to our data analysis, the association of crime geographic relationship is best modeled using the following formula:

$$\text{Association} = \begin{cases} 1, \text{ if distance } d \text{ is less than or equal 2} \\ \frac{1}{(d-1)^{1.5}}, \text{ if d} > 2 \end{cases}$$

4 Experiment and Evaluation

4.1 Data Set

The data set of our experiment consists of 1504 solved robbery cases from a local police department. Each record has fields such as location, date, time, crime cause, crime habit, preparation action, crime method, and crime tool, etc. Crime locations and modus operandi were extracted for experiment. Among 1504 cases, 23.07 % or 347 cases were committed by 3.33% or 23 criminals each committed more than 10 cases. Among the 23 criminals, 14 of them committed more than 10 consecutive cases before being caught by police and they were chosen to evaluate the proposed association model.

4.2 Association by Combining Modus Operandi and Geographic Relationship

The best strategy for combining associations of modus operandi and geographic relationship is tested using ratios 4:1 (P4D1), 2:1 (P2D1), 1:1 (P1D1), 1:2 (P1D2), and 1:4 (P1D4). For examples, the weight of modus operandi is 4 times of that of geographic relationship in P4D1 and the weight of modus operandi is half of that of geographic relationship in P1D2. Both modus operandi and geographic relationship associations are scaled to a range between 0.0 and 1.0 before combining them and the combined association is also normalized to the same range.

4.3 Experiment Design

The associations for each of 14 criminals and every of the 1504 cases were computed and used as measures to retrieve cases (from the 1504 cases) for each criminals. The retrieval effectiveness is represented in terms of recall, precision, and F-measure, using association levels (thresholds) 0.9, 0.8, 0.7, and 0.6. Recall is the number of cases committed by a criminal that also retrieved divided by the number of cases committed by a criminal, precision is the number of retrieved cases which are committed by a criminal divided by the number of cases retrieved for each criminal, and F-measure is: 2 *Recall*Precision /(Recall + Precision).

4.4 Experiment Results

The recall, precision, and F-measure for each of the 14 criminals were computed and their averages were summarized in Table 1, 2, 3. The average F-measures were also plotted in Fig. 1.

Table 1. Average Precision for Each Association level and Weight Ratio

Association Level Weight Ratio	0.9	0.8	0.7	0.6
P1D4	0.43	0.33	0.32	0.30
P1D2	0.49	0.41	0.37	0.32
P1D1	0.50	0.46	0.40	0.37
P2D1	0.50	0.50	0.50	0.41
P4D1	**0.62**	0.54	0.50	0.38

Table 2. Average Recall for Each Association level and Weight Ratio

Association Level Weight Ratio	0.9	0.8	0.7	0.6
P1D4	0.46	0.74	0.76	**0.79**
P1D2	0.34	0.52	0.71	0.78
P1D1	0.32	0.45	0.54	0.71
P2D1	0.26	0.36	0.54	0.59
P4D1	0.23	0.34	0.49	0.56

Table 3. Average F-measures for Each Association level and Weight Ratio

Association Level / Weight Ratio	0.9	0.8	0.7	0.6
P1D4	0.45	0.45	0.45	0.44
P1D2	0.40	0.46	0.48	0.45
P1D1	0.39	0.46	0.46	0.48
P2D1	0.34	0.42	**0.52**	0.48
P4D1	0.34	0.42	0.50	0.45

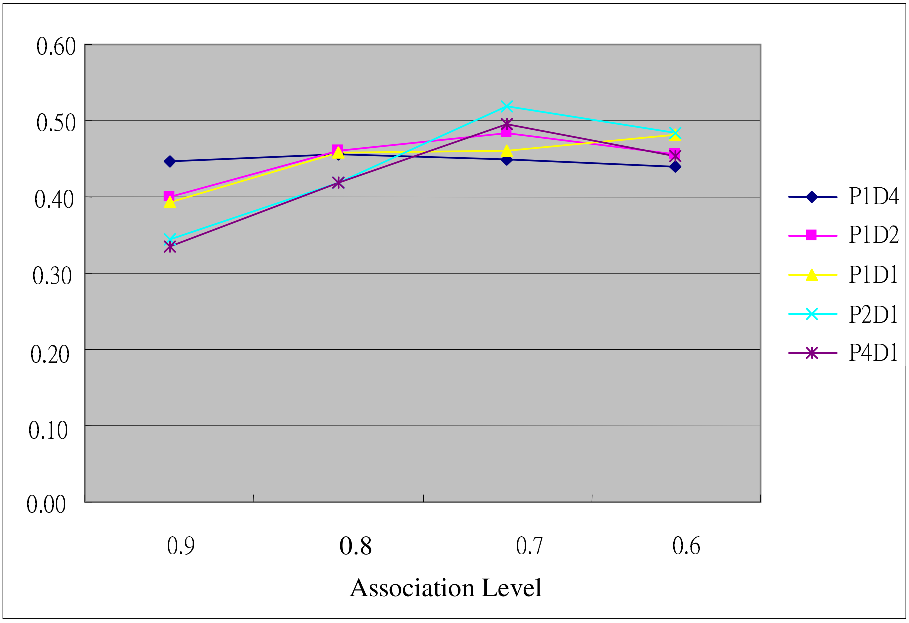


Fig. 1. The Effectiveness of Retrieval Using Implicit Link Modeling

4.5 Discussion

The best F-measure is 0.52, while recall and precision at 0.54, 0.50 respectively, and the association threshold for retrieval is set at 0.7 in P2D1. In other words, the proposed model can retrieve 54% of the cases committed by a criminal while maintaining a precision level at 50%. The best precision is 0.62 with recall at 0.23 in P4D1 and association level at 0.9. The best recall is 0.79 with precision at 0.30 in P1D4 and association level at 0.6. In terms of crime investigation, the implicit link modeled by this research can be used to establish good links to an unsolved case if it is indeed committed by the criminal.

5 Conclusions

This paper proposed and developed an association model for implicit crime link construction based on modus operandi and geographic relationship of crime. The model is based on statistic and probability distribution to compute association and establish links between criminals and crime cases. A data collection from a local police department is used to evaluate the proposed model. The experiment results shows that the proposed association model can be used to establish good implicit links to an unsolved case if it is indeed committed by the criminal.

Besides official structured criminal records, semi or unstructured data will also be added in future analysis and the model should also be expanded to include more variables, such as time of day, for better association modeling. Furthermore, subject evaluation will be conducted to validate the proposed approach.

References

1. Berry, M.J., Linoff, G.: Data Mining Techniques: For Marketing, Sales, and Customer Support. John Wiley & Sons, Chichester (1997)
2. Chen, H.C., Zeng, D., Atabakhsh, H., Wyzga, W., Schroeder, J.: COPLINK: Managing Law Enforcement Data and Knowledge. Communications of the ACM 46(1) (2003)
3. Chen, R.Z.: Locality Profiling and Its Application in Serial Street Robbery Investigation, Master Thesis, Department of Criminal Investigation, Central Police University Taiwan (2004)
4. Graycar, A.: Local Government and Crime Prevention. In: Proceeding of the character, Impact and Prevention of Crime in Regional Australia Conference, Townsville (2001)
5. Han, J., Kamber, M.: Data Mining: Concepts and Techniques, 2nd edn. Morgan Kaufmann, San Francisco (2006)
6. Hsu, C.C.: Criminology, 2nd edn. Central Police University, Taiwan (1996) (in Chinese)
7. Lee, H.C.: Crime Scene Investigation. Central Police University Press, Taoyuan (1994)
8. Palmiotto, M.: Crime Pattern Analysis: An Investigative Tool. In: Critical Issues in Criminal Investigation, 2nd edn., Pilgrimage (1988)
9. Peng, Y.T., Wang, J.H.: Link Analysis Based on Webpage Co-occurrence Mining – a Case Study on a Notorious Gang Leader in Taiwan. In: Proceedings of the IEEE International Conference on Intelligence and Security Informatics (2008)
10. Shannon, C.E., Weaver, W.: The Mathematical Theory of Communication. University of Illinois Press, Urbana (1949)
11. Salton, G.: Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison Wesley, Reading (1989)
12. Turvey, B.E.: Criminal Profiling: An Introduction to Behavioral Evidence Analysis, 3rd edn. Academic Press, London (April 30, 2008)
13. Wang, J.H., Lin, B.T., Lin, C.C.: Application of the Vector Space Model on Criminal Record Retrieval. In: Proceedings of the 1997 International Scientific Symposium on Police. Central Police University, Taiwan (1997)
14. Wasserman, S., Faust, K.: Social Network Analysis and Applications, New York and Cambridge (1994)
15. `http://en.wikipedia.org/wiki/Locard's_exchange_principle`

Design and Deployment of a National Detecting Stolen Vehicles Network System

Weiping Chang[1] and Chingwei Su[2]

[1] Department of Criminal Investigation, Central Police University
Kueishang, 333 Taiwan
una024@mail.cpu.edu.tw
[2] Information Office, National Police Administration
Taipei, 100 Taiwan
frank@npa.gov.tw

Abstract. This paper deals with the design and deployment of a National Detecting Stolen Vehicles Network System (NDSVNS) for a country and a police management authority. NDSVNS mainly uses the pattern recognition techniques to setup a Stolen Vehicle Identification System (SVIS), and integrates multiple heterogeneous databases of different city and county police departments. Each database functions as a stand-alone entity, but they are also connected by Police VPN. These heterogeneous databases include stolen vehicle data and license plate recognition records from 26 city and county police departments across this island. To share the valuable information, firstly Criminal Investigation Bureau (CIB) unified the data format of license plate recognition records and exchange formats on Police VPN. Secondly, some appropriate locations are selected in all cities and counties and toll stations of highways to install SVIS. In this paper, system design and deployment for a National Detecting Stolen Vehicles Network System is presented.

1 Introduction

One of the most challenging problems facing most countries is the increasing phenomenon of vehicle thefts. The impact of vehicle thefts results in property loss, inconvenience and in fear to the vehicle owners. Every country government regards fighting vehicle thefts as a very important mission [2, 7]. Many researchers applied technologies to combat vehicle thefts and proposed many prototypes [1, 4, 5, 6, 8, 10, 11]. However, vehicle thefts are growing at alarming rate around the world in recent years [7]. Increasing the possibility of detecting and finding stolen vehicles is very important to both police and vehicle owners.

The number of vehicle theft cases was dramatically growing in the 1990s in Taiwan, but from 2000 to 2009 the phenomenon has changed. After 2004, the number of vehicle theft cases was significantly decreasing. The number of vehicle theft cases reached the lowest point for the past 19 years in 2009 [9].

Taiwan police agencies have studied how to use IT to combat vehicle theft over the past decade. NDSVNS, National Detecting Stolen Vehicles Network System, was built in 2003 by CIB. NDSVNS mainly uses the pattern recognition techniques to setup

H. Chen et al. (Eds.): PAISI 2010, LNCS 6122, pp. 22–27, 2010.
© Springer-Verlag Berlin Heidelberg 2010

SVIS, Stolen Vehicle Identification System, and integrates multiple heterogeneous databases of different city and county police departments. They are all based on a Police VPN. Each database functions as a stand-alone entity, but they are also connected by Police VPN. These heterogeneous databases include stolen vehicle data and license plate recognition records from 26 city and county police departments across this island. To share the valuable information, firstly CIB established a common data format of license plate recognition records and a format for data exchange. Secondly, some appropriate locations are selected in all cities and counties and toll stations of highways to install SVIS, Stolen Vehicle Identification System. Stolen vehicle data from every police department are sent to CIB through the Police VPN. This enables CIB's Stolen Vehicle Database to always reflect current data. Furthermore, the license plate recognition records from all city and county police departments will also be sent to the centralized database at CIB via Police VPN. Therefore, different police departments could share the records of the Stolen Vehicle Database and the driving track records from license plate recognition record. NDSVNS improves the isolated system's shortcomings of not knowing what is occurring in neighboring communities by sharing data. When the stolen vehicles pass through the SVIS, the nearest police station will be notified and a police officer will come to arrest the thief CIB has expanded NDSVNS and SVIS systems and increased SVIS locations every year since 2004.

2 System Architecture

The architecture diagram of National Detecting Stolen Vehicles Network System is shown in Fig. 1. NDSVNS is built on Police VPN which links police agencies across the country. Stolen vehicle related information is sent and shared by all police departments via Police VPN. Police VPN also provides three basic requirements for information security; confidentiality, integrity and availability.

The system architecture of NDSVNS is described below:

2.1 NDSVNS Server

The NDSVNS server is installed in the CIB Computer Center and provides police officers of all police agencies with capability to add, query, and modify license plate information and other functions.

2.2 License Plate Recognition Workstation

The License Plate Recognition Workstation is responsible for license plate pattern recognition, and comparing suspected stolen vehicle information to the Stolen Vehicle Database (SVD). If the vehicle is reported as stolen in SVD, the system will display a warning message on the screen and automatically sound the alarm to inform the police on duty. The police may proceed to arrest the thief.

2.3 Camera

Cameras installed on poles along major roads in all cities and counties are responsible for capturing image data of vehicle license plates and sending data to the License Plate Recognition Workstation.

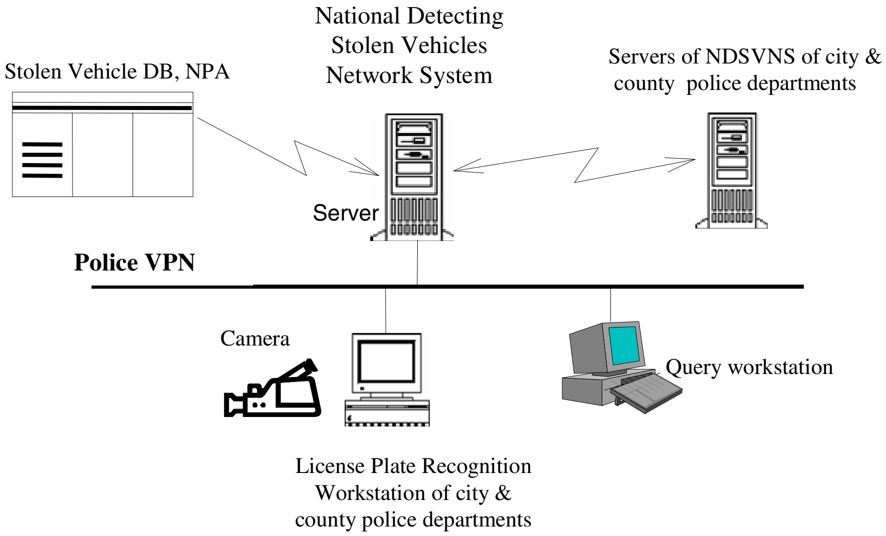


Fig. 1. Architecture diagram of National Detecting Stolen Vehicles Network System

3 Design of the System Blocks

According to the software functions, NDSVNS is divided into a Stolen Vehicle Data Management Website, a Data Conversion Subsystem, a License Plate Pattern

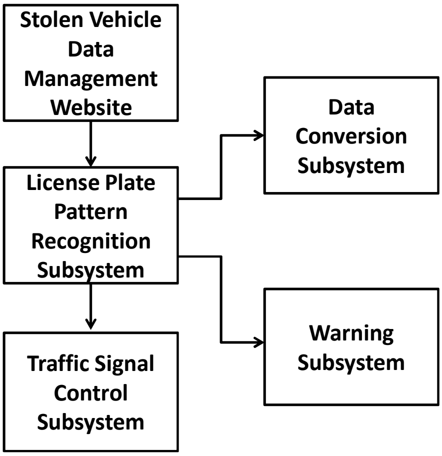

Stolen Vehicle Data Management Website
License Plate Pattern Recognition Subsystem
Data Conversion Subsystem
Warning Subsystem
Traffic Signal Control Subsystem

Fig. 2. Block diagram of National Detecting Stolen Vehicles Network System

Recognition Subsystem, a Warning Subsystem, and a Traffic Signal Control Subsystem, as Fig. 2.

The subsystems function as follows:

3.1 Stolen Vehicle Data Management Website

1) Data entry and updating the license plate data of stolen vehicle
2) Query of driving track records from license plate recognition record

3.2 License Plate Pattern Recognition Subsystem

License plate number is recognized and compared to Stolen Vehicle Database in 0.2 second. When a vehicle drives into the detectable area of the system, the system will automatically search the license plate image. The license plate image will be sent to the License Plate Recognition Workstation to be recognized. The recognizable vehicle speed is from 0 to 100 KM/H.

3.3 Traffic Signal Control Subsystem

In addition to sending a message to the Warning Subsystem to inform the police on duty, the system will also send signals to the Traffic Signal Control Subsystem. After receiving the signal, the Traffic Signal Control Subsystem will control and change the traffic light to red. Not knowing the traffic signals are controlled to enable an arrest, the thief is arrested and the stolen vehicle is seized by the police more safely.

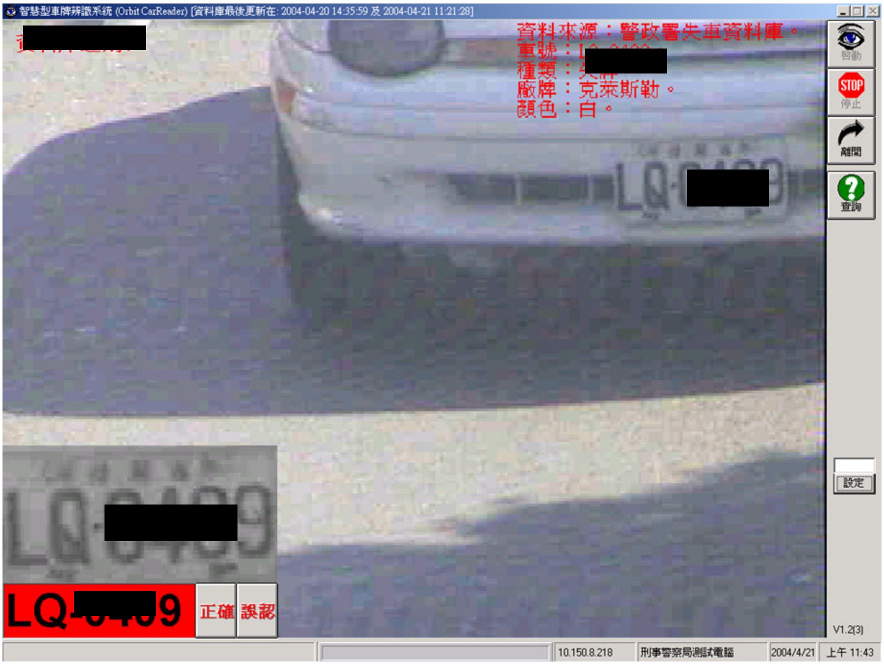
3.4 Data Conversion Subsystem

The Data Conversion Subsystem contains the capture and delivery function. It functions as follows:

1) Capture function: The Capture Function automatically captures and aggregates the plate recognition records, including plate number, date, and time, from all License Plate Recognition Workstation on a daily basis.
2) Delivery function: According to the schedule, delivery function automatically receives the data from Stolen Vehicle Database and distributes it to each License Plate Recognition Workstation.

3.5 Warning Subsystem

When the system detects a stolen vehicle, the system will automatically display a warning message on the screen and inform the police on duty via voice in Chinese, as Fig. 3. The police may then proceed to make an arrest.



智慧型車牌辨識系統 (Orbit CarReader) [資料庫最後更新在: 2004-04-20 14:35:59 及 2004-04-21 11:21:28]

[redacted]

資料來源：警政署失車資料庫。
車號：[redacted]
種類：
廠牌：克萊斯勒。
顏色：白。

[redacted]

[redacted]

LQ[redacted]9

Fig. 3. Warning message of a stolen vehicle being detected and informed the police

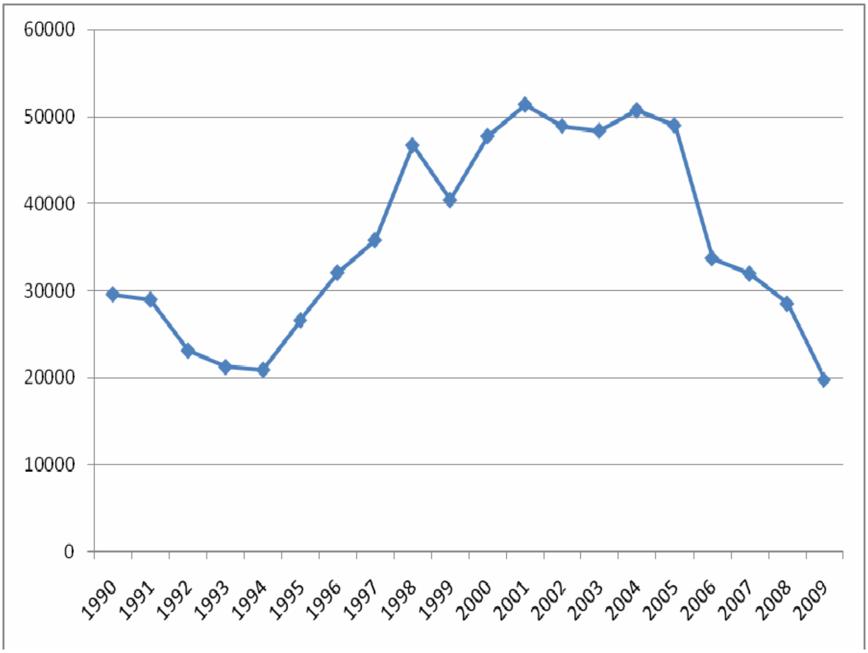


Fig. 4. The number of vehicle theft cases from 1990 to 2009 in Taiwan

4 Conclusion

In this paper, a National Detecting Stolen Vehicles Network System is designed and deployed. The system is based on Police VPN to share stolen vehicle related data between city and county police departments in Taiwan. From 1990 to 1999, the number of vehicle theft cases increased from 29,587 to 40,410, for a growth rate of 36.58% in 10 years. To fight against the high growth rate of vehicle theft cases, Taiwan police applied IT to detect stolen vehicles, developed and setup the National Detecting Stolen Vehicles Network System in 2003. In 2004, the number of vehicle theft cases was significantly decreasing. The NDSVNS successfully detected 1,139 stolen vehicles in 2004 [3]. The number of vehicle theft cases reached the lowest point for the past 19 years in 2009. The number of vehicle thefts cases dropped to 19,755 in 2009 from 47,750 in 2000 [9], with a growth rate of -58.63%, as shown in Fig. 4. The application of IT by Taiwan is one of the leading strategies for fighting against vehicle theft. The success experienced in Taiwan to combat vehicle thefts is worthy of being recommended to other Asian country police.

References

1. Al-Hmouz, R., Challa, S.: Intelligent Stolen Vehicle Detection using Video Sensing. Information, Decision and Control. In: IDC ’07, pp. 302–307 (2007)
2. Arizona Criminal Justice Commission: Arizona Auto Theft Study. Arizona Criminal Justice Commission (2004)
3. Chang, W., Su, C.: Benefit Analysis of Stolen Vehicle Detecting System. In: Proceedings of 9th Information Management Research and Police Application Conference, pp. 215–222 (2005)
4. Cherbonneau, M., Copes, H.: ’Drive It Like You Stole It’: Auto Theft and the Illusion of Normalcy. The British Journal of Criminology 46(2), 193–211 (2006)
5. Global Design News: Centralized network slows auto theft. Global Design News (2001)
6. Kursun, O., Reynolds, K., Eaglin, R., Chen, B., Georgiopoulos, M.: Development of an artificial intelligence system for detection and visualization of auto theft recovery patterns. In: Proceedings of the 2005 IEEE International Conference on Digital Object Identifier, pp. 25–29 (2005)
7. Michigan Automobile Theft Prevention Authority: 2002 ATPA Annual Report. Michigan Automobile Theft Prevention Authority (2002)
8. Mustafa, M., Behnam, M., El-Tarhuni, M.: A Wireless Embedded System for the Tracking of Stolen Vehicles. In: IEEE/ACS International Conference on Computer Systems and Applications, pp. 818–825 (2006)
9. National Police Administration: Criminal Cases Statistics. National Police Administration (2009)
10. Nagaraja, B., Rayappa, R., Mahesh, M., Patil, C., Manjunath, T.: Design & Development of a GSM Based Vehicle Theft Control System. In: 2009 International Conference on Digital Object Identifier, pp. 148–152 (2009)
11. Wang, J.: Study of a conformal hidden wire antenna used for the detection of stolen cars. In: Electromagnetic Compatibility, Proceedings of the 2008 IEEE International Symposium on Digital Object Identifier, pp. 1–6 (2008)

Fighting Cybercrime: A KM Perspective

Weiping Chang

Department of Criminal Investigation, Central Police University
Kueishang, 333 Taiwan
una024@mail.cpu.edu.tw

Abstract. Cybercrime is one of the most difficult challenges facing law enforcement agencies. The police need to learn vast skills and accumulate much knowledge to fight against cybercrime. Because law enforcement agencies possess knowledge to fight against cybercrime in various forms and formats, knowledge management (KM) is needed. In this paper, the author applies KM into cybercrime investigation to manage cybercrime knowledge.

1 Introduction and Related Work

Cybercrime is one of the most difficult challenges to every government. Law enforcement agencies invest enormous resources to fight against cybercrime, but the effectiveness is limited. This is because cybercrime is different from conventional crime. Conventional criminal investigation skills and knowledge could not successfully combat cybercrime. Many scholars have studied how to combat cybercrime. They proposed many tools [2, 3]. However, to solve the cybercrime cases, the police need to learn more skills and knowledge, such as criminal investigation stages [9], Criminal Code, intrusion detection, computer forensics [5], etc. To manage the vast knowledge well, KM is needed.

2 KM Perspective

The recommended procedure for managing cybercrime investigation knowledge is:

1) Determining knowledge needs (KN) for the general crime investigation (GCI) [7, 9].
2) Determining knowledge needs for the general cybercrime investigation (GCCI), such as computer forensics [5, 11], wireless and VoIP forensics [8], etc.
3) Determining knowledge needs for each type of cybercrime investigation (CCI), such as Phishing [1].
4) Assessing 1, 2, and 3 knowledge needs (KNs) and classifying them [6].
5) Identifying agency's cybercrime investigation knowledge (CCIK) [10].
6) Finding the knowledge gap [13] of cybercrime investigation between CCIK and KNs.
7) Filling the knowledge gap of cybercrime investigation, such as recruiting experts, training, and outsourcing depends on types of knowledge [13].

The workflow chart of managing cybercrime investigation knowledge is as Fig. 1.

H. Chen et al. (Eds.): PAISI 2010, LNCS 6122, pp. 28–30, 2010.
© Springer-Verlag Berlin Heidelberg 2010

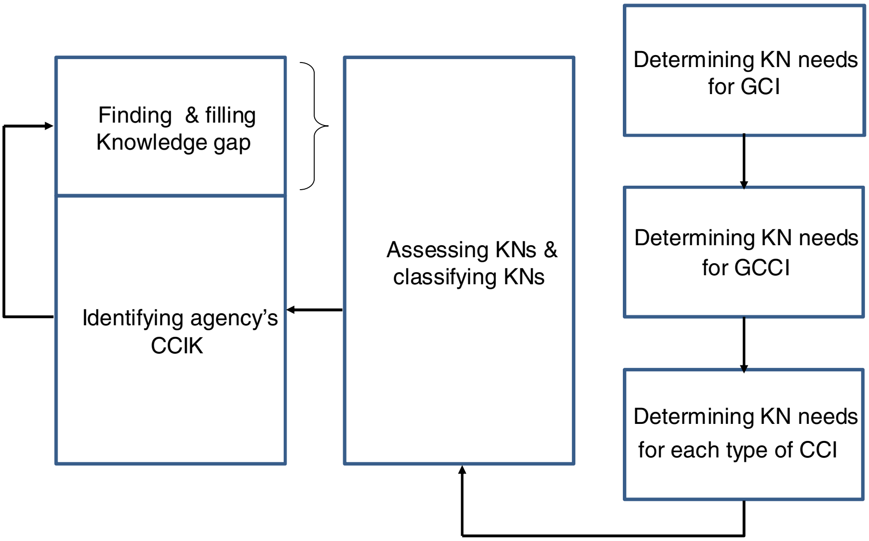

Finding & filling Knowledge gap
Identifying agency's CCIK
Assessing KNs & classifying KNs
Determining KN needs for GCI
Determining KN needs for GCCI
Determining KN needs for each type of CCI

Fig. 1. The workflow chart of managing cybercrime investigation knowledge

3 Future Work

The procedure of managing cybercrime investigation knowledge proposed in this paper provides the capabilities to law enforcement agencies. However, in addition to the procedure of managing cybercrime investigation knowledge, law enforcement agencies must also establish a good strategy, create a culture of knowledge sharing, and apply information technology (IT) to setup a knowledge sharing website for knowledge management of cybercrime investigations. When everyone is willing to share cybercrime investigation knowledge and law enforcement agencies manage cybercrime investigation knowledge well, law enforcement will win against cybercrime.

References

1. Birk, D., Gajek, S., Grobert, F., Sadeghi, A.: Phishing Phishers - Observing and Tracing Organized Cybercrime. In: Second International Conference on Internet Monitoring and Protection, pp. 3–8 (2007)
2. Chen, H., Zeng, D., Atabakhsh, H., Wyzga, W., Schroeder, J.: COPLINK- Managing Law Enforcement Data And Knowledge. Communication of the ACM 46(1), 28–34 (2003)
3. Donalds, C., Osei-Bryson, K.: Criminal Investigation Knowledge System: CRIKS. In: Proceedings of the 39th Hawaii International Conference on System Sciences, pp. 155–164 (2006)
4. Nordin, M., Pauleen, D., Gorman, G.: Investigating KM antecedents: KM in the criminal justice system. Journal of Knowledge Management 13(2), 4–20 (2009)
5. Park, H., Cho, S., Kwon, H.: Cyber Forensics Ontology for Cyber Criminal Investigation. LNICST, vol. 8, pp. 160–165 (2009)

6. Polanyi, M.: Personal Knowledge: Towards a Post-Critical Philosophy. University of Chicago Press, Chicago (1958)
7. Ratcliffe, J.H.: Crime mapping and the training needs of law enforcement. European Journal on Criminal Policy and Research 10(1), 65–83 (2004)
8. Stephens, P., Induruwa, A.: Cybercrime Investigation Training and Specialist Education for the European Union. In: Digital Forensics and Incident Analysis, WDFIA 2007, pp. 28–37 (2007)
9. Swanson, C.R., Chamelin, N.C., Territo, L.: Criminal Investigation, 8th edn. McGraw-Hill, Boston (2003)
10. Yang, A., Liu, A.: The Importance of Knowledge Identification in Developing Organizational Core Competences. In: Proceedings of the 13th Asia Pacific Management Conference, Melbourne, Australia, pp. 1005–1014 (2007)
11. Yeager, R.: Criminal Computer Forensics Management. In: InfoSecCD Conference'06, Kennesaw, GA, USA, pp. 168–174 (2006)
12. Zack, M.: Developing a Knowledge Strategy. In: Choo, C., Bontis, N. (eds.) The Strategic Management of Intellectual Capital and Organizational Knowledge. Oxford University Press, New York (2002)

Trajectory Similarity of Network Constrained Moving Objects and Applications to Traffic Security

Sajimon Abraham[1] and Paulose Sojan Lal[2]

[1] School of Management & Business Studies, Mahatma Gandhi University, Kerala, India
[2] Mar Beselios Institute of Technology and Science, Kothamangalam, Kerala, India-686 693
sajimabraham@rediffmail.com, padikkakudy@gmail.com

Abstract. Spatio-Temporal data analysis plays a central role in many security-related applications including those relevant to transportation infrastructure, border and inland security. In several applications, data objects move on pre-defined spatial networks such as road segments, railways, and invisible air routes, which provides the possibility of representing the data in reduced dimension. This dimensionality reduction gives additional advantages in spatio-temporal data management like indexing, query processing, similarity and clustering of trajectory data etc. There are many proposals concerning trajectory similarity problem which includes Euclidian, network, time based measures and concepts known as Position of Interest(POI), Time of Interest(TOI) etc. This paper demonstrates how these POI and TOI methods could be advantages in security informatics domain suitable to work with road network constrained moving object data, stored using a binary encoding scheme proposed in a previous PAISI paper.

Keywords: Moving objects on road networks, spatio-temporal databases, traffic security, spatio-temporal distance measure, trajectory similarity.

1 Introduction

Trajectory database management has emerged due to the profusion of mobile devices and positioning technologies like GPS or recently the RFID (Radio Frequency Identification). Studying people and vehicle movements within some road network is both interesting and useful especially if we could understand, manage and predict the traffic phenomenon. In essence, by studying the mass flow of traffic data we can monitor the traffic flow and discover traffic related patterns. Intelligent Transportation Systems (ITS) have been developed allowing better monitoring and control of traffic in order to optimize traffic flow. ITS collect data utilizing various technologies such as sensors, live cameras and cellular phone data and in turn, they propose reroute of traffic to avoid congestion. Beyond transportation perspective, information technology provide a variety of applications and technologies for gathering, storing, analyzing and providing access to traffic data.

In order to search moving object trajectories, some methods of existing research have been proposed in Euclidean space. However, Euclidean distance is not appropriate for road network space defined along a road. We investigate several differences

H. Chen et al. (Eds.): PAISI 2010, LNCS 6122, pp. 31–43, 2010.
© Springer-Verlag Berlin Heidelberg 2010

between Euclidean space and road network space. First, while moving object trajectories in Euclidean space are expressed to a sequence of points in (*x*,*y*,*t*) space, those of road networks are represented as a set of (l*oc* ,*t*),where loc is a road sector identifier. Therefore, the distance between two points is calculated more easily on road networks defined along road sectors than Euclidean space. Second, moving object trajectories in Euclidean space have a linear interpolation problem.

The challenge is to express trajectory similarity by respecting network constraints, which is also a strong motivation for the following real and practical applications in field of security Informatics.

i. By identifying similar trajectories, effective data mining techniques (e.g., clustering) can be applied to discover useful patterns. For example, a dense cluster is an indication of emerge traffic measures, future road expansions, traffic jam detection, traffic predictions, etc.

ii. Trajectory similarity can also help in several road network applications such as, routing applications which support historical trajectories, logistic applications, city emergency handling, drive guiding systems, flow analysis, etc. In such applications, efficient indexing and query processing techniques are required.

iii. Knowledge and prediction of the road traffic: Given that the number of vehicles increases on the roads, information related to the density on the network becomes very useful for many purposes as navigation, trip planning, etc.

iv. Car-sharing: In these last years, the massive use of the private means of transport caused many problems, namely the pollution and also the raising of oil prices. Car-sharing appears as an interesting alternative for the security of the persons traveling since a quiet lot of problems in lonely travel could be avoided. Identifying the similar trajectories or even sub-trajectories becomes very useful for such types of applications.

v. Transport planning: At the moment of its creation, each road is planned for certain utilization. Reporting trajectory groups allows assessing the suitability of the road infrastructure with its actual use.

vi. As dark web is the concept of terrorist web site activities, trajectory can be treated as the path formed by series of web clicks. Trajectory similarity of moving objects resembles path similarity of user click-streams in the area of web usage mining. By analyzing the URL path of each user, we are able to determine paths that are very similar, and therefore effective caching strategies can be applied.

In particular trajectory similarity problem has specific applications in the area of security informatics which has been identified as an additional extension of the paper [1] which deals with the concepts of automatic security alarming schemes.

vii. The paper suggests in giving security alarm trigger to vehicles entering into an emergency area. There is no direction further to these vehicles about how to go further or to deviate to which direction or to stop the movement at all. To give better suggestion to the triggered vehicles the system has to know the traffic patterns of the alternate routes. Clustering of trajectories in these routes at specified time period will give traffic patterns and the system can automatically suggest the possible way to go further.

viii. We are proposing Clustering algorithm for Interested Points on the roads to find the sequence of vehicles which typically choose these routes and also at Interested time also.

ix. Since we are using a binary coded data for representing road location using which we can easily classify districts, sectors, roads etc , the section wise query processing can be made more simple by excluding the unnecessary data at the first scan itself.

The rest of this paper is organized as follows. In section 2, the related works in the problem area of trajectory similarity is discussed. A brief description on Binary Encoding Method to store Roads and Relative Locations, proposed in the author's previous PAISI 2008 paper is given in section 3. In section 4 the various similarity measures with algorithms discussed in detail. Experimental study with real data set and the results were analyzed in section 5. Section 6 concludes the paper with a note on future directions to be carried out.

2 Related Works

Regarding the works related to the similarity of moving objects trajectories, we first mention those in the free moving trajectory context and then for constrained trajectories. Yanagiswa et al. [2] focused on the extraction of the individual moving patterns of each object from the trajectories considering both time and location. Their approach uses the shape similarity between lines to retrieve required objects. Shim and Chang [7] considered the similarity of sub-trajectories and proposed a distance 'K - Warping' algorithm. We also find similar approaches in Valachos et al. [8]), Sakurai et al. [10], and Chen et al. [11]. Valachos et al. [9] presented an investigation for analysis of spatio-temporal trajectories for moving objects where data contain a great amount of outliers. Therefore, they propose the use of a non metric distance function that is based on the Longest Common Sub Sequences (LCSS) algorithm in conjunction with a Sigmoid Matching function to increase the performance of Euclidean and Time Warping Distance. Zeinalipour- Yazti et al. [12] introduce a distributed spatio-temporal similarity search based on the LCSS distance measure and propose two new algorithms offering good performances.

All these methods are inappropriate for similarity calculation on road networks since they use the Euclidian distance as a basis rather than the real distance on the road network. This point has motivated the proposition of Hwang et al. [3][4] that were the first to propose a similarity measure based on the spatiotemporal distance between two trajectories using the network distance. The algorithm of similar trajectory search consists of two steps: a filtering phase based on the spatial similarity on the road network, and a refinement phase for discovering similar trajectories based on temporal distance. This work seems to be the first research work studying trajectory similarity on networks where the authors propose a simple similarity measure based on Points Of Interest (POI) and Time Of Interest(TOI). They retrieve similar trajectories on road network spaces and not in Euclidean spaces. In our research we discuss modifications of the algorithms proposed in Hwang research work. Tiakas et al. [13]

and Chang et al. [14] also use the same spatiotemporal distance, based on the road network, in their algorithm of similar trajectory search. One of the recent works in trajectory similarity problem for network constrained objects can be found in [5]. The paper introduces new similarity measures that should be employed to express similarity between two trajectories that do not necessarily share any common sub-path. They define new similarity measures based on spatial and temporal characteristics of trajectories, such that the notion of similarity in space and time is well expressed, and moreover they satisfy the metric properties. In addition, it demonstrates the similarity range queries in trajectories that are efficiently supported by utilizing metric-based access methods, such as M-trees.

3 A Brief Description on Binary Encoding Method to Store Roads and Relative Locations

As in the paper[1] here also we assume that the database stores the complete history of moving objects through time and must answer queries about any time in the history of objects. We assume that each database record has the format (oid, location, time*),* where oid identifies an object, location is the spatial coordinates(x,y) represented as binary string, and time indicates the time in which the object remained at position (x,y). A typical domain where such a model fits is mobile device tracking, e.g., of GPS, PDA, or wireless phone devices. The model reflects a real-world application constraint where assuming an object follows a linear trajectory between data points may lead to incorrect, and unacceptable, assumptions. For example, in security/monitoring applications, a person could be mistakenly assumed to have entered a restricted area, instead of gone around it, because his/her movement was interpolated. Our model can be viewed as a step-wise interpolation instead of a linear interpolation. That is, as long as the object's position is not updated in the database it is assumed to remain stationary in its last observed position.

In conventional approaches, the location information of moving objects were expressed as a geometric coordinate (x,y) in two-dimensional space. However, instead, [1] propose to express location information using both hierarchical administrative district and road network in one-dimensional space that fits real world better. For instance, if a moving object is in a building at a coordinate of latitude = 125.58 and longitude = -37.34, then it can be expressed as a set of fields according to an administrative district such as city, road-name, road-block (e.g., Seoul, Main road, 165th block). Furthermore, by converting the fields into a binary string that has efficient ways to process queries, we focus the following advantages.

(i) Storage cost can be reduced as the proposed scheme requires to store one-dimensional data against multi-dimensional. (ii) The complexity in managing large scale multi-dimensional spatio-temporal data for indexing and query processing can be simplified. (iii) In real world, moving objects can only follow along the ‘‘roads’’. However, if one expresses location information as geometric coordinates, then one may include spaces where moving objects can never move into, so-called dead space, incurring storage waste. (iv) Since the location information is specified in binary code, entire district or road-block can be easily addressed based on number of bits.

The dimension reduction of spatio-temporal data management [16] discusses two algorithms for binary encoding process, one for administrative district, road and location encoding and the second, for converting a position represented as geometric co-ordinate into an equivalent binary string. Since the proposed alarming scheme is based on this encoding method these basic algorithms are briefly discussed below.

The paper proposes two algorithms which describe how an administrative district or a road location can be represented and stored as a set of binary string. The method is a recursive procedure which will successively divide the entire region into sub-regions and finally map each district into a two-dimensional space and then assign a binary string to each district. To provide the relative position of districts the mapping is based on space-filling curves such as Z-ordering[15]. Another algorithm in the paper will fix the position of a moving object represented by geometric co-ordinates (x,y) as usually supplied by the GPS, onto roads. The algorithm uses an R-tree for roads to quickly convert a two-dimensional point into equivalent binary string.

Experimental validation proved that the overhead in converting (x,y) into binary string will be negligible for a typical LBS environment and tolerable even for an environment with a huge number of locations. The total time spent for the conversion increases almost linearly with the number of two-dimensional locations and more than seven million locations can be processed per second.

The characteristics of the binary encoding scheme which made base line for the proposed alarming scheme in [16] are (i) It will be easy to find out the lowest common administrative district by extracting the longest common prefix of a given set of binary strings, and (ii) a district containing a set of lower districts can be represented by the range of binary strings; for example, county ''A'' will be represented by the range [00000, 00111]. These advantages make it easy in addressing the whole country, whole district or of a single road by identifying the common prefix in the binary string representing the location of the object. For example if the object's location on the road encoded as 001010000110010 where first 2 bits for the country, 4 bits for the district, 5 bits for road and 4 bits for relative location on the road. Then in all locations if first 2 bits are 00 then they all belongs to the same country.

In the proposed scheme, moving object data base is a relational database , where the base table contains the details of the moving objects currently on roads under consideration. Its structure contains data as given below.

Structure of Base Table

Ob-Id	Location(Bit String)	Time
O1	001010000110010	10:00
O2	001010000110011	10:01
O3	001010000100010	10:02
O2	001010000110111	10:10
O1	001010000110011	10:15

The object will update this table at frequent time intervals on the assumption that the system will have a continuous network connectivity and sufficient capacity to handle large amount of data. Our proposed methods will work on this table.

4 Trajectory Similarity in Spatial Networks

The common characteristic of the aforementioned approaches and research works is that objects are allowed to move freely in 2D or 3D space, without any motion restrictions. However, in a large number of applications, objects are allowed to move only on pre-defined paths of an underlying network, resulting in constraint motion. For example, vehicles in a city can only move on road segments. In such a case, the Euclidean distance between two moving objects does not reflect their real distance. Objects moving in a spatial network follow specific paths determined by the graph topology, and therefore arbitrary motion is prohibited. This means that two trajectories which are similar regarding the Euclidean distance may be dissimilar when the network distance is considered. The majority of existing methods for trajectory similarity assume that objects can move anywhere in the underlying space, and therefore do not support motion constraints. Most of the proposals are inspired by the time series case, and provide translation invariance, which is not always meaningful in the case of spatial networks.

4.1 Finding Similar Trajectories

Let T be a set of trajectories in a spatial network, in which each trajectory is represented as

$$T = ((b_1,t_1),(b_2,t_2),(b_3,t_3),\ldots\ldots\ldots\ldots,(b_n,t_n))$$

where n is the trajectory description length, bi denotes a location in binary string and t_i is the time instance (expressed in time units, e.g. seconds) that the moving object reached node bi, and $t_1 < t_i < t_m$, for each $1 < i < m$. It is assumed that moving from a node to another comes at a non-zero cost, since at least a small amount of time will be required for the transition.

We identify the following similarity types related to road network environment.

(i) **Finding Objects moving through certain Points of Interest:** This will be useful in finding out the movements of objects through known locations of interest, which may be terrorist locations, points of emergency or list of strategically important locations.
(ii) **Finding Objects moving through Certain Times of Interest:** Useful for objects moving through known times of interest, which may be emergency time, explosion happened time or festival season time etc.
(iii) **Finding Objects moving through certain Points of Interest and Time of Interest:** Here we consider both spatial and temporal features together coming as spatio-temporal distance measure.
(iv) **Finding objects moving similar routes based network distance measure:** To find out the traffic congestion in a particular route useful in planning new transport schedules or reducing existing one.
(v) **Finding objects whose distance differed by Euclidian measure:** To locate objects affected by a Bomb blast, disturbance of Pollution like sound, wind, infectious diseases in which the effect spread in Arial distance.

In each of these methods we use existing algorithms already available in the literature and each of them is modified to suit our binary encoding method and also applicable to security Informatics domain. We will discuss the first three methods based on [3][4] with its modifications in the following section. Methods (iv) and (v) can be explored using the concept discussed in [5] suitable for security informatics and we set apart it for future work.

(i) **Finding Objects moving through certain Points of Interest**

Here we use the method proposed in [3] which is briefly described as follows.

Most previous methods considered only spatial similarity in measuring the similarity between moving object trajectories. For example, if two trajectories pass through the same points at different time intervals on road networks, we understand by spatio-temporal intuition that they are not similar to each other. However, previous methods asserted that two trajectories are similar to each other. To solve these problems concerning previous methods, we define spatial and temporal similarity based on road networks. In general, moving objects on road networks are represented as locations and times obtained by GPS. We are interested in the movement of objects through selected locations or times and we consider POI (Points Of Interest) on road networks.

The paper suggest a two step process of filtering trajectories based on spatial similarity and then refining similar trajectories based on temporal distance. For this the following two definitions are given.

Definition 1. Spatial Similarity between Trajectories on road network space Suppose that P is a set of POI's on a given road networks. Then spatial similarity between two trajectories TR_A and TR_B is defined as

$$Sim_{POI}(TR_A; TR_B; P) = \begin{matrix} 1 \text{ if } \forall p \text{ in } P;\ p \text{ is on } TR_A \text{ and } TR_B \\ 0;\ \text{otherwise} \end{matrix}$$

Temporal similarity can be defined as the inverse of temporal distance. When a POI is given, as the difference between the times two objects passed the same POI as follows:

Definition 2. Temporal Distance between Trajectories for one POI. Suppose that p Є P, and P is the set of POI. Then the temporal distance between two trajectories TR_A and TR_B is

$$dist_T(TR_A; TR_B; p) = |t(TR_A, p) - t(TR_B, p)|$$

If neither TR_A nor TR_B pass through p, the temporal distance is considered as infinity.

If we consider $t(TR; p_i)$ as the time the i^{th} POI, was passed each trajectory, TR, is plotted as a point $t(TR) = (t(TR, p_1), t(TR, p_2), \ldots\ldots\ldots, t(TR, p_k))$ in a k-dimensional space where k is the number of POIs. Then the temporal distance between two trajectories for a set of POIs is defined as the L_P distance of this k-dimensional space as follows:

Definition 3. Temporal Distance between Trajectories for a set of POIs.

Suppose that P is a set of POI and TR_A and TR_B are two trajectories. Then the temporal distance between TR_A and TR_B is

$$dist_T\ (TR_A, TR_B,P) = Lp(TR_A, TR_B,p) =(\sum_{i=1}^{k}(|pi(TR_A)-pi(TR_B)|^{p})^{1/p}$$

Using above two definitions the algorithm for searching similar trajectories is explained as below as a two stage process. In the first stage, which is a filtering process, the algorithm proposed in [3] has the following draw back.

The similarity in space with definition (1 = similar, 0 = dissimilar) does not take into account any notion of similarity percentage or similarity range. Therefore, we cannot determine how similar two trajectories are in space. We have modified this by introducing a threshold (threshold-s in algorithm 1) to determine how similar the trajectories with reference to how many points it pass through and then this measure is used to filter based on spatial distance. So all trajectories which satisfies this threshold will be considered in the next stage of refinement. The binary encoded location data will have the advantage of an initial filtering stage based on administrative district kind of search or road kind of search when the POI's are within that constrained area.

Algorithm 1. Searching Similar Trajectories moving through certain Points of Interest; Input: Input trajectories TR_{IN}, threshold-s ρ, threshold-t δ , query trajectory tr_Q, POI set P; Output: similar trajectories TR_{OUT}

```
Begin
   $TR_{Candidate}$= φ
   $TR_{OUT}$  = φ
   n= number of Points in POI
   For each $t_r$ in $TR_{IN}$,
     tr.k=0
     For each p in P
       If p is on $t_r$ then
          tr.k = tr.k+1
     End For
     If (tr.k/n)> ρ then
       $TR_{Candidate}$ =$TR_{Candidate}$ U{$t_r$}
    End For
   For each $t_r$ Є $TR_{Candidate}$
    If $dist_T$  ($tr_Q$, $t_r$, P) < δ then
      $TR_{OUT}$ = $TR_{OUT}$ U{$t_r$}
   End For
   return $TR_{OUT}$
End
```

(ii) **Finding Objects moving through Certain Times of Interest**

In terms of practical application, the meaning of distance between two time intervals can rarely be found. However, we are interested in time intervals of moving objects, TOI (Times of interest) is an important characteristic of road networks. If trajectories pass the same points at the same TOI on road networks, we consider that they are similar to each other. Therefore, we define temporal similarity based on TOI. For example, the heaviest traffic time intervals on a specific road network can be TOI. We filter trajectories using this definition. If two trajectories pass through the same TOI, they are considered similar by the following definition:

Definition 4. Temporal Similarity between Trajectories on Road Networks.

Suppose that T is a set of TOIs on a given road networks. Then, temporal similarity between two trajectories TR_A and TR_B is defined as

$$\mathrm{Sim}_{\mathrm{TOI}}\,(TR_A, TR_B, \mathrm{T}) = \begin{cases} 1 & \text{if } \forall t \text{ in T, } t \in [t_s(TR_A), t_e(TR_A)] \;\&\&\; t \in [t_s(TR_B), t_e(TR_B)] \\ 0; & \text{otherwise} \end{cases}$$

Definition 5. Spatial Distance between Trajectories.

Suppose that t Є T, and T is the set of TOIs. Then the spatial distance between two trajectories TR_A and TR_B is defined as

$$dist_s\,(TR_A, TR_B, T) = \Sigma dist_s(p(TR_A, t_i), p(TR_B, t_i))$$

Algorithm 2. Objects moving through Certain Times of Interest Searching based on Temporal Filter and Spatial Distance

Input:Input trajectories $\mathrm{TR}_{\mathrm{IN}}$, threshold-s ρ, threshold-t δ, query trajectory tr_{Q}, TOI set T,time interval t; Output: similar trajectories $\mathrm{TR}_{\mathrm{OUT}}$

```
Begin
   $TR_{Candidate}$ = φ
   $TR_{OUT}$ =φ
   nt= number of Time Points in TOI
     For each $t_r$ in $TR_{IN}$,
      set ts= $t_r$.t∩ $tr_Q$.t
      If (tr.k/n)> ρ then
        $TR_{Candidate}$ = $TR_{Candidate}$ U { $t_r$}
     End for
   For each $t_r$ Є $TR_{Candidate}$
    If $dist_s$ ($tr_Q$, $t_r$, T) < δ then
      $TR_{OUT}$ = $TR_{OUT}$ U{$t_r$}
   End For
   return $TR_{OUT}$
End
```

(iii) **Finding Objects moving through certain Points of Interest and Time of Interest.**

In earlier two methods we follow two stage process with initial filtering and then refinement. In this Method, spatial and temporal similarity together in the filtering step. Afterwards, we refine similar trajectories using spatio-temporal distance based on POI and TOI. We regard spatio-temporal distance as the sum of temporal distance and spatial distance, which is defined as follows:

Definition 6. Spatio-Temporal Distance between Trajectories

Suppose that TR_A and TR_B are two trajectories. Then the spatio-temporal distance between TR_A and TR_B is

$$\text{dist}_{ST}(TR_A, TR_B) = \text{dist}_T(TR_A, TR_B) + \text{dist}_S(TR_A, TR_B)$$

To use this definition, the equivalence between temporal distance and spatial distance is defined so that 1 second = m meters. Moving objects on road networks move with various speeds. With this observation, we solve the equivalence problem between temporal distance and spatial distance using the speed of moving objects. That is, the equivalence problem between temporal distance and spatial distance is solved by the following formula.

$$\text{Convt}_S(TR_A, TR_B) = | (V_{TRA} - V_{TRB}) | * \text{dist}_T(TR_A, TR_B)$$

The above formula converts temporal distance into spatial distance. Applying this formula to definition 5, the spatio-temporal distance between two trajectories is defined as follows:

$$\text{dist}_{ST}(TR_A, TR_B) = \text{Convt}_S(TR_A, TR_B) + \text{dist}_S(TR_A, TR_B)$$

Algorithm 3. Searching based on Spatio-Temporal Filter and Spatio-Temporal Distance
Input: Input trajectories TR_{IN}, threshold-s1 ρ1, threshold-t1 δ1 , threshold-s2 ρ2, threshold-t2 δ2, query trajectory $tr_{Q,}$ POI set *P*, TOI set *T*, time interval *t*
Output: Similar trajectories TR_{OUT}

```
Begin
     $TR_{Candidate}$=φ
     $TR_{OUT}$ = φ
     np= number of Locations in POI
     nt= number of Time Points in TOI
     For each $t_r$ in $TR_{IN}$,
        tr.k=0;
        For each p in P
         If p is on $t_r$ then
          $t_r$.k=$t_r$.k+1
        End for
        set $t_r$.s= $t_r$.t∩$tr_Q$.t
        If ($t_r$.k/np)>ρ1 and($t_r$.s/nt)>δ1then
        $TR_{Candidate}$ = $TR_{Candidate}$ U{$t_r$}
     End for
     For each $t_r$ ∈ $TR_{Candidate}$
      If (dist$_T$($tr_Q$,$t_r$,P)<ρ2)and(dist$_s$($tr_Q$,$t_r$,T)<δ2)then
       $TR_{OUT}$ =$TR_{OUT}$ U{$t_r$}
     End For
     return $TR_{OUT}$
End
```

Advantages of threshold-s1 ρ1, threshold-t1 δ1 are that we can see the degree of similarity if the trajectories does not passes through all points in POI and all time Points in TOI. This measure could be used later for trajectory clustering purpose.

In all the above algorithms the advantages with binary encoding is that we can do initial filtering by using binary code component of the district or road. Thus a large number of trajectories could be pruned out at the initial stage itself when the query requirement is restricted to a district or road.

5 Experimental Evaluation

We have taken a real-world data set for experimentation purposes, namely a fleet of trucks available in [6] trajectory data set. The data set consists of 276 trajectories. From this original dataset we exported a subset of 20 trajectories belonging to two district clusters. Trucks dataset consists of 50 trucks delivering concrete to several construction places around Athens metropolitan area in Greece for 33 distinct days. The structure of each record is as follows:

{obj-id, traj-id, date(dd/mm/yyyy), time(hh:mm:ss), lat, lon, x, y}, where (lat, lon) is in WGS84 reference system and (x, y) is in GGRS87 reference system.

The average search time in query processing for our three methods are compared with number of POI's. In the experimental evaluation done in [4], method 2 involves so many trajectories including meaningless trajectories at the filtering stage as the time interval of a query trajectory is smaller than the total life span for all moving objects. But in our experiments we could pruned out large number of unwanted trajectories when POI or TOI has taken in a specific district or road. Method 2 takes less time when number of POI is less. But it is not higher as in the experimentation [4]. As we could do a preliminary filtering based on administrative district or road the average search time in all the three methods are less compared to the earlier implementation. Also the additional thresholds used in the filtering step provides the possibility of clustering the trajectories for future data mining applications.

6 Conclusion

The similarity problem in trajectory data base for moving objects on road networks has many applications in security informatics area like traffic security, identification of traffic congestion and re-routing etc. It has also applications in other areas like logistics, supply chain management, geo-marketing. In this paper we have identified an earlier work on security informatics as a baseline concept, and demonstrated how the binary encoding scheme of location data is advantages over existing trajectory similarity algorithms for network constrained moving objects. We have modified these algorithms in finding percentage of similarity which could be used as a measure for trajectory clustering applications. As a continuation work we are planning to use these measures in different types of clustering algorithms that will have lot of applications in security related applications like segregation of objects with specific moving characteristics, emergency area clusters, dark web links analysis, spreading of infectious diseases.

Acknowledgements

This research is being supported by University Grants Commission(UGC) under Ministry of Human Resources, Government of India under the scheme of Grant to Academic Research for Faculty members in Colleges and Universities vide order No. 37-633/2009 (SR).

References

1. Abraham, S., Sojan Lal, P.: Trigger Based Security Alarming Scheme for Moving Objects on Road Networks. In: Yang, C.C., Chen, H., Chau, M., Chang, K., Lang, S.-D., Chen, P.S., Hsieh, R., Zeng, D., Wang, F.-Y., Carley, K.M., Mao, W., Zhan, J. (eds.) ISI Workshops 2008. LNCS, vol. 5075, pp. 92–101. Springer, Heidelberg (2008)
2. Yanagisawa, Y., Akahani, J., Satoch, T.: Shape-Based Similarity Query for Trajectory of Mobile Objects. In: Proc. of the 4th Intl. Conf. on MDM, pp. 63–77 (2003)
3. Hwang, J.-R., Kang, H.-Y., Li, K.-J.: Spatio-temporal Similarity Analysis between Trajectories on Road Networks. In: Akoka, J., Liddle, S.W., Song, I.-Y., Bertolotto, M., Comyn-Wattiau, I., van den Heuvel, W.-J., Kolp, M., Trujillo, J., Kop, C., Mayr, H.C. (eds.) ER Workshops 2005. LNCS, vol. 3770, pp. 280–289. Springer, Heidelberg (2005)
4. Hwang, J.-R., Kang, H.-Y., Li, K.-J.: Spatio-temporal similarity analysis between trajectories on road networks. In: Akoka, J., Liddle, S.W., Song, I.-Y., Bertolotto, M., Comyn-Wattiau, I., van den Heuvel, W.-J., Kolp, M., Trujillo, J., Kop, C., Mayr, H.C. (eds.) ER Workshops 2005. LNCS, vol. 3770, pp. 282–295. Springer, Heidelberg (2005)
5. Tiakas, E., et al.: Searching for similar trajectories in spatial networks. J. Syst. Software (2009) doi:10.1016/j.jss.2008.11.832
6. Theodoridis, Y.: R-Tree Portal (validation, February 2007), `http://www.rtreeportal.org`
7. Shim, C.-B., Chang, J.-W.: Similar Sub-Trajectory Retrieval for Moving Objects in Spatiotemporal Databases. In: Proc. of the 7th EECADIS, pp. 308–322 (2003)
8. Vlachos, M., Gunopulos, D., Kollios, G.: Robust Similarity Measures of Mobile Object Trajectories. In: Proc. of the 13 th Intl. Workshop on DEXA, pp. 721–728. IEEE Computer Society Press, Los Alamitos (2002)
9. Vlachos, M., Kollios, G., Gunopulos, D.: Discovering Similar Multidimensional Trajectories. In: Proc. of the 18th ICDE, pp. 673–684. IEEE Computer Society Press, Los Alamitos (2002)
10. Sakurai, Y., Yoshikawa, M., Faloutsos, C.: FTW: Fast Similarity Search under the Time Warping Distance. In: PODS, pp. 326–337 (2005)
11. Chen, L., Ozsu, M.T., Oria, V.: Robust and Fast Similarity Search for Moving Object Trajectories. In: ACM SIGMOD, pp. 491–502 (2005)
12. Zeinalipour-Yazti, D., Song Lin, S., Gunopulos, D.: Distributed Spatio-Temporal Similarity Search. In: CIKM, pp. 14–23
13. Tiakas, E., Papadopoulos, A.N., Nanopoulos, A., Manolopoulos, Y.: Trajectory Similarity Search in Spatial Networks. In: Proc. of the 10th IDEAS, pp. 185–192 (2006)

14. Chang, J.-W., Bista, R., Kim, Y.-C., Kim, Y.-K.: Spatio-temporal similarity measure algorithm for moving objects on spatial networks. In: Gervasi, O., Gavrilova, M.L. (eds.) ICCSA 2007, Part III. LNCS, vol. 4707, pp. 1165–1178. Springer, Heidelberg (2007)
15. Orenstein, J.A., Merrett, T.H.: A class of data structures for associative searching. In: 3rd ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, Waterloo, Ontario, pp. 181–190 (1984)
16. Lee, S.Y., Park, S., Kim, W.-C.: An efficient location encoding method for moving objects using hierarchical administrative district and road network. Information Sciences 177, 832–843 (2007)

An Emergent Traffic Messaging Service Using Wireless Technology

Patrick S. Chen, Yong-Kuei Liu, and Chia-Shih Hsu

Dept. of Information Management, Tatung University
chenps@ttu.edu.tw, alex51204612@yahoo.com, kobe0621@hotmail.com

Abstract. Real-time information about traffic flow, accidents, or road availability is very important for drivers, and it is useful to implement a messaging system for supporting inter-vehicular communication. Aided by positioning instruments and *ad hoc* network, important messages can be sent from vehicle to vehicle to regulate traffic flow, enhance driving safety and distribute alert messages. We identified various types of broadcasting and analyzed the circumstances for broadcasting. The usefulness of broadcasting was considered in which a message about an accident is important for vehicles approaching the accident, and the information that an ambulance is coming is also important for vehicles before the ambulance. LAN and 3G provide a communication platform for messaging. Broadcasting, P2P communication and SMS are a choice for the messaging system. A prototype of the system was implemented and its application was discussed as well.

Keywords: VANET, GPS, Risk Management, Messaging Service.

1 Introduction

Vehicle Management System is an integration of global positioning system (GPS), geographic information system (GIS), mobile communication system and Internet technology, offering users valuable, real-time information. With a GPS-equipped vehicle, important information like current location, speed, direction, vehicle condition, etc. is sent to the control center. The manager is able to know oil consumption, personnel in vehicle, goods delivery, and anomalous state of the vehicle, and to make right decisions to enhance performance. A vehicle management system can:

(1) Organization of the caravan: update the state of the caravan and the personnel;
(2) Real-time monitoring: control the vehicle based on the latest information, including the condition of the vehicle;
(3) Retrieval of historic data: check the current and historical operation data;
(4) Statistics: provides several statistical reports including transportation fee, overspeed records, oil consumption, mileage, etc.
(5) Vehicle positioning: facilitate query by street, milestone, administrative area and other geographical conditions;
(6) Instant messaging: broadcast message to vehicles using general mode and group mode:
(7) Presentation of vehicle states: present the current condition of a specified vehicle.

H. Chen et al. (Eds.): PAISI 2010, LNCS 6122, pp. 44–51, 2010.
© Springer-Verlag Berlin Heidelberg 2010

Some industries like taxis require high flexibility to offer high-quality service. Taxi drivers shuffle either in downtown or in countryside to offer fast service to customers. To know the traffic flow and accidents is of great importance for them. Currently, they rely on pagers to share information within the caravan. But this simple equipment cannot meet the requirements of advanced management. The main problem is that it cannot distribute detailed information, coverage of the pager system, exactness of information like location. In this study, we will propose a novel communication system for caravans with vehicles equipped with handset to facilitate inter-vehicular communication.

In an *ad hoc* network, broadcasting is used to distribute information such as emergency, actual traffic flow, business decisions, even entertainment materials. Though simple pager system can afford to distribute information within the caravan, it has several drawbacks like noise, interception, narrow bandwidth, etc. With the development of mobile techniques and smart handheld equipment, better communication service can be offered.

There are three different types of vehicle *ad hoc* network (VANET): inter-vehicle communication (IVC), roadside-to-vehicle communication (RVC) and hybrid-vehicle communication (HVC). Management modes of VANET can be centralized or distributed. In this study, HVC is considered. IVC is used when all vehicles concentrate in a closed area. RVC is used when vehicles are far away from each other. This will increase the reliability of the entire system. Furthermore, as vehicles undergo rapid locative change, we adopt a distributed management mode, so a restructuring of the vehicles group is not necessary.

In instant point-to-point messaging, each site is both a server and a client. As shown in Fig. 1, all vehicles can communicate with each other; failure of any site will not corrupt the network. Managing a taxi caravan, the server can be the management center.

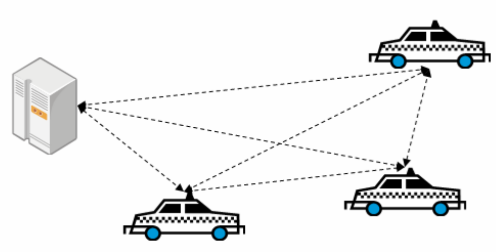


Fig. 1. A fully-connected communication network

This study, proposing a framework to facilitate inter-vehicular communication based on broadcast protocol, aims to reduce broadcasting storms, reduce repeated broadcast, and determine transfer destination. Section 2 is literature review, in which the reason of selecting P2P is explained. Section 3 describes the proposed framework. Section 4 presents the system prototype. And the final section is conclusion.

2 Literature Review

Along with the development of mobile technology, wireless communication network based on the protocol of IEEE 802.11x has become a new alternative for supporting intelligent transportation system (ITS). The P2P communication system forms a distributed framework in which connected nodes sharing a common topic constitute a network without intermediary or server.

According to IEEE 802.11 DCF (Distributed Coordination Function), types of broadcasting can be centralized or distributed. In order to improve the distributed system, we may put some vehicles in groups and appoint a node as coordinator to manage the channel. There are four types of broadcasting: simple flooding, probability-based broadcasting, counter-based broadcasting, and location-based broadcasting [2].

Researchers have proposed various ways for broadcasting. They distinguish among dense traffic regime, carry forward, retry and carry forward [5]. A decision tree can be constructed for DV-CAST in well-connected, sparsely-connected and totally disconnected network [4]. We distinguish between dense traffic, sparse traffic, and regular traffic.

In case of emergency, it is important to inform the vehicles coming from the opposite direction. Targeted routing of the message is of great importance. For instance, information of accident is needed by the vehicles coming behind the site of the accident while the information that an ambulance is approaching is important for vehicles before the accident. Broadcast direction has great impact on the network efficiency [7]. The coordinate system and broadcast direction can be used to classify emergency information to determine where to broadcast. It enhances the applicability of the system [3].

In order to reduce the probability of sending the same message, There are three kinds of broadcasting: weighted p-persistence broadcasting, slotted 1-persistence broadcasting, and slotted p-persistence broadcasting [6]. These methods are able to reduce redundant packets or avoid loss of packets, but position information is occasionally needed. Besides the position information, local traffic situation is needed as well on the highway or in the city. It is impossible to forward any message if no neighboring nodes are available in an area.

3 The Model

In order to provide the vehicles with necessary information, we propose a real-time, distributed decision-support system for two purposes: one is management, aiming to offer the caravan an information platform to convey real-time information; another is communication, aiming to control the channel and minimize repeated sending. For better understanding, we illustrate it with the aid of two scenarios: with or without GPS.

A message can be forwarded to vehicles via base station. If Vehicle A knows there is an accident, it then passes the information to Vehicles B, C, D and E. We may determine a vehicle by Eq. 1 to store the message based on the principle of weighted p-persistence broadcasting. All vehicles in the coverage of AP will calculate their probabilities, and the vehicle with the highest probability is responsible for storing the message.

$$P_{ij} = \frac{D_{ij}}{R} \tag{1}$$

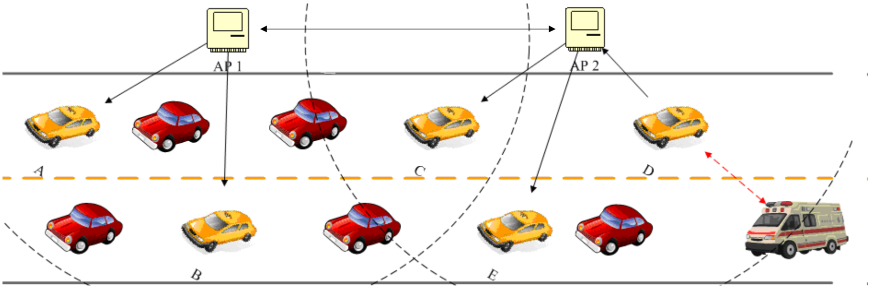


Fig. 2. Scenario I: Information classification

Message category is vital in determining the direction of broadcasting: if it is about an accident, the information is important for vehicles after it. If the message is about the coming of an ambulance, it is more important for the vehicles before the ambulance. We may attach a sign (+,-) to determine the direction of messaging. For example, Vehicle A in Fig. 2 sends out a message, and Vehicles B, C, D and E will receive the message with the sign "+". By contrast, Vehicle D in Fig. 4 senses an ambulance coming from behind and sends out a message, Vehicles A, B C, and E will receive the message with the sign "-". Directional information is very useful in practical use in that it helps to determine the probability of data storage.

In case no GPS signals are available, e.g., vehicles inside along tunnel, location-based messaging can be used instead [1]. As shown in Fig. 3, Vehicle A finds an accident and sends the information to the control center, whicc in turn will forward it to specified areas. The road should be divided into several sections beforehand.

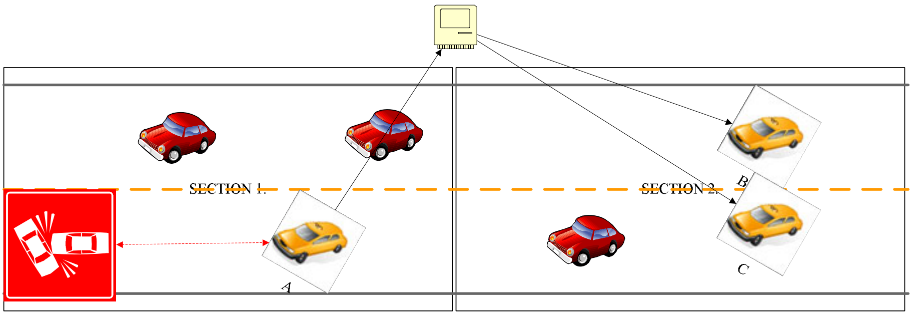


Fig. 3. Scenario II: Messaging in specified area

4 System Analysis

The system architecture for communication in Scenario I is illustrated in Fig. 4, where two peer vehicles determine the IP of their counterpart and communicate through Port 8080. They all make use of the Listener mechanism of the web browser to wait for messages sent by their counterpart.

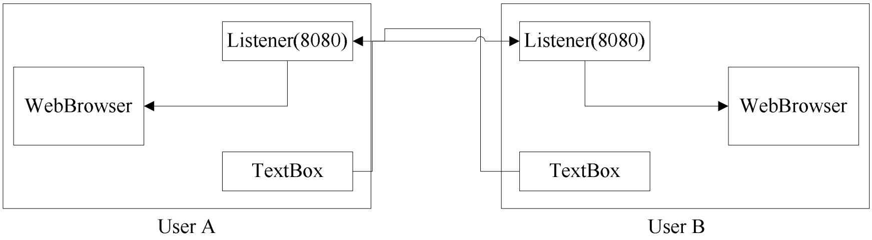


Fig. 4. P2P Communication architecture

Formula (2) is a string of GPS signals consisting of 15 parts in which 2501.9891, N, 12133.8101, E are used to extract location information. In the formula, L stands for longitude (or latitude), D for degree, M for minute, and S for second.

```
$GPGGA,095031.254,2501.9891,N,12133.8101,E,1,07,7.0,
123.9,M,15.0,M,0.0,0000*74                                (2)
```

$$\mathrm{L} = \mathrm{D} + \left(\mathrm{M} * 60 + S/10000 * 60\right)/3600 \tag{3}$$

If all vehicles are equipped with GPS and there are access points (AP) on roadside, we use P2P communication either over LAN based on IEEE802.11x or mobile communication through 3G. If vehicles from the opposite direction would not help to carry a message, we then use disconnected mode for sparse traffic and regular traffic. Therefore, only well-connected mode and disconnected mode will be considered. In order to avoid a broadcasting storm, the system provides the user a function to decide whether to forward the message. In case no GPS signals are available, we use information push system as an auxiliary. An information push system is suggested to include three functions: message generation module, receiver allocation module, and message transmission module [1]. But their system is used for short message service; what we need for inter-vehicle communication is much more complex. A great extension of the system is expected.

Information involved in inter-vehicle communication includes event, vehicle direction, street, distance to an accident, traffic flow, IP of the sender, and remarks from the user. The Euclidian distance is determined as Eq. 4.

$$D = \sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2} \tag{4}$$

After receiving a message, a user then calculates his location and probability, and decides where to forward this message. If he is located in a disconnected situation (e.g., in sparse traffic), the server will locate an area and broadcast in his area.

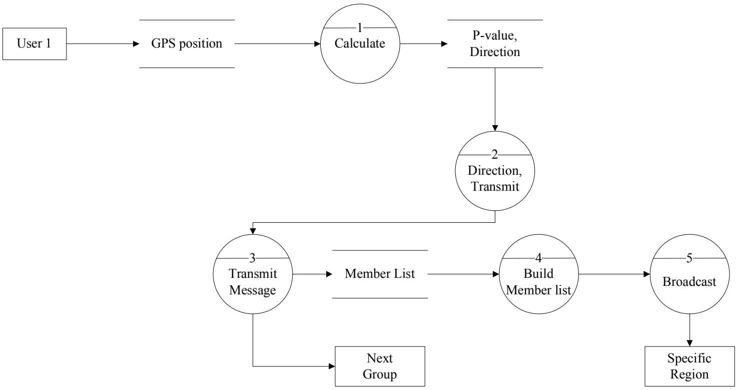


Fig. 5. Main modules of the inter-vehicle communication system

The system development environment was Visual Studio 2005 .NET Framework. We used VB to design the system functions and user interface. SQL Server 2005 was used to manage all necessary data. WiFi (802.11a) and 3G is the platform for communication, Blue Tooth or USB receiver was used to connect GPS, and all real-time data is displayed for supporting user's decision.

5 System Design and Implementation

User interface was intended for use in moving environment. The user sends out a message by touches the screen. Eight kinds of information are available. Output is in the form of text as shown in Fig. 6. The user interface is divided into 3 blocks: The upper block, which includes driver identification and car plate number; the middle block, which includes icons for 8 different kinds of information including car and road conditions; the lower block, which documents information.

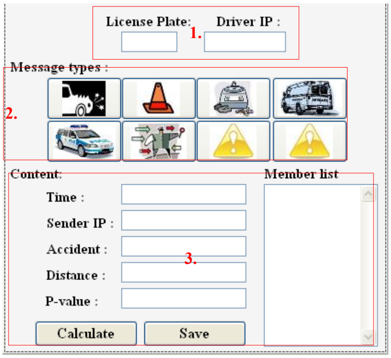
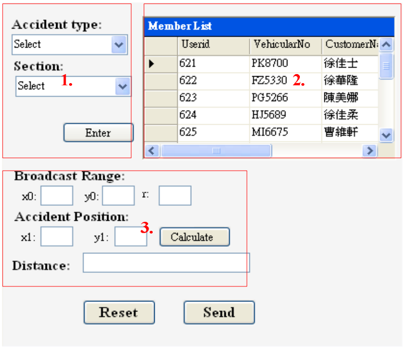


Fig. 6. User interface



Accident type: Select

Section: Select

	Userid	VehicularNo	CustomerN:
▶	621	PK8700	徐佳士
	622	FZ5330	徐華隆
	623	PG5266	陳美娜
	624	HJ5689	徐佳柔
	625	MI6675	曹維軒

Fig. 7. Administrator's interface

In case no GPS signals are available, roadside access points are responsible for transmitting message. The message is initiated by driver and forwarded by the APs. At the control center, the administrator's interface is divided into 3 blocks: the left upper block is for administrator to enter information about an event; the right upper block is a list of members, and the lower block is a specification of broadcast.

The system is usually in IDLE state. When a message comes, it changes to GetMsg state to receive message. Then it tries to check GPS signals. If there are signals, then calculate its probability, the p-value. If the p-value does not exceed the upper limit, it then enters the state of KMsg and keeps the message. Otherwise, it changes into Rebroadcast state. Before sending the message, it checks the neighboring state. If the state is well connected, it broadcasts the message. If the state is disconnected, it forwards the message to the control center.

To activate the system, we first store the basic data of car plate number, IP, vehicle ID etc. In case there are no GPS signals available, the user should first determine the types of event and calculate the distance, and then set up a list of receivers and broadcast the message.

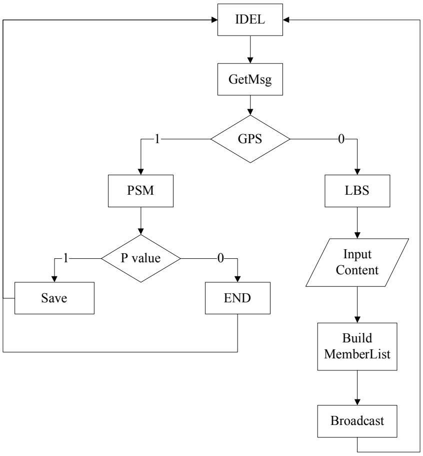


Fig. 8. Data flow of the system

6 Conclusion

Besides the conventional message broadcasting, the proposed system helps to manage a caravan taking into account message usefulness and avoidance of a broadcasting storm. In case no GPS signals are available, location-based broadcasting will be used to carry forward the message. The system is useful for managing taxis, police wagons, fire-fighters, ambulance, etc. In subsequent study, we will administer questionnaires to know its applicability when the system is put into practice. The Unified Theory of Acceptance and Use of Technology proposed by Venkatesh et al. (2003) will be the research structure of the empirical study. It is sugested to know the benefits of the system from the drivers in order to manage the caravan more efficiently.

References

1. Chen, P.S., Tzeng, J.R., Taso, W.H.: Location-Based SMS Information Push System – Its Architecture, Design, and Application. In: Multi-Conference on Systemic, Cybernetics and Informatics WMSCI (2007)
2. Fan, P.: Improving Broadcasting Performance by Clustering with Stability for Inter-Vehicle Communication. In: IEEE 65th Vehicular Technology Conference, VTC 2007, pp. 2491–2495 (Spring 2007)
3. Fukuhara, T., Warabino, T., Ohseki, T., Saito, K., Sugiyama, K., Nishida, T., Eguchi, K.: Broadcast Methods for Inter-Vehicle Communications System. In: IEEE Communications Society WCNC, vol. 4, pp. 2252–2257 (2005)
4. Tonguz, O., Wisitpongphan, N., Bait, F., Mudaliget, P., Sadekart, V.: Broadcasting in VANET. In: IEEE Mobile Networking for Vehicular Environments, pp. 7–12 (2007)
5. Wisitpongphan, N., Tonguz, O., Parikh, J., Bai, F., Mudalige, P., Sadekar, V.: On the Broadcast Storm Problem in Ad hoc Wireless Network. In: 3rd International Conference on Broadband Communications, Networks and Systems, BROADNETS 2006, pp. 1–11 (2006)
6. Wisitpongphan, N., Tonguz, O.K., Parikh, J.S., Mudalige, P., Bai, F., Sadekar, V.: Broadcast storm mitigation techniques in vehicular ad hoc networks. IEEE Wireless Communications 14, 84–94 (2007)
7. Yang, Y., Chou, L.: Position-Based Adaptive Broadcast for Inter-Vehicle Communications. In: IEEE International Conference on Communications Workshops, ICC Workshops ’08, pp. 410–414 (2008)

A Model for Detecting “Global Footprint Anomalies” in a Grid Environment

Pramod S. Pawar[1] and Srinath Srinivasa[2]

[1] Centre for Development of Advanced Computing
68, Electronics City, Bangalore 560100, India
pramod@ncb.ernet.in
[2] International Institute of Information Technology
26/C, Electronics City, Bangalore 560100, India
sri@iiitb.ac.in

Abstract. Grid computing environments pose unique security concerns that are not generally relevant for conventional data management systems. An event that appears as benign on a grid node, may actually be part of a larger incident hazardous to the grid. Since a node only sees the local footprint of an event, it cannot know the contribution of this event at a global scale. The focus of this work is on detecting such kinds of anomalous behaviors that we call *global anomalies*. In this paper, we propose two classes of global anomalies, and a model for detection of a class of global anomalies that we call *global footprint anomalies*. The main challenge here is to detect anomalous behavior, which looks normal locally at any individual grid node, but when observed globally, the anomalous behavior is apparent.

1 Introduction

One of the biggest challenges in current day Intrusion Detection Systems (IDS) over grid computing environments is to handle distributed attacks [10]. Several models of distributed attacks exist, like: Distributed Denial of service (DDoS)[1], Race condition exploitation, etc. A commonly used method to detect attacks is to use *anomaly detection*. Anomaly detection entails detecting behaviors that are in some way “abnormal” while distinct from “noise” or “novel” behaviors [3]. A common approach to anomaly detection are unsupervised learning strategies with the underlying assumption that most activities of the system are normal. Anomaly detection in such cases, involves profiling the normal behavior of the system and detecting deviations from the normal behavior in order to flag anomalies.

In a grid computing environment, we propose a classification of anomalies as shown in Figure 1. we propose a distinction of anomalies, which can be of two kinds: *local* anomalies or *global* anomalies. Local anomalies affect a single node in the grid and are usually detected and handled locally. Global anomalies represent anomalous behavior whose impact goes beyond any single node and for which distributed algorithms are

[1] http://tools.ietf.org/pdf/rfc4732,
http://www.cert.org/homeusers/ddos.html

H. Chen et al. (Eds.): PAISI 2010, LNCS 6122, pp. 52–64, 2010.
© Springer-Verlag Berlin Heidelberg 2010

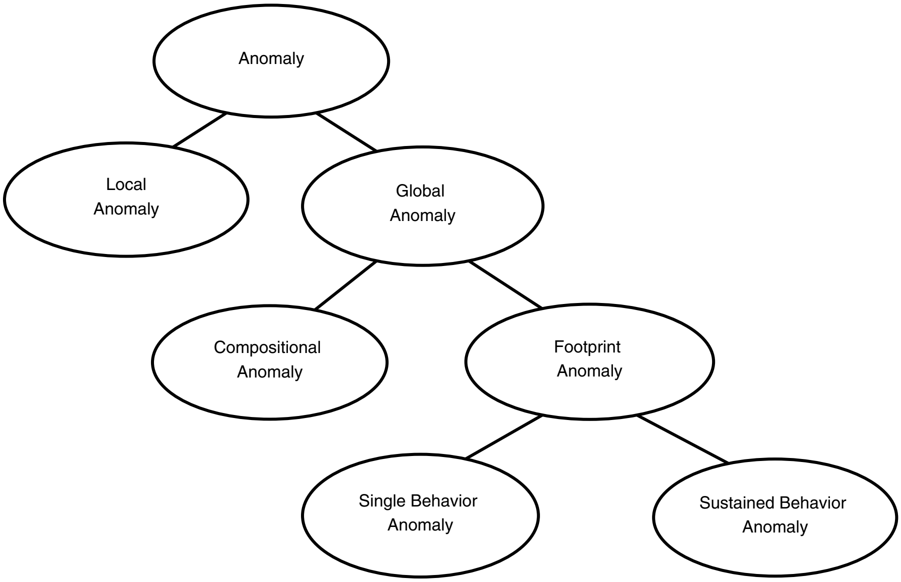


Fig. 1. Classification of Anomalies

required for detecting or handling such behaviors. Worm attacks and distributed denial of service (DDoS) attacks are examples of global anomalies. As the size of the grid and its activity level grows, global anomaly detection becomes very challenging. Detecting global anomalies requires collecting profile information from several nodes and aggregating them on a continuous basis. This becomes infeasible as the number of nodes in the grid becomes large. Hence global anomaly detection has to resort to event-based approaches where the communication costs of both profiling and detection are manageable. For flagging global anomalies, this usually means that some node in the grid has to start the process by suspecting that something is wrong. The trigger for this usually is the detection of a local anomaly by a grid node. Based on the nature, severity or other characteristics of the anomalous behavior, the grid node suspects that the anomaly is actually on a bigger scale, and flags off a global detection algorithm. Such an approach has been successfully used for worm detection in distributed environments [6]. However, in this paper we focus on a class of anomalies that are global in nature, but which appear *locally normal* or *benign* at every node. With the absence of any local trigger, no node can initiate a global anomaly detection process. Some examples of such anomalies are as follows:

1. Consider different people carrying different liquids into an airplane. Each of these liquids may be considered safe in isolation, but their combination can be potentially dangerous. Each individual would be passed through security if the global picture is not considered.
2. An ATM machine validates a transaction by comparing the magnetic id of the card along with the PIN provided by the user. However, if a card with the same magnetic

id and PIN were to be inserted in two different ATM's separated by thousands of kilometers within a matter of minutes, the behavior is clearly anomalous. However, since the machine performs its validation in isolation, such anomalies are unlikely to be detected.

3. Consider the evaluation of answer papers for a nation-wide examination. If each centre has a different evaluator, then similar or identical answer sheets across the centres go undetected since each evaluator performs an integrity check on the answer-sheets in isolation.

In our exploration for application cases involving global anomalies, we have been able to classify global anomalies into two broad classes that we call the following:

1. Compositional anomalies, and
2. Footprint anomalies

A *compositional anomaly* is one where the anomaly results from some form of a *composition operation* over local events. The first scenario in the list above is an example of a compositional anomaly. The events that make up a compositional anomaly are typically of disparate event types.

A *footprint anomaly* is one that is characterized by *anomalous correlations* between multiple local events of a given type. The second two scenarios in the list above are about the footprint anomaly. Footprint anomalies typically involve events of a single type, that make them somewhat simpler to address than compositional anomalies.

The focus of this paper is only on footprint anomalies. Also, the class of problems addressed in this paper does not include any of the specific examples discussed above. All three examples denote global anomalies that are due to a single behavioral instance. Instead, the focus of this paper is on global footprints denoting a *sustained* anomalous behavior as shown in the Figure 1. Some examples include: subtle changes in application usage patterns across a wide-area banking network, that are apparent only at a global scale; spurious cooccurrences of ailments or diseases of a given type across a nation-wide medical grid, that are too low to trigger a local alarm, but may nevertheless signify a possible pandemic threat.

In the late 1980s and the early 1990s, enormous amounts of research have gone into the problem of *distributed snapshots* or *global states* of a distributed system. Global state detection addresses several problems like distributed commits, distributed termination detection, distributed deadlock detection, virtual time approximation, etc. Chandy [4] and Mattern [13] provide good points of entry into research literature on global state detection. More recently, global state detection has also been used in mobile and ad hoc networks for problems like topology discovery and localization [1,18].

However, we found global state detection algorithms to be largely unsuitable for the challenges of global anomaly detection. Much of the work in distributed snapshots have concentrated on obtaining consistent snapshots where consistency is defined in terms of causal or even sequential consistency. In our case, the consistency requirements are more relaxed as only statistical measures of footprints are required. On the other hand, while distributed snapshot algorithms typically require traversal over the entire distributed system, it would be very inefficient over large grids comprising of hundreds or even thousands of nodes separated over wide geographical areas. Such algorithms become infeasible when a large set of event classes have to be observed on a continuous basis over the entire grid.

The proposed solution resorts to distributed hashing techniques to reduce communication overhead; and graph encoding techniques based on Markov Random Fields to compute the essence of a global footprint, that can be efficiently communicated across the grid and used for comparison in the detection phase.

2 Related Work

One of the earliest papers on intrusion detection, by Denning [7] has articulated several techniques that can be adopted for detecting intrusions. This paper proposed a model of real time intrusion-detection expert system capable of detecting break-ins, penetrations and other forms of computer abuse. The model includes profiles for representing behavior of subjects with respect to objects in terms of metrics and statistical models, and rules for acquiring knowledge about this behavior from audit records and for detecting anomalous behavior.

AAFID [17] is a distributed intrusion detection architecture and system that proposed the use of autonomous agents for intrusion detection. The paper describes the AAFID architecture and the existing prototype, as well as some design and implementation experiences. In 1998 DARPA published an evaluation data for intrusion detection systems and the Master’s thesis by Kenndal [10] describes all the attacks that were the part of this evaluation data. It describes the simulation network used to collect this data. It focuses on different types of attacks that were developed and presents a taxonomy of computer attacks. There are several efforts to apply statistical techniques and one such effort include application of Chi-square technique [21] for anomaly detection. In this paper the author identifies a number of variables in a computer and network system, group them into several categories. A two-stage process is used, first to build a normal profile during training and use it to detect anomalous activities during testing.

The need for IDS within grid environments is a topic of pertinent discussions at the Open Grid Forum[2]. Kenny and Coghlan [11] propose a solution for efficiently accessing audit data from grid but there is no mention on how to use the data to identify intrusions. Choon [5] describes a grid-based IDS architecture that consists of agents located at grid nodes responsible for collecting and sending host audit data to storage and analysis but it is a centralized solution. The Grid Intrusion Detection Architecture (GIDA) proposed by Tolba [19] solves the scalability problem by distributing the intrusion detection problem among several analysis servers. Both Choon et al. [5] and Tolba et al. [19] concentrate on the detection of anomalies in the interaction of grid users with resources which result of misuse. But they lack detection of unauthorized access, exploits detection. Fang-Yie et al. [22] addresses some of these issues and also network denial of service attacks. Schulter et. al [16] propose a distributed IDS architecture integrating the detection of typical host-specific attacks with grid specific attacks and user behavior anomalies. This paper comes with their class of grid intrusions like (a). Unauthorized access, (b). Misuse, (c). Grid exploits, and (d). Host or network specific attacks.

[2] www.ggf.org

Slow propagating attacks are difficult to detect which can be hidden under the veil of normal traffic. Dash et al. [6] describe a method for detecting such attacks using distributed probabilistic inferences. The main contribution of this paper is the probabilistic framework that aggregates (local) beliefs to perform network wide inferences. Local detectors (LD) detect order of magnitude slower worms at local nodes and the global detectors (GD) taking aggregated views from LDs determine if the network as a whole is in anomalous state. Li et al. [15] also focus on zero-day slow scanning worms. For effective intrusions the method uses a host organization based on the concept of regions and consider dependency among hosts within each region. There are 3 kinds of detectors: local detectors, regional detectors and global detectors. Local detectors which are weak in detecting capability reside on each end hosts. Regional detectors diagnose potential problems at neighborhood levels. Global detectors uses sequential hypothesis testing to detect any intrusions at the global level. Huang et al. [9] describes network anomaly detection using distributed PCA (Principal Component Analysis) techniques for data collected and processed over a large distributed system. In this paper they consider set of local monitors each of which collect locally observed time series data streams. A central coordinator node does the global collection and makes global decision concerning network-wide health. The focus here is to detect volume anomalies that refers to unusual traffic load levels that may be caused due to worms, DDoS attacks etc.

In all of the above techniques a basic assumption is that, at least one of the local detectors has to flag an anomaly in order to trigger the global anomaly detection algorithm. Also the global detector takes into account the aggregated view of the local anomaly detected to decide the global anomaly. In our work, we make no such assumption of local anomaly detection.

3 Global Anomaly Detection Model

This section presents the proposed model for detecting footprint global anomalies. The underlying assumption is that of a single event type, whose co-occurrence footprint when viewed globally can be flagged as anomalous. The proposed system is based on an unsupervised learning model with the underlying assumption that most events are normal. Based on this, the model comprises the following steps: (a). building a reference global footprint profile, of all event classes in the grid that are to be observed, (b). building the detection global footprint profile from event instances observed, and (c). determining if there is any significant deviation between the reference global profile and the detection global profile. The system initially works in a profiling mode to build a reference global footprint profile of the grid event class. Once the profiling is completed for a configured amount of time, the system switches into a detection mode. In detection mode, the detection global profile is compared with the reference global profile and upon any deviations an anomaly is flagged.

3.1 Computing the Local Footprint

The profiling phase of the anomaly detection algorithm observes a specified set of event classes over a sufficiently long period and builds their "local" and "global" footprint profiles. The process adopted for this is explained below.

Let

$$E = \{e_1, e_2, e_3 \dots e_n\}$$

be the set of *event classes* of interest, for the entire grid. An event class is a class of any activity of interest that takes place on one or more nodes of the grid. User login of a given user, ATM card transaction of a given customer, Starting of an application, Acceptance of a network socket connection, etc. are all examples of event classes.

Each grid node records observations of events as a function of time called *epoch*. An epoch is a unit of time whose duration typically has specific relevance to the event class to be profiled. An epoch could be a day or week or month depending on the semantics of the event being profiled. An epoch in turn is divided into suitable time units, to form time intervals from $t_1, \dots, t_m$. Thus an epoch T with time intervals $t_1, t_2, \dots t_m$ can represented a sequence:

$$T = (t_1, t_2, \dots, t_m)$$

An *event signature* or the *local footprint profile* of event class e is a histogram over T, showing either the number of occurrences or the probabilities of the occurrence of e in each of the time interval $t_i \in T$. We use the notation $\mathbf{e}_j^i$ to denote the *footprint vector* of event class e_j at grid node i, and the $e_j^i(t)$ to denote the *interval occurrence* of event class e_j recorded for time interval t at grid node i.

The *interval probability* of event class e_j for time t at grid node i, denoted as $Pr[e_j^i(t)]$, is calculated as:

$$Pr[e_j^i(t)] = \frac{|e_j^i(t)|}{\sum\limits_{\forall t \in T} |e_j^i(t)|} \tag{1}$$

where $|e_j^i(t)|$ denotes the number of occurrences of events of class e_j in time interval t on grid node i. The interval probability hence denotes the relative importance of time interval t for the event class, in an epoch.

A *local interestingness threshold* function $\delta_j \in [0, 1]$ is used to identify an interval as either interesting or uninteresting as far as the event class e_j is concerned, on a given grid node.

3.2 Computing the Global Footprint

In order to compute the global footprint of a given event class e_j, it is required to exchange local footprint profiles across all nodes in the grid. This is a costly operation over large grids with hundreds or thousands of nodes distributed over wide geographical areas. Given that the global footprint is a central piece of information during the detection phase, it is necessary to compute and communicate the gist of the global footprint of an event in an efficient fashion. This is achieved by modeling global footprints as Markov Random Fields (MRFs) [12] over the grid. The process is explained in detail below.

A grid-wide distributed hash table (DHT) is maintained where the identifier of each event class e_j is used as the hash key. Let τ_j denote the number of epochs that constitute the profiling phase, where the footprint of event class e_j is constructed. Each grid node records events and constructs local footprints for each specified event class e_j. At the end of every epoch after the first τ_j epochs (i.e. after the profiling phase is over), grid nodes update their local footprints for event class e_j over the DHT. Let grid node k be in charge of maintaining local footprints of event class e_j. Based on the received set of local footprints, the grid node computes an *event correlation graph* G_j for each event class that it maintains. An event correlation graph $G_j = (V, E)$ for event class e_j is a graph where V is the set of grid nodes, and $E \subseteq V \times V \times [-1, 1]$, relates grid nodes based on correlations between their local footprints of event class e_j.

In this work, correlations between local footprints is computed as the *joint probability of cooccurrence* across time intervals that make up a footprint. Let $T_j^a = (a_i, a_2, \ldots, a_m)$ and $T_j^b = (b_a, b_2, \ldots, b_m)$ be the local footprints from nodes a and b respectively for event e_j. Let δ_j be the interestingness threshold for event j. In order to compute the joint probability of cooccurrence, we first set all interval probabilities $a_i \in T_j^a$ and $b_i \in T_j^b$ to either 0 or 1, depending on whether they are below the threshold δ_j or at least as high as δ_j respectively.

Now that both T_j^a and T_j^b are reduced to bit vectors, the joint probability of cooccurrence is computed as:

$$J^{a,b} = \frac{|T_j^a \wedge T_j^b|}{|T_j^a \vee T_j^b|} \tag{2}$$

where $\wedge$ and $\vee$ are the logical AND and OR operators respectively, and $| \cdot |$ counts the number of 1s in the given bit vector. An example event correlation graph is shown in Figure 2 over a grid comprising of five nodes.

Another thresholding function Δ_j called as the "correlation interestingness threshold" is used to remove all edges from the correlation graph whose joint probability is too low.

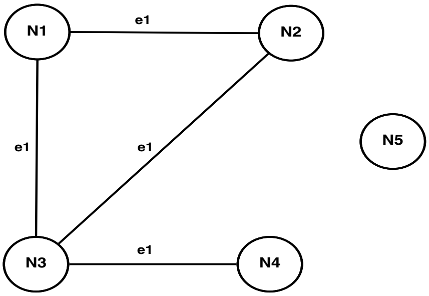


Fig. 2. Event Correlation graph for event type e_1

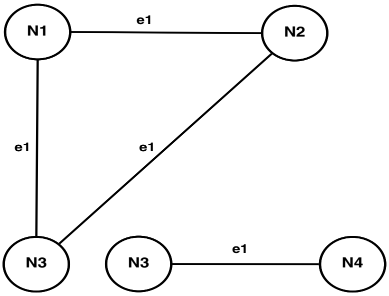


Fig. 3. Maximal cliques C_1 & C_2 for event e_1

The global footprint of an event is now defined in terms of potential functions (also called "energy" functions in this paper), over maximal cliques of the event correlation graph. Maximal cliques for the event correlation graph of Figure 2 is shown in Figure 3.

Potential functions over maximal cliques are computed using the energy model proposed in Rachakonda and Srinivasa [14] and is explained as follows. The energy (E_C)

associated with a maximal clique is directly proportional to the mean (μ_C) of all the edge weights between the nodes of the cliques and inversely proportional to its variance (σ_C). A high mean indicates that the events co-occur together a lot and high variance of the edge-weights indicate that the set has some elements that do not belong to this profile. This is given as:

$$E_C = \frac{\mu_C}{(1 + {\sigma_C}^2)} \tag{3}$$

The global profile for an event class, is represented as a set of clique and energy pairs. Given a grid G comprising of a set of nodes, the global footprint profile of event class e_j is given by:

$$\mathcal{P}_j = \{(C, E_C) \mid C \in 2^G\} \tag{4}$$

Thus for the event class e_1 the global profile $P_{e_1} = \{(C_1, E_{C_1}), (C_2, E_{C_2})\}$ as shown in Figure 3.

Triadic closure: Graphs that capture co-occurrence information display a *triadic closure* property resulting in *clustered graphs* that is a characteristic feature of social acquaintance networks [8,20].

An undirected graph $G = (V, E)$ is a clustered graph if for any $a, b, c \in V$, if $\{a, b\}, \{a, c\} \in E$ it implies with high probability that $\{b, c\} \in E$.

For instance, given any three grid nodes a, b and c, if the joint correlation probabilities (Equation 2) $J^{a,b}, J^{a,c} \geq 0.6$, we see that the triadic closure necessarily holds whenever $\Delta_j \leq 0.2$. In such cases, the global footprint of the event class e_j would be a single clique or a set of disjoint cliques with no grid nodes in common. Depending on the application context, the latter may also indicate separate footprint instances (for example, two unrelated deployments of a distributed application).

Once the global footprint of an event is generated, it is sent to all grid nodes involved in the global footprint.

3.3 Detection Phase

As noted earlier, for any event class e_j, the number of epochs for which profiling is done, is represented by τ_j. The local footprints are hashed onto the DHT after τ_j epochs following which, the global footprint is computed and communicated to all grid nodes that are part of the footprint.

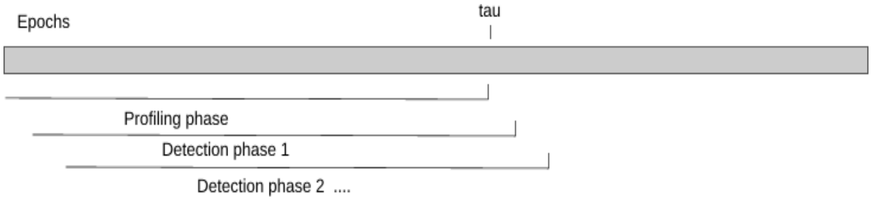


Fig. 4. Sliding epoch window across profiling and detection phases

The detection phase starts immediately after the profiling phase is over. A sliding window is used over the set of epochs where, after the end of the first epoch in the detection phase, the last $\tau - 1$ epochs from the profiling phase are used to compute the interval probabilities. This is schematically shown in Figure 4. Computation of the local footprint and interval probabilities in the detection phase, follow the same mechanism as is used in the profiling phase. Computation of the global footprint for an event class e_j is done at grid node i only when the node i has been detected to be a part of the global footprint of e_j. In such a case, the node requests the latest local footprint information from all the other participating nodes and computes the global footprint. Note that since the scale of an epoch is typically very large (days, weeks, etc.) a global clock (or calendar) can be safely assumed to exist, for synchronizing epoch numbers of local footprints.

Let $\mathcal{P}_j$ be the global footprint profile for event class j and let $\mathcal{D}_j$ be the detected global footprint profile. The profiles can be viewed as a vector comprising of one or more $(clique, energy)$ pairs. Given this, the L_1 distance between the two is computed to check for deviations from the profiled footprint. An anomaly is flagged if $L_1(\mathcal{P}_j, \mathcal{D}_j)$ exceeds a specified threshold.

Global anomaly detection is performed asynchronously and independently by each of the grid nodes. Since the global footprint is based on correlations across local footprints, anomaly detection will not be affected due to changes in overall behavioral patterns – like holidays or rush hour usages, etc. Also, global footprint anomalies cannot be detected at real-time. It requires at least one epoch before which a suitably sustained anomalous can be detected. This makes the detection of global anomalies all the more significant, since it depicts not just anomalous behavior, but *sustained* anomalous behavior.

Superset attacks: The proposed model detects anomalous behaviors on a global scale even when the behavior appears locally benign at all grid nodes. However, there is a class of distributed attacks that can foil the proposed global footprint based anomaly detection technique.

This is what we term the “superset attack.” Consider an event e_j (for example, financial activity in a particular account) to be comprising of the following global footprint:

$$\mathcal{P}_j = \{(C_1, E_1), (C_2, E_2), \ldots (C_n, E_n)\}$$

Now, a malicious adversary may create a global footprint:

$$\hat{\mathcal{P}_j} = \{(\hat{C_1}, \hat{E_1}), (\hat{C_2}, \hat{E_2}), \ldots (\hat{C_m}, \hat{E_m})\}$$

such that $m \geq n$ and/or there exists at least one i where $\hat{C_i}$ is a strict superset of C_i, without altering the local signatures of the event on the nodes of C_i in $\hat{C_i}$. Since each grid node in C_i requests the global signature from other grid nodes that are part of the clique C_i, the new footprint that is a strict superset of C_i goes undetected.

Superset attacks can be addressed by making profiling a continuous process, with thresholds to tradeoff between comprehensiveness of the anomaly detection with communication and computation complexity. Given a threshold k, grid nodes hash their local profiles for event class e_j after every $\tau_j + nk$ epochs where $n = 1, 2, \ldots$ Sustained superset attacks can be detected after at most k epochs in such cases.

3.4 Complexity of Anomaly Detection

Suppose that the grid has N nodes, a reasonable DHT implementation requires $O(\log N)$ lookups for both hashing and retrieval. If the grid represents a wide-area distributed network, usually event footprints span over a small subset of the grid nodes $n \ll N$. Given this, the overall communication complexity of computing the global footprint would be $O(n \log N)$.

The main computational complexity comes from computing maximal cliques. Clique enumeration is known to be NP-hard [2]. Even though the number of nodes involved in a global footprint n is likely to be much smaller than the grid size N, clique enumeration still presents a bottleneck in computational complexity.

However, for most applications involving sustained global footprints of events, the correlation threshold required is usually high. This sets in the triadic closure property of the cooccurrence graph, making clique enumeration much more efficient.

If the triadic closure property is known to hold, then a clique of n nodes can be detected with just n pairwise comparisons, drastically bringing down the complexity of clique enumeration to $O(n)$. Hence if $\{a, b, c, d\}$ is a clique of cooccurring footprints and the node e is known to have a high joint probability with at least one of a, b, c or d, it can be added to the clique if triadic closure is known to hold.

4 Experimental Results

Our Experimental setup includes a Grid Simulator which has been implemented to simulate the grid environment which has facilities to define, the *number of nodes* required in the grid, *number of event classes* with the *interarrival rate* for each event class, *number of time epochs* for profiling, *number of time intervals* for each epoch, the *threshold* Δ_j , the *threshold* δ_j and the *deviation threshold*. The grid simulator works in two modes a) Profiling mode b) Detection mode. In profile mode, randomly events are generated on every node based on the inter-arrival rate of that event class. Frequency is recorded for each event generated on the node and is updated in the appropriate time interval. Using the local information the profiler of the simulator creates the global signature in the form of graph and the cliques of the graphs. The simulator automatically switches to the detection mode after the profile period and creates the detection profiles.

In this particular experiment for which the result are mentioned, we aim to detect the accuracy of profiling the global behavior. For this experiment, we give input a synthetic graph to the simulator. The weights of the input graph is the joint probability between the nodes of the edge. The cliques & clique energy of this graph depicts the global profile of a single event class which is known in advance. We generate random events in the simulator and whenever a event is generated on any node of the graph, additional events are generated on the neighbors of the node with the conditional probability Pr(B/A) = $\frac{2J^{A,B}}{(1+J^{A,B})}$ where A and B are neighboring nodes. The events are generated for the duration of the profile period. After the profiling period is over, the local signatures are exchanged and the global profile is derived from the correlation graph. This experiment is run for different values of *threshold* Δ_j and for every Δ_j a correlation graph is obtained. During a single run, for any correlation value higher than the *threshold* Δ_j, results in an edge to be established between the nodes. The accuracy of the profiling is measured in terms of false positives and false negatives. We say a false positive (F_p) when there are extra edges in the global profile graph which are not in the Synthetic graph and a false negative (F_n) when one or more edges from the synthetic graph are missing in the generated global profile graph.

$$F_n = |E_s - E_p|, F_p = |E_p - E_s| \tag{5}$$

E_s = Edges of Synthetic graph E_p = Edges of the profiled graph

Figure 5 & Figure 6 shows the result obtained with a graph of 18 and 20 edges respectively with the maximum of 10 nodes. A graph with n nodes can have max of $\frac{n(n-1)}{2}$ edges. For a graph with 18 edges there can be maximum 27 more edges that can lead to false positive and 18 edges that can lead to false negative. In a graph of 20 edges there can be maximum of 25 edges that can lead to false positive and 20 edges that can lead to false negative. The Figure 5 & Figure 6 shows the sum of false positive and false negative obtained for different Correletion interestingess threshold Δ_j. The observation shows that at some threshold the false positive (F_p) and the false negative (F_n) is zero, which means that we have obtained the global profile graph same as the synthetic graph.

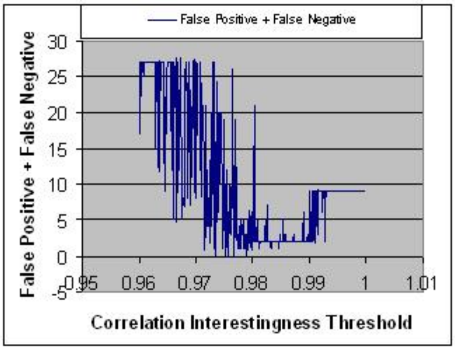


Fig. 5. Graph with 18 edges

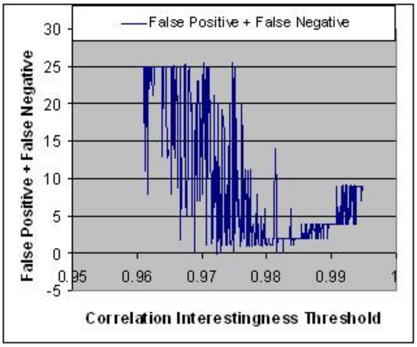


Fig. 6. Graph with 20 edges

5 Conclusion and Future Work

In this work, we have described the two class of Global anomalies a) Footprint anomalies b) Compositional anomalies and proposed a model for Global footprint anomaly detection in the grid environment. The proposed model is simulated and the initial results of profiling the global behavior of an event have included in the experimental section.

The future work involves identifying the deviation threshold upon which we can classify the event class as normal or anomalous. Also the current work have been carried, considering the correlation using joint probability, which leads to loss of the time importance. We intend to use the Pearson's correlation or other similar correlation techniques to have correlation graph of a event class which will take into account the time factor.

Acknowledgment

The method proposed in this paper has been accepted for Intellectual Ventures Pvt. Ltd. as patent No. 21421/0209362-US0 and the authors have been authorized to publish the method.

References

1. Agbaria, A., Sanders, W.H.: Distributed Snapshots for Mobile Computing Systems. In: Second IEEE International Conference on Pervasive Computing and Communications, PerCom'04 (2004)
2. Akkoyunlu, E.A.: The Enumeration of Maximal Cliques of Large Graphs. SIAM Journal on Computing 2, 1–6 (1973)
3. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: A survey. ACM Computing Survey 41(3) (July 2009)
4. Chandy, K.M.: Distributed Snapshots: Determining Global States of Distributed Systems. ACM Transactions on Computer Systems 3, 63–75 (1985)
5. Choon, O., Samsudin, A.: Grid-based intrusion detection system. In: Proceedings of 9th Asia-Pasicific Conference on Communications, September 21-24, vol. 3, pp. 1028–1032 (2003)
6. Dash, D., Kveton, B., Agosta, J.M., Schooler, E., Chandrashekar, J., Bachrach, A., Newman, A.: When gossip is good: Distributed probabilistic inference for detection of slow network intrusions. In: Proceedings of the 21st National Conference on Artificial Intelligence (2006)
7. Denning, D.: An Intrusion-Detection Model. IEEE Transactions on Software Engineering SE-13(2), 222–232 (1987)
8. Granovetter, M.: The Strength of Weak Ties. American Journal of Sociology 78(6), 1360–1380 (1973)
9. Huang, L., Nguyen, X.L., Garofalakis, M., Jordan, M., Joseph, A., Taft, N.: Distributed PCA and network anomaly detection. Technical Report UCB/EECS-2006-99, Electrical Engineering and Computer Science Department, University of California Berkeley (July 2006)
10. Kendall, K.: A database of computer attacks for the evaluation of intrusion detection systems. Master's thesis, MIT (1999)
11. Kenny, S., Coghlan, B.: Towards a grid-wide intrusion detection system. In: Sloot, P.M.A., Hoekstra, A.G., Priol, T., Reinefeld, A., Bubak, M. (eds.) EGC 2005. LNCS, vol. 3470, pp. 275–284. Springer, Heidelberg (2005)
12. Kindermann, R., Snell, L.J.: Markov Random Fields and Their Applications. American Mathematical Society. MR0620955 (1980), http://www.ams.org/online_bks/conm1/ ISBN 0-8218-5001-6
13. Mattern, F.: Efficient Algorithms for Distributed Snapshots and Global Virtual Time Approximation. Journal of Parallel and Distributed Computing 18, 423–434 (1993)
14. Rachakonda, A.R., Srinivasa, S.: Incremental Aggregation of Latent Semantics Using a Graph-Based Energy Model. LNCS. Springer, Heidelberg (2006)
15. Li, J., Sollins, K., Lim, D.-Y.: Dependency-based Distributed Intrusion Detection System. In: Proceedings of the DETER Community Workshop on Cyber Security Experimentaion and Test (2007)
16. Schulter, A., Reis, J.A., Koch, F., Westphall, C.B.: A Gird-based intrusion detection system. In: Proceedings of the international conference on Networking, International Conference on Systems and International conference on Mobile Communications and Learning Technologies ICNICONSMCL' 06 (2006)

17. Spafford, E.H., Zamboni, D.: Intrusion detection using autonomous agents. Computer Networks: The International Journal of Computer and Telecommunications Networking 34(4), 547–570 (2000)
18. Srinivasa, S., Patil, S.: A Symmetric Localization Algorithm for MANETs Based on Collapsing Coordinate Systems. In: Bader, D.A., Parashar, M., Sridhar, V., Prasanna, V.K. (eds.) HiPC 2005. LNCS, vol. 3769, pp. 73–82. Springer, Heidelberg (2005)
19. Tolba, M., Abdel-Wahab, M., Taha, I., Al-Shishtawy, A.: GIDA: Toward enabling grid intrusion detection systems. In: Proceedings of 5th IEEE/ACM Int. Symp. on Cluster Computing and the Grid (CCGrid 2005), Cardiff, UK (May 9-12, 2005)
20. Watts, D.J.: The New science of networks. Annual review of sociology 30, 243–270 (2004)
21. Ye, N., Chen, Q., Emran, S.M., Noh, K.: Chi-square Statistical Profiling for Anomaly Detection. In: Proceedings of the 2000 IEEE Workshop on Information Assurance and Security United States Military Academy, West Point, NY, June 6-7 (2000)
22. Fang-Yie, L., Jia-Chun, L., Ming-Chang, L., Chao-Tumg, Y.: Performance-Based Intrusion Detection System. In: Proceedings of 29th Annual IEEE International Computer software and Applications Conference (COMPSAC 2005), July 26-38, pp. 525–530 (2005)

Secure Anonymous Routing for MANETs Using Distributed Dynamic Random Path Selection

Vakul Mohanty[1], Dhaval Moliya[1], Chittaranjan Hota[1], and Muttukrishnan Rajarajan[2]

[1] Computer Sc. & Info Systems Group, Birla Institute of Technology and Science, PilaniHyderabad Campus, Hyderabad, Andhra Pradesh, India
[2] School of Engineering & Mathematical Sciences, City University, London
{vakul.mohanty,moliyadhaval}@gmail.com,
hota@bits-hyderabad.ac.in, r.muttukrishnan@city.ac.uk

Abstract. Most of the MANET security research has so far focused on providing routing security and confidentiality to the data packets, but less has been done to ensure privacy and anonymity of the communicating entities. In this paper, we propose a routing protocol which ensures anonymity, privacy of the user. This is achieved by randomly selecting next hop at each intermediate. This protocol also provides data security using public key ciphers. The protocol is simulated using in-house simulator written in C with OpenSSL crypto APIs. The robustness of our protocol is evaluated against known security attacks.

Keywords: Anonymity, routing, ad hoc networks, mobile, public key.

1 Introduction

Mobile Ad Hoc Networks (MANETs) are autonomous collection of mobile nodes without any fixed infrastructure that communicate over relatively bandwidth constrained wireless links, establishing dynamic communication. The nodes in a MANET may change its' position, adjust transmission and reception parameters causing links to be broken and re-established. A malicious node or an attacker can easily eavesdrop into the wireless channels and infer communication. A malicious node may even drop packets it had otherwise agreed to forward earlier. It may even go the extent of creating denial of service, exploited by injecting large number of unwanted packets into the network. So far, researchers in MANETs have generally studied the routing problem in a non-adversarial network setting, assuming a trusted environment; relatively little research has been done in a more realistic setting in which an adversary may attempt to disrupt the communication.

In this paper we present an anonymous routing protocol for a MANET. The protocol seeks to achieve anonymity with the minimal use of encryption and nullify the requirement of padding of data packet to prevent traffic analysis. In the protocol the next hop is dynamically selected by the router. This makes traffic analysis for a malicious router difficult as the traffic flow is erratic and confuses the adversary.

The rest of this paper is organized as follows: we present related work in Section 2; in Section 3, we present system model of our approach; and using this model, in Section 4, we present our Anonymous routing algorithm; in Section 5 & 6, we present the simulation results and analysis of our protocol respectively; finally, in Section 7, we summarize our work and point out several future research directions.

H. Chen et al. (Eds.): PAISI 2010, LNCS 6122, pp. 65–72, 2010.
© Springer-Verlag Berlin Heidelberg 2010

2 Related Work

Due to the nature of wireless environment and unavailability of fixed infrastructure, achieving security in MANET routing is a complex task. Onion routing [1,6] uses multiple layers of encryptions wrapped around the message. Each router in the path of the onion receives a message, performs a set of cryptographic operations on the message and then forwards it. Each router uncovers a layer of encryption using its private key, this allows it to access routing instructions for the next router. This process continues until the message reaches the last router. Papadimitriou and Haas [2] proposed a secure routing protocol for MANETs using a security association between source and destination to validate the integrity of a discovered route. Sanzgiri et al. [3] have proposed cryptographic ways to secure routing in MANETs wherein every intermediate node verifies the integrity of the message and then forwards it to the next node. Certificates are used by source and destination nodes to get the public key of each other. ASR [4] uses anonymous virtual circuit in routing and data forwarding where each node does not know its immediate upstream nodes and immediate downstream nodes. Using a special anonymous signaling procedure, the node only knows the physical presence of neighboring ad hoc nodes. The session key of the route between every pair of the intermediate nodes is determined when a node forwards reply packet to its upstream nodes. Although the above mentioned anonymous routing techniques can provide a certain level of anonymity, an external adversary can still monitor the transmitted packets to identify the communication peers [5].

3 System Model

We explain here the notations, assumptions and the system model. An example of our approach is shown in fig. 1.As depicted in fig. 1, every node in the network maintains ART and ARC. Destination maintains PIT and IRT as well. Source node starts with the route discovery message (routeRequest) by flooding it to all neighbors.Request id is embedded in it.

Notations used:

S	:Sender	**R**	: Receiver
M	:Message	**D**	: Data
X	:Intermediate node	**E**	: Encrypt function
ccCount	:Criss-cross count	**ccTimer**:	Criss-cross timer
PU$_N$	:Public key of N	**PR**$_R$	: Private Key of N
D	:Decrypt function	**ccTable**	: Criss cross Timer Table
ART	:Anonymous Routing Table		
IRT	:Intermediate Routing Table<pathID, path_of_message>		
PIT	:Path Info Table <pathID, nodeID, nextHop>		
∪	:Set inclusion, modeled as appending at the end of set (array)		
ARC	:Anonymous Routing Cache <reqID, ccCount>		
exists(x,z)	:returns true if table z has record mapped to x, otherwise false.		
getCnt(x)	: returns ccCount value from the record <x, ccCount> of ARC, if no such record found then return false.		
setCnt(x, y)	: sets ccCount value from the record <x, ccCount> of ARC to y.		
expired(x)	:returns true if timer mapped to x is expired else false.		

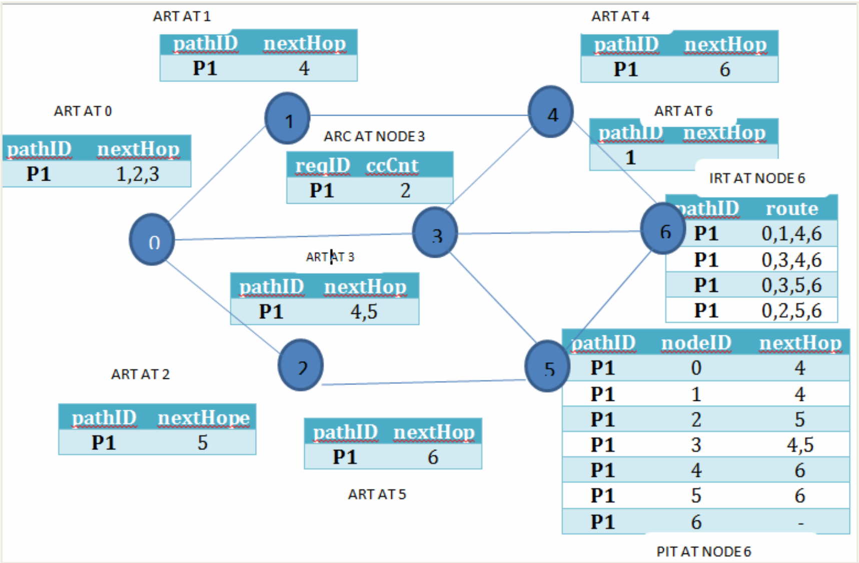


Fig. 1. An Example of Secure Anonymous Routing Protocol

Data portion D of the message contains pathID, Source(S), Destination (D) and Nonce$_s$ which is encrypted using the public key of destination (PU$_D$). Message also contains I, set of all nodes traversed by routeRequest message to reach destination, and PU$_D$.Every entry in I is encrypted using PU$_D$.

ccCnt indicates the number of more routeRequest messages with same request id that can be flooded by the same node. Initially assigned value to ccCnt is a parameter set by network administrator, subject to tuning. Upon receiving routeRequest message with given reqID first time, node makes an entry in ARC(as shown in fig 1, at node 3), inserting reqID and ccCnt. For subsequent receipts of routeRequest message with same reqID, node checks whether value of ccCnt is zero or not. If zero then ccCnt limit is reached and packet is discarded there itself. If not zero then ccCnt is decremented by 1 and message is forwarded to neighbors by appending its id(encrypted with PU$_D$) in I field of the message.

Upon receiving the route request message, destination node takes action as explained in previous paragraph with following additional steps: it sets ccTimer for given reqID. ccTimer also acts as a threshold like ccCnt, but it is used to filter optimal paths. It is also subject to tuning by network administrator. Destination decrypts D and I part of message using its private key(PR$_D$) to extract NONCE$_S$, pathID and all en-route nodes which are entered into IRT (i.e. <P1, <0,1,4,6>> of IRT at node 6 in fig. 1). There is one to one mapping between pathID and reqID. Whenever ccCnt becomes zero or ccTimer expires for a given reqID, node creates PIT from IRT to be used for updating routing tables of en-route nodes. As depicted in fig. 1, for IRT entry <P1, < 0, 1, 4, 6>>, it updates PIT entries as <P1, 0, <1>>, <P1, 1, <4>>, <P1, 4, <6>>, <P1, 6, <>>. For <P1, < 0, 3, 4, 6>>, it updates PIT as <P1, 0, <1, 3>>, <P1, 3, <4>>, <P1, 4, <6>>, <P1, 6, <>>. This is used to construct routeReply messages, composed of routing table updates of en-route nodes. Destination node encrypts these

PIT entries with the public keys of en-route nodes, in the sequence marked in the routeRequest message and onion routing [1, 6] is used to forward these updates to en-route nodes. $NONCE_R$, $F(NONCE_S)$ are also added to message encrypted using public key of Source.

Upon receiving the routeReply message, the intermediate node removes outermost layer from the onion, does appropriate cryptographic operations on it, updates its ART from the update received, and forwards the message to the next hop. Upon receiving the routeReply message, the source node updates it's ART as explained for en-route nodes. It receives $NONCE_R$ and $F(NONCE_S)$, which are used for authentication and preventing the replay attack. For regular data transfer, source uses the pathID, and selects the next hop randomly from its ART. Every en-route node also does the same for selection of the next hop.

Here we assume that any node leaving the network does not cause the partition in the MANET. Every node(X) sends a beacon to its neighbor, and updates the status of neighbors depending upon the reply. If any node $N_L$discovers change in the topology of network then it searches <pathID, Z> in ART such that $N_L \in Z$. If such entry is found then it sends the update message to all nodes in Z and removes N_L from Z. After removing the entry, if Z is empty then node floods the route invalidate message with corresponding pathID. Upon receiving the update message, node updates its ART. In case the node receiving the invalidate message is the one that started the communication with the corresponding pathID, it re-initiates route discovery.

4 Proposed Algorithm

4.1 Path Discovery Phase

```
Source initiates with routeRequest message<reqID, E(PU_{R,D}), I> ;D=, <pathID, sourceID,
destinationID, NONCE_S>, I={E(PU_R, S)} by sending to all neighbors.

X (≠R), an intermediate node receives routeRequest<reqID, E(PU_R,D), I>message:
if ( $exists(reqID, ARC) \wedge getCnt(reqID) \neq 0$)then
    setCnt(reqID, getCnt(reqID) - 1)          /* decrement the ccCount */
    $I \leftarrow I \cup \{ E(PU_R, X)\}$                               /* Append ID to the message*/
    forward  <reqID, E(PU_R, D), I>  to neighbors except the one from it received.
elif($exists(reqID, ARC) \wedge getCnt(reqID) = 0$)then
    Discard routeRequest message as ccCnt limit reached.
else
    ARC $\leftarrow ARC \cup$ {<reqID, ccCntUL>}      /* Make entry in ARC */
    $I \leftarrow I \cup \{ E(PU_R, X)\}$                               /* append ID in message */
    forward<reqID, E(PU_R, D), I> to neighbors.
endif
Send acknowledgement to the node (sourceID) from which message is received.

R receives routeRequest message M_{RQ}<reqID, sourceID, destinationID, E(PU_R, D), I>:
Decrypt each entry of I private key, store decrypted values in I.
if ( $exists(reqID, ARC) \wedge expired(ccTimer_{reqID})$ )then
            Discard routeRequest message.    /* Timer Expired */
elif ( $exists(reqID, ARC) \wedge \neg expired(ccTimer_{reqID}) \wedge getCnt(reqID) = 0)$)then
            Discard routeRequest message.   /* Criss-cross count limit reached */
elif ( $exists(reqID, ARC) \wedge \neg expired(ccTimer_{reqID}) \wedge getCnt(reqID) \neq 0)$)then
```

```
            setCnt ( reqID, getCnt(reqID) - 1)
            pathID← D($PR_R$, E($PU_R$, D))
            IRT ←IRT∪ {<pathID, I ∪ {R} >}
else        (¬exists(reqID, ARC), ARC)     /* No entry found in ARC for reqID */
            ccTable←ccTable∪ {<pathID, $ccTimer_{Mrq}$>}           /* Set ccTimer*/
            ARC ←ARC∪ {<reqID, 5>}
            pathID← D($PR_R$, E($PU_R$, D)
            IRT ←IRT∪ {<pathID, I ∪ {R} >}
endif
```

4.2 Construction of Routing Table Entries for Intermediate Nodes

```
/*Process entries in IRT with reqID for with ccCnt is zero or ccTimer is expired*/
for each <reqID, I> in IRT  do
            for each $x_i$ in I;$x_i$≠R,do
                        if exists(<pathID, $x_i$,  Z>, PIT)then
                                    Update Z←Z∪{ $x_{i+1}$} in PIT
                        else
                                    PIT ←PIT∪ {<pathID,$x_i$, { $x_{i+1}$}> }
                        endif
            end for
end for
```

4.3 Updating ART of Intermediate Nodes

```
Constructing and sending Reply Message:
for each <pathID, I>in IRT do
            I'= ϕ                                   /* initialize I' as Null*/
            for each $x_i$ in I, i=n…1 do           /* Reverse the path for reply message */
                        I'=I' ∪ {$x_i$}
            endfor
            temp=< $NONCE_R$ , F($NONCE_S$)>
            for each $x_i$ in I', $x_i$≠R do
                        Search <pathID, $x_i$ , Z> in PIT
                        if  i=1 then                /* for source node's case */
                                    msg = msg + <S , E($PUx_i$, <<pathID, $x_i$ , Z>, temp>) >
                        else        msg = msg + <$x_{i-1}$ , E($PUx_i$, <pathID, $x_i$ , Z>) >
                        endif
            end for
Send routeReply message $M_{RP}$ <R, $x_{n-1}$, msg> to $x_{n-1}$      /*<source, to, msg_data> */
end for

X receives the routeReply message $M_{RP}$<Y, X, msg>:
/* extract the routing info sent by the destination and update ART */
<<pathID, X , Z >, nextHop, E($PU_{nextHop}$ , msg) > = D($PR_x$, msg)
ART = ART ∪ {<pathID,  Z>}                         /*Update routing table*/
if X=S then
            <pathID, $NONCE_R$ , F($NONCE_S$)> = D($PR_S$, msg)
            Send F($NONCE_R$) to the destination.
else
            forward$M_{RP}$<X, nextHop, msg>
endif
```

4.4 Data Communication Phase

Source-destination pair exchanges session key for regular data transfer. Source sends message with pathID prepended to the message. Every intermediate node will choose the next hop dynamically from its ART corresponding to the pathID in the message. We have employed acknowledgement mechanism for detection of passive nodes.

5 Simulation Results

We have written our simulator using C in UNIX. All cryptographic operations are performed using OpenSSL Crypto API. MANET is constructed using 50 nodes, initially uniformly distributed. Source destination pairs are chosen randomly. Mobility of nodes is random, with constant speed. Once node becomes immobile, it waits there for fixed time. Maximum number of communicating pairs in MANET at a given time is assumed to be 20, chosen randomly. We use cc_cnt, and cc_timer metrics as global tunable parameters which are set by network administrator. As depicted from the cc_cnt vs delay graph in fig. 2, by increasing the value of ccCnt, the number of paths discovered is more. However few of these paths might be longer ones. So the delay incurred on an average to reach the destination also increases.

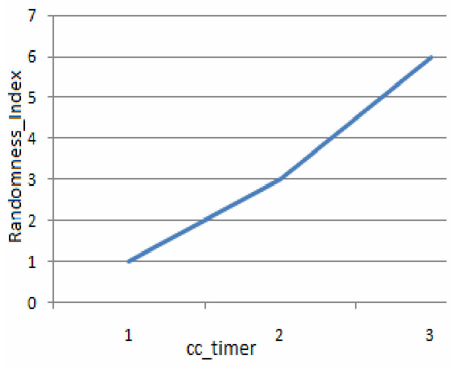


Fig. 2. cc_timer vs Randomness_index

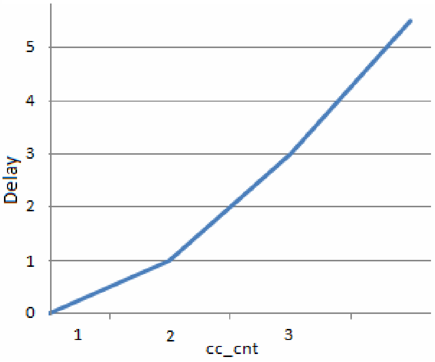


Fig. 3. cc_cnt vs Delay

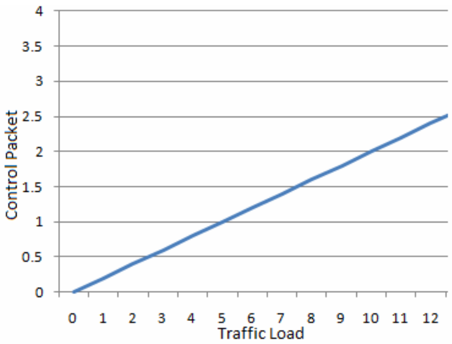


Fig. 4. Traffic Load vs Control Packets

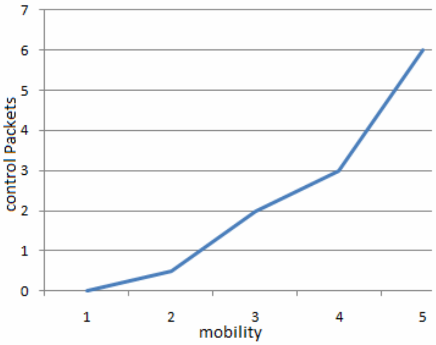


Fig. 5. Mobility vs Control Packets

Fig. 3 shows the cc_timer vs randomness_index, the average number of nodes in each intermediate node's ART, for a given pathID. The linear increase in this randomness_index guarantees higher anonymity because of the fact that more number of paths is now available to make traffic analysis increasingly difficult in the MANET. Fig. 4 shows the mobility vs control packet and fig. 5 shows traffic load vs control packets. Both the figures show that increase in mobility increases number of the control packets. This is because of flooding many packets that include path invalidation, path update and route rediscovery messages.

6 Simulation Analysis

6.1 Anonymity Analysis

Identity Privacy: In our protocol, the identities of source and destination are known to only two communicating parties, as we are using them only in the route request message and with encryption, thereby not revealing them to intermediate nodes. Hence identity privacy is ensured.

Route Anonymity: In our protocol, no adversary can trace a flow of packet because of random selection of next hop and thereby leading to dynamic path selection. Any adversary on the route has no information about the path other than the next hop. As we have employed fixed size padding, we can introduce several dummy packets and reshuffling of actual packets in the buffer to eliminate the possibility of temporal analysis as defined in [12]. Thus all the requirements of route anonymity are satisfied.

6.2 Possible attacks

Route Rediscovery Attack: One possible attack is that adversaries send fake route update or route invalidate packets to fool the intermediate nodes or source to begin route rediscovery process. In our protocol, only the nodes whose routing table has entry for the node leaving the network, can send the route invalidate, route rediscovery or route update messages whichever applicable as explained in the algorithm. So our proposed protocol is less vulnerable to the route rediscovery attack.

Selfish Nodes or Byzantine nodes: Byzantine nodes can intercept packets, create routing loops, selectively drop packets, or purposefully delay packets. Our protocol uses the acknowledgement mechanism. If any node is dropping the packet then acknowledgement will not be sent to sender. Even in presence of live communication link, if node is dropping packets then it can be detected as selfish node. And as we are choosing next hope dynamically at any intermediate node for routing, we can exclude this selfish node from the ART.

6.3 Cryptographic Overhead

In our protocol, we use cryptosystem of the form onion only for path discovery. For data communication, data is encrypted by source with the destinations' public key, i.e.

end to end encryption; onion routing is not used here. So there is not much cryptographic overhead involved for normal data communication phase that leads to computational advantage.

7 Conclusions

This has paper has proposed a new routing protocol in mobile ad hoc networks with anonymity and provable security. We have stressed upon the anonymity, which is becoming one of the most important aspect in securing the next generation mobile ad hoc networks. The developed protocol has been evaluated with respect to anonymity and known security threats. Simulation results give the performance of our protocol. Our future work will aim at overcoming Distributed Denial of Service (DDoS) attack, and estimating the cryptographic computation overhead in this type of environment. We will also focus on improving security by adopting strong peer to peer authentication in the route discovery phase using extensive simulations.

References

1. Reed, M., Syverson, P., Goldschlag, D.: Anonymous connections and Onion Routing. IEEE J. Selected Areas in Commun. 16(4), 482–494 (1998)
2. Papadimitratos, Haas: Secure Routing for Mobile Ad hoc Networks. In: SCS Communication Networks and Distributed Systems Modeling and Simulation Conference (CNDS 2002), San Antonio, TX, January 27-31 (2002)
3. Sanzgiri, K., Dahill, B., Levine, B.N., Shields, C., Belding-Royer, E.M.: A Secure Routing Protocol for Ad Hoc Networks. In: Proceedings of 2002 IEEE International Conference on Network Protocols, ICNP (November 2002)
4. Zhu, B., Van, Z., et al.: Anonymious secure routing in mobile ad hoc networks. In: 29th IEEE International conference on local computer networks LCN'04, pp. 102–108 (2004)
5. Huang, D.: Traffic analysis-based unlinkability measure for ieee 802.11b-based communication systems. In: Proceedings of ACM Workshop on Wireless Security, WiSe (2006)
6. Ren, J., Wu, J.: Survey On Anonymous Communications In Computer Networks. Comput. Commun. 33(4), 420–431 (2010)

MAD-IDS: Novel Intrusion Detection System Using Mobile Agents and Data Mining Approaches

Imen Brahmi[1], Sadok Ben Yahia[1], and Pascal Poncelet[2]

[1] Faculty of Sciences of Tunis, Tunisia
sadok.benyahia@fst.rnu.tn
[2] LIRMM Montpellier, France
Pascal.Poncelet@lirmm.fr

Abstract. Intrusion Detection has been investigated for many years and the field reached the maturity. Nevertheless, there are still important challenges, *e.g.*, how an Intrusion Detection System (IDS) can detect distributed attacks. To tackle this problem, we propose a novel distributed IDS, based on the desirable features provided by the mobile agent methodology and the high accuracy offered by the data mining techniques.

Keywords: Intrusion Detection System, Mobile Agents, Data Mining Techniques.

1 Introduction

As an important gatekeeper of network, *Intrusion Detection Systems* (IDS)s must have the ability to detect and defend intrusions more proactively in shorter period. However, most current IDSs are centralized and thus a central analyzer presents a favorable target to the attackers.

In this paper, we investigate another way of tackling this problem. Moreover, according to the advantages of mobile agent technology, which include: *reducing network overload*, *overcoming network latency*, *system scalability*, etc, this technology seems to be very suitable to solve intrusion detection in a distributed environment [2]. Thus, we introduce a new distributed IDS, called **MAD-IDS** (*Mobile Agent using Data mining based Intrusion Detection System*). The MAD-IDS system integrates the data mining techniques and the mobile agent methodology in order to detect both known and unknown attacks.

2 The MAD-IDS System

Figure 1 provides an overall architecture of MAD-IDS. Its distributed structure comprises different agents which are able to move from one station to another, called respectively: *Sniffer, Filter, Misuse Detection, Anomaly Detection, Rule Mining* and *Reporter Agent*.

Each of these agents will be individually described in the following subsections.

H. Chen et al. (Eds.): PAISI 2010, LNCS 6122, pp. 73–76, 2010.
© Springer-Verlag Berlin Heidelberg 2010

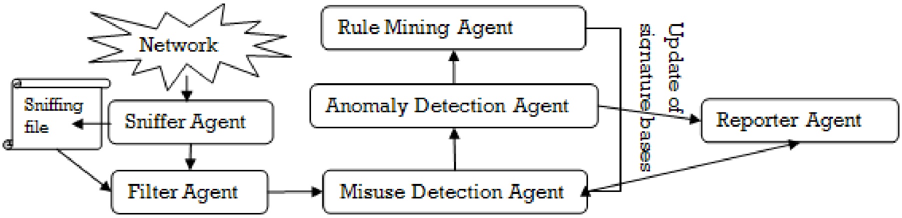


Fig. 1. The architecture of MAD-IDS

2.1 The Sniffer Agent

The Sniffer Agent collects the network packets and stores them in a "*sniffing file*". The benefits of this kind of agent include: *i*) the cloning and the distribution throughout the network; and *ii*) the duplication in order to lighten the network charge.

2.2 The Filter Agent

The Filter Agent aggregates and merges events stored in the sniffing file. It performs its tasks as a pretreatment phase, which precedes the analysis phase.

2.3 The Misuse Detection Agent

The Misuse Detection Agent detects known attacks in network connections. Hence, if there is a similarity between the filtered packets and attacks signatures, then the agent raises an alert to the Reporter Agent, and then removes these anomalous packets from further analysis.

Although the known attacks are detected, it remains nevertheless the problem of the new attacks detection. One answer to the problem could rely on data mining techniques.

2.4 The Anomaly Detection Agent

The Anomaly Detection Agent provides the combination of distributed IDS with clustering techniques. The clustering-based anomaly detection algorithm is based on the steps described as follows:

Step 1 (Initialization): Partition the training data into k clusters;
Step 2 (Assignment): Assign each instance to its closest center;
Step 3 (Updating): Replace each center with the mean of its members;
Step 4 (Iteration): Repeat Steps 2 and 3 until there is no more updating;
Step 5 (Anomaly finding): For each test instance Z:

1. Compute the Euclidean distance between Z an initial cluster C_i;
2. Find cluster C_i that is closest to Z;
3. Classify Z as an anomaly or a normal instance using a pre-defined *Threshold*.

2.5 The Rule Mining Agent

The Rule Mining Agent provides the construction of a summary of anomalous connections detected by the Anomaly Detection Agent. To mine association rules, we apply the *Informative Generic Basis* ($\mathcal{IGB}$) [1]. In addition to the elimination of redundancy, the application of the $\mathcal{IGB}$ basis during an intrusion detection process provides: *the increase of the overall coverage of detectable attacks* and *the maximum convey of useful knowledge*, while being *the information lossless* [1]. Therefore, the database of signatures of the Misuse Detection Agent can be updated regularly by the addition of the extracted rule set.

2.6 The Reporter Agent

The Misuse and the Anomaly Detection Agents send their findings as alerts to the Reporter Agent which transmits to the system administrator.

3 Experimental Results

During experiments, we partly use the traffic data DARPA[1]. Table 1 shows the distributions of record types in training and testing datasets, used during our experiments.

Table 1. The considered datasets at a glance

Record Type	Training Set	Testing Set
Normal	48886	27322
Intrusion	37804	23009

In fact, to evaluate the performance of an IDS, two interesting metrics are usually of use: the *Detection Rate* (DR), which indicates the number of correctly detected intrusions; and the *False Positive Rate* (FR), which calculates the number of instances that were incorrectly considered as attacks.

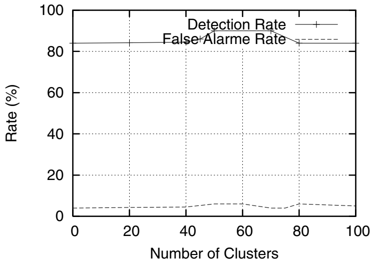


Fig. 2. DR and FR using the clustering-based anomaly detection algorithms

[1] Available at: `http ://www.ll.mit.edu/IST/ideval/data/data_index.html`

Figure 2 plots the DR and FR for the considered datasets, using our clustering-based anomaly detection algorithm.

As shown in Figure 2, the best performance of our anomaly clustering-based algorithm was obtained when DR = 89.89% and FR = 1.00%.

4 Conclusions

In this paper, a novel distributed multi-agent IDS architecture, called MAD-IDS was presented. MAD-IDS integrates the mobile agent methodology and the data mining techniques to accommodate the special requirements in distributing IDS. The preliminary experimental results indicated that the data mining algorithms used in MAD-IDS are feasible for detecting attacks within a distributed environment.

References

1. Ben Yahia, S., Gasmi, G., Mephu Nguifo, E.: A New Generic Basis of Factual and Implicative Association rules. Intelligent Data Analysis (IDA) 13(4), 633–656 (2009)
2. Jaisankar, N., Saravanan, R., Swamy, K.D.: Intelligent Intrusion Detection System Framework using Mobile Agents. International Journal of Network Security and its Applications (IJNSA) 1(2), 72–88 (2009)

Fuzzy Biometric Signcryption Scheme with Bilinear Pairings in the Standard Model

Mingwu Zhang[1,2,3], Bo Yang[1], Tsuyoshi Takagi[2],
Yanping Shen[1], and Wenzheng Zhang[3]

[1] Department of Computer Science and Engineering,
College of Informatics, South China Agricultural University,
Guangzhou 510642, China
csmwzhang@gmail.com

[2] Faculty of Mathematics, Kyushu University, Fukuoka, 819-0395, Japan

[3] National Laboratory for Modern Communications, Chengdu 610041, China

Abstract. A new cryptographic primitive, called fuzzy biometric signcryption scheme which can perform confidentiality and authentication in a logic step while the sender and receiver identities are based on the fuzzy biometric string, is proposed. The proposed scheme possesses the error-tolerance property that allows a user with the private key for fuzzy biometric string b to decrypt a message for string b' if and only if b and b' are within a distance t. After given the scheme model and security definitions, a fuzzy biometric signcryption scheme in the standard model is proposed and then then security are analyzed including indistinguishable against adaptively chosen ciphertext attack and existentially unforgeable against adaptively chosen message attack.

Keywords: Biometric cryptographic, signcryption, threshold secret sharing, standard model.

1 Introduction

In 2004, Boneh and Boyen [1] first proposed a short signature scheme without using random oracles from bilinear groups, which is fully secure with the support of existentially unforgeable under a chosen message attack where its security model reduces to q-strong Diffie-Hellman assumption that is a stronger assumption compared with the standard computational Diffie-Hellman(CDH) assumption. Their construction is too inefficient to be of practical use and then they fixed and expanded the previous scheme in [2]. Boneh and Boyen [3] improved upon this result by constructing an efficient scheme to be secure in the selective identity model, and then they described a scheme that is fully secure without random oracles [4]. However, their construction is also inefficient. Based on Boneh and Boyen's primary ideal and consider, Waters [5] first constructed an efficient and provable security encryption scheme under without random oracle model, at the same time he proposed an open issue to regard compact public parameters. Recently, several signature/signcryption schemes using the bilinear

H. Chen et al. (Eds.): PAISI 2010, LNCS 6122, pp. 77–87, 2010.
© Springer-Verlag Berlin Heidelberg 2010

pairings without random oracles have been proposed that are efficient and at the same provable security in the standard model [7,8,9,10].

Identity based cryptosystems were first introduced by Shamir [11] where participators use their identities information such as name, identity card number, IP address or email address as his public key, which is a good alternative for certificate-based systems from the viewpoint of efficiency and convenience. Provable security is the basic model for identity based cryptographic schemes. In [12], the authors first proposed the fuzzy identity based encryption where allows a user with the private key for identity b to decrypt a ciphertext encrypted for identity b' if $\mathrm{dis}(b, b') < t$, but the proposed scheme was security in the random oracles model such that the random oracles are instantiated with concrete hash functions. So Gentry [6] proposed an IBE scheme that is fully secure in the standard model with short public parameters and a tight security reduction, where the ciphertext does not leak the identity of the recipient. Motivated by Waters scheme without random oracles [5], Yang et al. [13] proposed a signature scheme in the standard model.

In this paper, we first proposed a fuzzy biometric based signcryption scheme to perform digital signature and encryption simultaneously at lower computational costs and communication overheads than sign-then-encrypt way to obtain private and authenticated communications while it supports the fuzzy identity parse such that one can decrypt the ciphertext if and only if fuzzy identity string is closer for signcryptor biometric b and descryptor biometric b' that satisfies $\mathrm{dis}(b, b')$ less than designated threshold t. We also give the concrete signcryption model and provide the security analysis including indistinguishability against chosen ciphertext attacks and unforgeability against chosen message and identity attacks.

Roadmap. The rest of the paper is organized as follows: section 2 gives some preliminaries and blocks including threshold secret sharing, bilinear pairing, security assumption and security notations and definitions. The concrete model of fuzzy biometric signcryption scheme is described in section 3. Security analysis including confidentiality and unforgeability are provided in section 4, and conclusion is drawn in section 5.

2 Preliminaries and Blocks

2.1 Shamir Threshold Secret Sharing

Shamir's Secret Sharing scheme is an algorithm in cryptography using a polynomial interpolation. It is a form of secret sharing, where a secret is divided into parts, giving each participant its own unique part, where some of the parts or all of them are needed in order to reconstruct the secret.

Let GF(q) be a finite field with $q \geq n$ elements, and let $s \in$ GF(q) be the secret to be shared. The dealer picks a polynomial $f(x)$ of degree $t-1$ randomly, and the constant of $f(x)$ is secret s where

$$f(x) = s + \sum_{i=1}^{t-1} a_i x^i \tag{1}$$

After we assign each user $\mathcal{U}_i$ with a unique element ϑ_i, the dealer sends him/her the secret share $s_i = f(\vartheta_i)$. If a subset of users $N \subset P$ such that $|N| >\geq t$, they can cooperate to recover the secret $s = f(0)$ by

$$f(x) = \sum_{\mathcal{U}_i \in N} \varDelta_{\vartheta_i,N}(x) f(\vartheta_i) = \sum_{\mathcal{U}_i \in N} \varDelta_{\vartheta_i,N}(x) s_i \tag{2}$$

where

$$\varDelta_{\vartheta_i,N}(x) = \prod_{\mathcal{U}_j \in P, j \neq i} \frac{x - \vartheta_j}{\vartheta_j - \vartheta_i} \tag{3}$$

Lemma 1. (Biometric Distribution Assumption) *Let H be the distance function in a metric Hamming space. Suppose b_i and b_j are the reference biometric templates for Alice and Bob, respectively. There is a threshold value t, the probability that $H(b_i, b'_j) > t$ is close to 1 and the probability that $H(b_i, b'_i) \leq t$ is close to 1, where b'_i and b'_j are the templates captured for Alice and Bob at any time.*

2.2 Bilinear Groups and Complex Assumptions

Definition 1. (Bilinear group) *Let $G =< g >$ be a cyclic group of prime order p with a bilinear map e from $G \times G$ to G_T, i.e., for all $u, v \in G$ and $a, b \in Z_p$, it holds that $e(u^a, v^b) = e(u, v)^{ab}$ and e is non-trivial, i.e., $e(g, g) \neq 1_{G_T}$. Note that $|G_T| = p$.*

Definition 2. (Computational Diffie-Hellman (CDH) assumption) *Given $(g, g^a, g^b) \in G^3$ for any $a, b \in Z_p$, to compute g^{ab} is a hard problem.*

Let Ω be a CDH parameter generator. We say that an algorithm B has the advantage $Adv_{\Omega,\mathrm{B}}(k)$ in solving the CDH problem for Ω in time at most $t(k)$ if for sufficiently large parameter k, i.e.,

$$Adv^{CDH}_{\Omega,B}(k) = Pr|[\mathrm{B}(g, g^a, g^b) = g^{ab} : a, b \leftarrow Z_p, g \leftarrow G] - \frac{1}{2}| \tag{4}$$

2.3 Formal of Fuzzy Biometric Identity Based Signcryption Scheme

A fuzzy biometric identity based signcryption scheme consists of the following four algorithms.

- Setup$(1^k, n, d)$: Given a security parameter k and biometric string length n together with threshold d, PKG generates a system master key s and common parameters *params*.
- Keyext(msk, b): Given a biometric string $b \in \{0, 1\}^n$, the PKG runs this algorithm to generate the private key D_b associated with b.
- Signcrypt$(params, D_{b_A}, b_B, m)$: To send a message m to Bob whose biometric string is b_B, Alice with biometric string b_A obtains a ciphertext C by computing $Signcrypt(m, S_{b_A}, b_B)$.

- Unsigncrypt($params, D_{b_B}, b_A$): After Bob receives the ciphertext C, he computes $Unsigncrypt$ (C, D_{b_B}, b_A) and obtains the message m or the symbol $\perp$ indicating that the ciphertext is invalid.

2.4 Security Notations

We formalize the bSC model with two security notations such as *confidentiality* for adaptive chosen ciphertext adversaries (IND-bSC-CCA2), and *unforgeability* for adaptive message attack adversaries(UNF-bSC-CMA2) .

Definition 3. Confidentiality. *A fuzzy biometric based signcryption scheme is said to have the indistinguishability against adaptive chosen ciphertext attacks property (IND-bSC-CCA2) if no polynomially bounded adversary has a non-negligible advantage in the following game. The game bSC for semantic security is defined as:*

- Setup: The distinguisher B runs the Setup algorithm with a security parameter k and n, d, and sends the public parameters $params$ to adversary A.
- Query 1: Adversary A performs key extract algorithm Keyext queries, Signcrypt queries, Unsigncrypt queries adaptively.
- Challenge: A chooses two plaintext m_0, m_1, sender fuzzy biometric string b_A^* and receiver fuzzy biometric string b_B^* on which he wants to be challenged. In this stage A cannot perform the key extract query corresponding to b_B^*. B picks a random c from $\{0, 1\}$ and computes $C^* = Signcrypt(m_c, D^*_{b_A}, b_B^*)$ and sends C^* to A.
- Query 2: The adversary A can do a polynomially bounded number of queries adaptively again as in the first stage with the restriction that he cannot make the key extraction query Keyext on b_B^* and Unsigncrypt query on C^*.
- Response: Finally, adversary A returns a bit c' and wins the game if $c' = c$.

Definition 4. *The bSC scheme is semantic security if adversary A obtains the advantage $Adv(A)$ is negligible in bSC game. The adversary's advantage is defined as*

$$Adv^{\texttt{IND-bSC-CCA2}}(\mathrm{A}) = |Pr[b' = b] - 1/2| \tag{5}$$

Definition 5. Unforgeability. *The bSC scheme based on fuzzy biometric is existentially unforgeable against chosen-message attack (EUF-bSC-CMA2) if no PPT forger $\mathcal{F}$ has a non-negligible advantage in the following game.*

- Challenger runs Setup just like in bSC game.
- Forger $\mathcal{F}$ adaptively performs a series of queries just like in bSC game.
- $\mathcal{F}$ outputs a ciphertext (C^*, b_A^*, b_B^*), and wins the game iff:
  (a)Ciphertext C^* is not produced by Signcrypt oracle,
  (b)$\mathcal{F}$ has not queried the b_A^*'s secret key, and
  (c)Unsigncrypt$(C^*, b_B^*, b_A^*) \neq \perp$.

The adversary's forgery goal is the existential forgery of a signcryption scheme. We give the adversary the power to choose the biometric identity on which wishes to forge a ciphertext, the power to request the Keyext algorithm adaptively. The adversary is also given access to a Signcrypt and Unsigncrypt oracle on any desired biometric strings.

3 Fuzzy Biometric Signcryption Scheme

In this section, we describe our fuzzy biometric based signcryption scheme called bSC in the standard model.

The decription that follows assumes that group G and G_T with the same prime order p and g is a generator of G, e is a bilinear pairing e : $G \times G \rightarrow G_T$. We also assume biometric string be sets of n elements of Z_p.

- Setup: Random pick a secret value $\alpha \in Z_p$ and $g_2 \in G$, compute $g_1 = g^{\alpha}$ and $A = \mathrm{e}(g_1, g_2)$; Let N be the set {1,...,n+1}, pick $t_1, ..., t_{n+1} \in G$ randomly; Define a function T, as

$$T(x) = g_2^{(x^n)} \prod_{i=1}^{n+1} t_i^{\Delta_{i,N}(x)} \tag{6}$$

  Next, random select a $m' \in G$ and a n_m vector $\mathbf{m} = (m_1, ...m_{n_m}) \in Z_p^{n_m}$ where n_m is the plaintext length, and compute $w_1 = g^{m_1}, ..., w_{n_m} = g^{m_{n_m}}$. The public parameters $params = (g_1, g_2, t_1, ..., t_{n+1}, m', w_1, ..., w_{n_m}, A)$ and the system master key $msk = g_2^{\alpha}$.
- Keyext: On inputting the fuzzy biometric string $b \in \{0,1\}^n$, pick a random d-1 polynomial q such as $q(0) = y$. Random pick $r_1, ..., r_n \in Z_p$ and return secret key $D_b = (\{u_{b_i}\}_{i \in b}, \{v_{b_i}\}_{i \in b})$ of biometric b, where

$$u_{b_i} = g_2^{q(i)} T(i)^{r_i} \tag{7}$$

$$v_{b_i} = g^{-r_i} \tag{8}$$

  Therefore, the sender Alice and the receive Bob's private keys are

$$D_{b_A} = (\{u_{A_i}, v_{A_i}\}_{i \in b_A}) = (g_2^{q_A(i)} T(i)^{r_{A_i}}, g^{-r_{A_i}})$$

$$D_{b_B} = (\{u_{B_i}, v_{B_i}\}_{i \in b_B}) = (g_2^{q_B(i)} T(i)^{r_{B_i}}, g^{-r_{B_i}})$$

- Signcrypt: To send a message $m \in G_T$ to Bob, the sender Alice does
  (a) Random pick $r_1, ..., r_{n_m} \in Z_p$, compute $t = \sum_{i=1}^{n_m} r_i$ and $C_1 = m \cdot A^t$;
  (b) Compute $C_2 = g^{-t}$;
  (c) Compute $C_3 = \{T(i)^{r_i}\}_{i \in b_B}$;
  (d) Compute $\widehat{m} = H(m) \in \{0,1\}^{n_m}$ which is a n_m-bit string, compute $C_4 = u_{A_i}(m' \prod_{j=1}^{n_m} w_j^{r_j} \widehat{m}_j)$;
  (e) Compute $C_5 = \{v_{A_i}\}_{i \in b_A}$;
  The ciphertext is $C = (C_1, C_2, C_3, C_4, C_5) \in G^5$.

- Unsigncrypt: On receiving a ciphertext $C = (C_1, C_2, C_3, C_4, C_5)$, Bob can decrypt the ciphertext using his secret key $(\{u_{B_i}, v_{B_i}\}_{i \in b_B})$ as follows
  (a) Compute $m = c_1 \frac{\prod_{i=1}^{n} \mathrm{e}(v_{B_i}, c_{3_i})}{\prod_{i=1}^{n} \mathrm{e}(u_{B_i}, c_2)}$;
  (b) Compute $\widehat{m} = H(m)$, check whether the following equation hold

$$\prod_{i=1}^{n} (\mathrm{e}(C_4^{(i)}, g)\mathrm{e}(C_5^{(i)}, g)\mathrm{e}(C_2, m' \prod_{j=1}^{n_m} w_j))^{\Delta_{i,S}(0)} = A \tag{9}$$

If above equation holds, accept plaintext m and return *valid*; otherwise output $\perp$ as failure.

4 Security Analysis

4.1 Confidentiality

Theorem 1. *Suppose A be a attacker that makes at most q_E key extract queries, q_S signcrypt queries, q_U unsigncrypt queries, and produces a valid ciphertext with probability ϵ in time t in IND-bSC-CCA2 game, then there exist an algorithm B that can solve CDH problem with advantage ϵ'*

$$Adv(B) = \epsilon + \frac{1}{8q_S(q_E + q_S)(n+1)n_m + 1},$$

Proof. Assume that the distinguisher $\mathcal{C}$ receives a random DBDH problem instance $(g, A = g^a, B = g^b, C = g^c, Z \in G_T)$, his goal is to decide whether $Z = e(g, g)^{abc}$ or not. $\mathcal{C}$ will run A as a subroutine and act as A's challenger in the IND-bSC-CCA2 game.

Setup.

- $\mathcal{C}$ sets $l_u = 2(q_E + q_S)$ and $l_m = 2q_S$;
- Random chooses two integers k_u and k_m where $0 \le k_u \le n_u$, $0 \le k_m \le n_m$;
- Picks an integer $x' \in Z_{l_u}$, and an n_u-dimensional vector $X = (x_i)(x_i \in Z_{l_u})$;
- Picks an integer $z' \in Z_{l_m}$, and an n_m-dimensional vector $Z = (z_j)(z_j \in Z_{l_m})$;
- Picks two integers $y', w' \in Z_p$, an n_u-length vector $Y = (y_i)$ $(y_i \in Z_p)$ and an n_m length vector $W = w_j(w_j \in Z_p)$.

We define the functions for an identity u and a message m as follows,

$$F(u) = (p - l_u k_u) + x' + \sum_{i \in \mathcal{U}} x_i, \tag{10}$$

$$J(u) = j_u + \sum_{i \in \mathcal{U}} y_i, \tag{11}$$

$$K(m) = (p - l_m k_m) + x' + \sum_{j \in \mathcal{M}} z_i, \tag{12}$$

$$L(m) = j_m + \sum_{j \in \mathcal{M}} z_i \tag{13}$$

Then the challenger assigns a set of public parameters as follows:

$$g_1 = A,\ g_2 = B, \tag{14}$$

$$u' = g_2^{p-l_u k_u + x'} g^{y'}, u_i = g_2^{x_i} g^{y_i} (1 \leq i \leq n_u) \tag{15}$$

$$m' = g_2^{p-l_m k_m + z'} g^{w'}, u_j = g_2^{z_i} g^{w_i} (1 \leq j \leq n_m) \tag{16}$$

Note that these public parameters have the same distribution as in the game between the challenger $\mathcal{C}$ and the adversary A. For any identity u and any message m, we have

$$u' \prod_{i \in \mathcal{U}} u_i = g_2^{F(u)} g^{J(u)}, \quad and$$

$$m' \prod_{j \in \mathcal{M}} m_j = g_2^{K(m)} g^{L(m)}.$$

Queries

- Extract queries. When the adversary A asks for the private key corresponding to an identity u. The challenger $\mathcal{C}$ first checks if $F(u) = 0$ and aborts in this situation. Otherwise, it chooses a random $r_u \in Z_p$ and gives A the tuple such that
  $d_u = (d_{u_0}, d_{u_1})$
  $= (g_1^{-J(u)/F(u)} (u' \prod_{i \in \mathcal{U}} u_i)^{r_u}, g_1^{-F(u)^{-1}} g^{r_u})$
- Signcryption queries. At any time, A can perform a signcryption query for a plaintext M and sender u_A and receiver u_B. If $F(u_A) \neq 0\ mod\ l_u$, $\mathcal{C}$ generates a private key for u_A just calling the extract query algorithm described above, and then runs $Signcrypt(m, d_A, u_B)$ to answer A's query. Otherwise, $\mathcal{C}$ aborts.
- Unsigncryption queries. At any time, the adversary A can perform an unsigncryption query on a ciphertext σ for identities u_A and u_B. If $F(u_B) \neq 0\ mod\ l_u$, $\mathcal{C}$ first generates a private key for u_B, and then runs $Unsigncrypt$ (σ, d_B, u_A) to answer A's query. Otherwise, $\mathcal{C}$ will simply abort.

Challenge. After a polynomially bounded number of queries, A chooses a pair of identities u_A^* and u_B^* on which he wishes to be challenged. Note that $\mathcal{C}$ fails if A has asked a key extraction query on u_B^* during the first stage. Then A submits two messages $M_0, M_l \in G_T$ and u_A^*, u_B^* to $\mathcal{C}$. $\mathcal{C}$ will abort if $F(u_B^*) \neq 0\ mod\ l_u$. Otherwise, $\mathcal{C}$ flips a fair binary coin β and constructs a signcryption ciphertext of M_β as follows:

- Computes $m_\beta = H(M_\beta)$, which is an n_m-bit string.
- If $K(m_\beta) = 0$, $\mathcal{C}$ aborts. Otherwise, $\mathcal{C}$ selects a random number $r_u \in Z_p$ and sets the ciphertext as

$$\sigma^* = (ZM_\beta, C, C^{J(u_B^*)}, g_1^{-J(u_A^*)/F(u_A^*)} (g_2^{F(u_A^*)} g^{J(u_A^*)})^{r_u} C^{L(m_\beta)}, g_1^{-1/F(u_A^*)} g^{r_u})$$

At the end of the simulation, A outputs a guess β' as guess. If $\beta' = \beta$, $\mathcal{C}$ answer 1 indicating that $Z = e(g,g)^{abc}$; Otherwise, $\mathcal{C}$ answers 0 to the DBDH problem.

For the simulation to complete without aborting, we require that

- All extraction queries on an identity u are not failed, i.e., $F(u) \neq 0 \ mod \ l_u$;
- All signcryption queries (u_A, u_B, M) are not failed, i.e., $F(u_A) \neq 0 \ mod \ l_u$;
- All unsigncryption queries (σ, u_A, u_B) are not failed, i.e., $F(u_B) \neq 0 \ mod \ l_u$;
- $F(u_A^*) \neq 0 \ mod \ l_u$, $F(u_B^*) \neq 0 \ mod \ l_u$ and $K(m_\beta) = 0 \ mod \ l_m$

Let $u_l, u_2, ..., u_{q_I}$ be the identities appearing in either extract queries or in signcryption queries not involving the challenge identity. Clearly, we will have $q_I \leq q_E + q_S$. Define the events A_i, A^*, B^* as:

$$\begin{aligned} &A_i : F(u_i) \neq 0 \ mod \ l_u, \\ &A^* : F(u^*) \neq 0 \ mod \ p, \\ &B^* : K(m_\beta^*) \neq 0 \ mod \ p. \end{aligned}$$

Then the probability ϵ of $\mathcal{C}$ not aborting is

$$\begin{aligned} Pr[\neg Abort] &\geq Pr[\bigwedge_{i=1}^{q_I} A_i \bigwedge A^* \bigwedge B^*] \\ &= Pr[\bigwedge_{i=1}^{q_I} A_i] \cdot Pr[\bigwedge A^*] \cdot Pr[\bigwedge B^*] \end{aligned}$$

It easy to see that the events A_i, A^* and B^* are independent for that the functions $F(.)$ and $K(.)$ are selected independently. Assume $l_u(n_u + 1) \leq p$ which implies $0 \leq l_u k_u \leq p$. It has $F(u) = 0 \ mod \ p \Rightarrow F(u) = 0 \ mod \ l_u$. Furthermore, this assumption implies that if $F(u) = 0 \ mod \ l_u$, there will be an unique k_u with $0 \leq k_u \leq n_u$ such that $F(u) = 0 \ mod \ p$.

$$\begin{aligned} Pr[\bigwedge_{i=1}^{q_I}] &\geq 1 - \bigwedge_{i=1}^{q_I} \neg A_i = 1 - q_I/l_u \\ &\geq 1 - (q_E + q_S)/l_u \\ Pr[A^*] &= Pr[F(u^*) = 0 \ mod \ p] \\ &= Pr[F(u^*) = 0 \ mod \ l_u] Pr[F(u^*) = 0 \ mod \ p | F(u^*) = 0 \ mod \ l_u) \\ &= \frac{1}{l_u(n_u + 1)} \\ Pr[B^*] &= Pr[K(m^*) = 0 \ mod \ p] \\ &= Pr[K(m^*) = 0 \ mod \ l_m] Pr[K(m^*) = 0 \ mod \ p | K(m^*) = 0 \ mod \ l_m] \\ &= \frac{1}{l_m(n_m + 1)} \end{aligned}$$

So, the challenger $\mathcal{C}$ obtain the DBDH result advantage

$$\begin{aligned} Pr[\neg Abort] &= (1 - \frac{q_E + q_S}{l_u})\frac{1}{l_u(n_u+1)}\frac{1}{l_m(n_m+1)} \\ &= \frac{1}{8q_S(q_E+q_S)(n_u+1)(n_m+1)} \end{aligned}$$

4.2 Unforgeability

Theorem 2. *Suppose $\mathcal{F}$ be a forger that makes at most q_E key extract queries and q_S signcrypt queries and forges a valid ciphertext against bSC scheme with probability ϵ in time t in UEF-bSC-CMA2 game, then there exists an algorithm B to solve the CDH problem in Z_p.*

Proof. Assume that an forger $\mathcal{F}$ for our scheme exists, we will construct a challenger $\mathcal{C}$, who runs $\mathcal{F}$ as a subroutine, to solve an instance of CDH problem. $\mathcal{C}$ is given a group G, a generator g and elements g^a and g^b. His goal is to compute g^{ab}.

$\mathcal{C}$ first sets the public parameters using the Setup algorithm and sets $g_1 = g^a$ and $g_2 = g^b$. It defines the functions for an identity u and a message m as follows,

$$F(u) = (p - l_u k_u) + x' + \sum_{i \in \mathcal{U}} x_i, \tag{17}$$

$$J(u) = j_u + \sum_{i \in \mathcal{U}} y_i, \tag{18}$$

$$K(m) = (p - l_m k_m) + x' + \sum_{j \in \mathcal{M}} z_i, \tag{19}$$

$$L(m) = j_m + \sum_{j \in \mathcal{M}} z_i \tag{20}$$

.

It sets the public parameters as

$$u' = g_2^{p - l_u k_u + x'} g^{y'}, u_i = g_2^{x_i} g^{y_i} (1 \leq i \leq n_u) \tag{21}$$

$$m' = g_2^{p - l_m k_m + z'} g^{w'}, u_j = g_2^{z_i} g^{w_i} (1 \leq j \leq n_m) \tag{22}$$

For any identity u and message m, it has

$$u' \prod_{i \in \mathcal{U}} u_i = g_2^{F(u)} g^{J(u)}, \tag{23}$$

$$m' \prod_{j \in \mathcal{M}} m_j = g_2^{K(m)} g^{L(m)} \tag{24}$$

Let $\mathcal{F}$ generates a valid ciphertext σ^* on message m^* where m^* has never been queried and sender identity u_A^* never been key extracted queried before.

If $F(u_A^*) \neq 0 \; mod \; p$ and $K(m^*) \neq 0 \; mod \; p$ then $\mathcal{C}$ aborts. Otherwise,

$$F(u_A^*) = 0 \; mod \; p, and \tag{25}$$

$$K(m^*) = 0 \; mod \; p \tag{26}$$

$\mathcal{C}$ computes and outputs

$$\begin{aligned} &\frac{\sigma_4^*}{\sigma_5^{J(u_A^*)}\sigma_2^{L(m^*)}} \\ &= \frac{g_2^{\alpha}(u'\prod_{i\in\mathcal{U}} u_i)^{r_A}(m'\prod_{j\in\mathcal{M}} m_j)^{r_m}}{g^{J(u_A^*)r_A}\cdot g^{L(m^*)r_m}} \\ &= g_2^{\alpha} = g^{ab}. \end{aligned}$$

5 Conclusion

In this paper, we proposed a fuzzy biometric based signcryption scheme, which The security model, including indistinguishable against adaptively chosen ciphertext attack and existentially unforgeable against adaptively chosen message attack, are analyzed and proved.

Acknowledgments

The authors would grateful thank the anonymous reviewers for their helpful and valuable comments.This work is supported by the National Natural Science Foundation of China under Grants 60773175 and 60973134, the Natural Science Foundation of Guangdong Province under Grants 9151064201000058, the Foundation of National Laboratory for Modern Communications under Grant 9140C1108020906, and the Support of JSPS PostDoctoral Research Project of Japan.

References

1. Boneh, D., Boyen, X.: Short signatures without random oracles. In: Cachin, C., Camenisch, J.L. (eds.) EUROCRYPT 2004. LNCS, vol. 3027, pp. 56–73. Springer, Heidelberg (2004)
2. Boneh, D., Boyen, X.: Short Signatures Without Random Oracles and the SDH Assumption in Bilinear Groups. Journal of Cryptology 21, 149–177 (2008)
3. Boneh, D., Boyen, X.: Efficient selective-id secure identity based encryption without random oracles. In: Cachin, C., Camenisch, J.L. (eds.) EUROCRYPT 2004. LNCS, vol. 3027, pp. 223–238. Springer, Heidelberg (2004)
4. Boneh, D., Boyen, X.: Secure identity based encryption without random oracles. In: Franklin, M. (ed.) CRYPTO 2004. LNCS, vol. 3152, pp. 443–459. Springer, Heidelberg (2004)
5. Waters, B.: Efficient identity-based encryption without random oracles. In: Cramer, R. (ed.) EUROCRYPT 2005. LNCS, vol. 3494, pp. 114–127. Springer, Heidelberg (2005)
6. Gentry, C.: Practical identity-based encryption without random oracles. In: Vaudenay, S. (ed.) EUROCRYPT 2006. LNCS, vol. 4004, pp. 445–464. Springer, Heidelberg (2006)

7. Tan, C.H.: A new signature scheme without random oracles. International Journal of Security and Networks 1(3-4), 237–242 (2006)
8. Ren, Y., Gu, D.: Fully CCA2 secure identity based broadcast encryption without random oracles. Information Processing Letters 109, 527–533 (2009)
9. Yu, Y., Yang, B., Sun, Y., Zhu, S.: Identity based signcryption scheme without random oracles. Computer Standards & Interfaces 31(1), 56–62 (2009)
10. Liu, Z., Hu, Y., Zhang, X., Ma, H.: Certificateless signcryption scheme in the standard model. Information Sciences 180(3), 452–464 (2010)
11. Shamir, A.: Identity-based cryptosystems and signature schemes. In: Blakely, G.R., Chaum, D. (eds.) CRYPTO 1984. LNCS, vol. 196, pp. 47–53. Springer, Heidelberg (1985)
12. Sahai, A., Waters, B.: Fuzzy identity-based encryption. In: Cramer, R. (ed.) EUROCRYPT 2005. LNCS, vol. 3494, pp. 457–473. Springer, Heidelberg (2005)
13. Yang, P., Cao, Z., Dong, X.: Fuzzy identity based signature, http://eprint.iacr.org/2008/002.pdf
14. Beak, J., Susilo, W., Zhou, J.: New construction of fuzzy identity-based encryption. In: ASIACCS'07, pp. 368–370. ACM, New York (2007)

Key Independent Decryption of Graphically Encrypted Images

Ram Ratan

Defence Research and Development Organisation
Scientific Analysis Group, Metcalfe House Complex, Delhi-110054
India
ramratan_sag@hotmail.com

Abstract. Cryptographically, an encryption algorithm should be strong enough so that one could not extract any information from encrypted data. A graphical encryption method proposed in [1] for the security of computer data is cryptanalysed in this paper. There are some regions left unchanged and clearly visible in graphically encrypted images. Key independent decryption of graphically encrypted images is proposed for recovery of intelligible information. Decryption scheme is based on neighbourhood similarity characteristics of adjacent pixels. Simulation results show that the decrypted images obtained by the proposed scheme are quite intelligible to understand. The graphical encryption method in present form is not suitable for security applications as encrypted images can be decrypted easily.

Keywords: Image Secrecy, Graphical Encryption, Image Decryption, Neighbourhood Similarity, Visual Perception.

1 Introduction

Security of information is an important issue in design and development of secure information management systems for reliable communication and storage of sensitive information. Nowadays in the digital world, digital images apart from other form of information are commonly used by modern societies. There are many ways: spread spectrum, steganography and cryptography for achieving security of such information. Encryption is the process which transforms the plain image with the use of encryption key into encrypted form which is unintelligible and looks like random mesh of pixels. Image encryption has wide applications in various areas like strategic communication, telemedicine, medical imaging and multimedia systems for secure management of visual information. The images are different from the normal text and it is not a wise idea to use traditional encryption schemes to encrypt them because of much encryption time of large image size. Moreover, decrypted text must be same as plain text but this is not necessary for images because of visual characteristics of human perception which tolerate small errors in decrypted images.

H. Chen et al. (Eds.): PAISI 2010, LNCS 6122, pp. 88–97, 2010.
© Springer-Verlag Berlin Heidelberg 2010

Various encryption schemes have been reported in the literature for achieving the security of images which can be classified in three types as position permutation, value transformation and visual transformation [1-5]. Present paper is concerned with value transformation based encryption in which the pixel values are transformed by drawing the lines randomly in inversion mode on image plane, i.e., black pixel becomes white or vice versa. This method of encryption is known as graphical encryption method for encryption of computer data [1].

Cryptanalysis of such encryption methods for recovery of plain information from encrypted images is very important in various applications like interception analysis for extraction of meaningful information and also in assessing security of encryption schemes. Decryption is the process by which plain information is recovered from given encrypted information with the use of decryption key. Decryption is very difficult in the absence of decryption key to recover plain information. Cryptanalytically, decryption should be simple and fast so that one can recover meaningful and intelligible information from given encrypted data with minimum efforts. Exploitation of weaknesses observed in encryption method, reduction of number of trails in exhaustive method of decryption and recovery of any plain information even in partial or in distorted form are also the appreciable achievements. Depending on the situations, following attacks can be considered in the cryptanalysis of encryption methods: (1) Known cipher image (2) Chosen cipher image (3) Known plain image and (4) Chosen plain image. Cryptographically, the attacks should not be applicable in any case on the encryption method developed for security applications.

A graphical method of encryption was cryptanalysed in [6] for obtaining original data from given ciphertext under known and chosen plaintext conditions. The mask key data is obtained from a pair of plaintext (chosen/known) and ciphertext by $'XOR'$ operation and this mask key data is $'XOR'$ with given ciphertext to obtain original data. This attack is applicable only when ciphertext of chosen/known plaintext and given ciphertext which is to be decrypted are obtained with same encryption key. Graphical encryption method in which horizontal/vertical lines along with few random lines were drawn was also analysed for recovery of text documents by computing line to line correlation and applying Fuzzy character recognition under known cipher image situation [7].

In this paper, we present some observations: presence of unchanged regions in encrypted images, inversion of pixels several times etc. and propose the key independent decryption scheme for recovery of graphically encrypted gray level images when only an encrypted image is known. The point, local and global operations like image smoothing, edge enhancement etc. for enhancement of images and also for extraction of features like edge extraction, region segmentation etc. for analysis of images which are useful in many image processing applications are available in the literature [8-10]. These operations are not applicable directly in decryption of encrypted images because of random messing and high distortion in such images. For decryption of such graphically encrypted images, we require to invert the pixels again along random lines as drawn during encryption. For this, we need seed of random number generator used during encryption to get same line points for drawing same lines and inverting pixels along drawn lines. In this paper, we decrypt such encrypted images without any knowledge of decryption key. Two cases, first one random lines drawing and another horizontal/vertical lines drawing used during encryption are considered for decryption. Decryption scheme proposed is based on neighbourhood similarity characteristics of adjacent pixels where point-to-point operation for random lines and line-to-line operation for horizontal/vertical lines are performed. Decryption scheme proposed for decryption of images encrypted with random lines drawing is also applicable for any kind of pixel inversion based encryption to decrypt such encrypted images.

The paper is organized as follows: First we give a brief introduction of graphical encryption method in Section 2. The cryptographic observations on graphical image encryption are presented in Section 3. Decryption scheme proposed for recovery of intelligible information from given graphically encrypted image is presented in Section 4. The simulation results obtained for some encrypted images are presented in Section 5. Finally, the paper is concluded in Section 6 followed by the references.

2 Graphical Encryption

Computer data which is to be encrypted is displayed on computer monitor. Graphical encryption method uses two computer functions: (1) random number generation, and (2) drawing of line in inversion mode with a pen of size $1 \times 1, 2 \times 2$ etc. between two points on the image plane. As per inversion, the black pixel becomes white and vice versa during graphical encryption. The seed of random number generator is known as the key which is needed during encryption and decryption. Random number generator generates a sequence of random numbers as coordinates of end points of lines which are to be drawn on image plane. As the number of lines increases during encryption, the intelligibility of encrypted image reduces. The process of drawing line in inversion mode is repeated until the intelligibility of image vahishes. Detailed description of this method can be found in [1]. This method is applied for encryption of gray level images where the darker pixels become brighter and vice versa [11]. Decryption of encrypted images is performed with the same key, pen size and number of lines as used in encryption to obtain the original images. As an example, encrypted images obtained for some plain images are shown in *figure 1.*

3 Observations on Graphical Encryption

Graphical encryption method has several good characteristics for encryption: (1) easy to implement, (2) easy assignment of encryption key, (3) non propagation of error, and (4) non expansion of encrypted data. Although graphical encryption method has above advantages but following observations show some disadvantages: (1) pixels are inverted many times depending on number of lines passing through these pixels, (2) unintelligibility of encrypted image depends on number of lines drawn, (3) Nearly half of the pixels in encrypted image remain

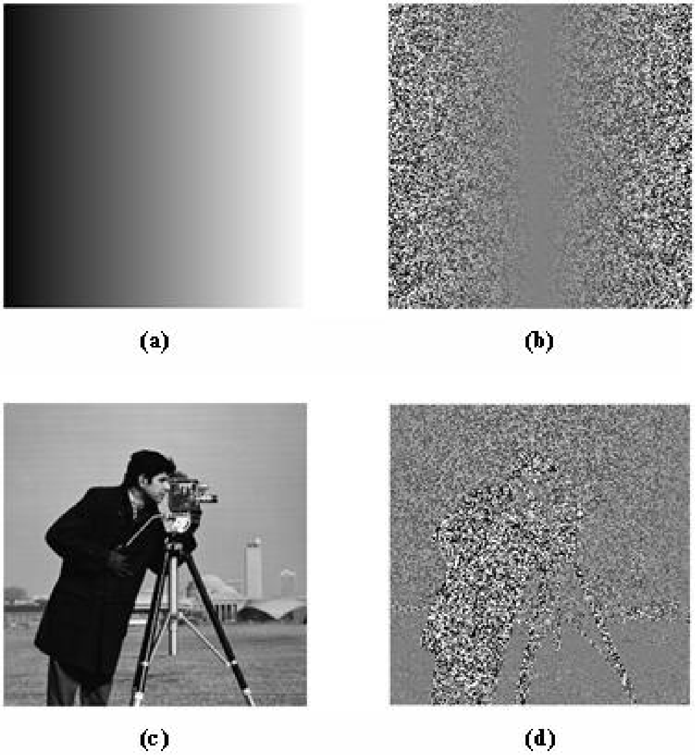


Fig. 1. Graphically encrypted images showing presence of unchanged regions : (a),(c) plain images; and (b),(d) encrypted images

unchanged, (4) Regions having pixel values near to middle of range in plain image remain visually unchanged and visible in encrypted image.

Let f be an image of size $M \times N$, $f(x, y), 1 \leq x \leq M, 1 \leq y \leq N$, be the pixel value at (x, y) position. $f(x, y)$ lies between 0 to L where L is the range of pixel values and this is 255 for 8 bit gray level image. Let encrypted image is f' which is obtained by applying graphical encryption on f. The regions of plain image which have $f(x, y)$ near to $L/2$ remain unchanged in f'. The *figure 1* shows unchanged regions in f. An image of *figure 1(a)* is the simulated picture where pixel values vary from 0 to 255 starting from left to right. And an image of *figure 1(c)* is the actual picture. In *figure 1*, it is seen that the vertically middle portion of *figure 1(a)* appears unchanged as shown in *figure 1(b)* and some patches of *figure 1(c)* appear unchanged as shown in *figure 1(d)*.

Similar results as of graphical encryption method can be obtained easily by generating a random binary matrix b of same size as of f through random number generator and inverting randomly selected pixel $f(x, y), 1 \leq x \leq M, 1 \leq y \leq N$, once only depending on $b(x, y)$ of b. $f(x, y)$ is to be inverted if $b(x, y) = 1(0)$ otherwise remains unchanged, i.e., if $b(x, y) = 1(0)$ then $f'(x, y) = L - f(x, y)$

else $f'(x,y) = f(x,y)$. In this manner, the pixels of f chosen randomly are to be inverted once only to get f'.

4 Image Decryption Scheme

We see in images that the value of pixels is normally varying smoothly in neighbourhood regions. This property of plain images can help us in decryption to recover intelligible information from given encrypted images. Divide and conquer attack can make the solution efficient to given complex problem by decomposing it into simple problems. Normally, we require decryption key (seed, pen size and number of lines as used in encryption) to decrypt such encrypted images, but decryption scheme proposed is independent of key. We take two cases, one is random lines drawing and another is horizontal/vertical lines drawing where key information is not available for decryption of graphically encrypted images.

As per divide and conquer attack, we process pixel-by-pixel and line-by-line for random lines drawing and horizontal/vertical lines drawing respectively. The neighbourhood similarity applied here is the difference between two adjacent pixels (lines) which is used to correct pixels (lines) by inverting function to decrypt encrypted image. Let f' is the given encrypted image and g is the decrypted image. The decryption process for recovery of information from graphically encrypted image with random and horizontal/vertical lines drawing is described as follows:

4.1 Decryption of Image Encrypted with Random Lines Drawing

```
//Processing of pixels starting from second pixel
 of column one//
  for x = 2 to M
  begin
     diff1 = |f'(x,1) - f'(x-1,1)|
     diff2 = |(L-f'(x,1)) - f'(x-1,1)|
     if (diff1 > diff2) then g(x,1) = L - f'(x,1)
     else g(x,y) = f'(x,y)
  end
//Processing of pixels starting from first row and
second column//
  for x = 1 to M
  for y = 2 to N
  begin
     diff1 = |f'(x,y)-f'(x,y-1)|
     diff2 = |(L-f'(x,y))-f'(x,y-1)|
     if (diff1 > diff2) then g(x,y) = L-f'(x,y)
     else g(x,y) = f'(x,y)
  end
```

4.2 Decryption of Image Encrypted with Horizontal/Vertical Lines Drawing

```
//Processing of rows starting from second row//
  for x = 2 to M
  begin
     diff1 = |row(x) - row(x-1)|
     //(pixel by pixel sum of absolute difference of
      pixel values)//
     diff2 = |L-row(x) - row(x-1)|
     if (diff1 > diff2) then row(x) = L-row(x)
  end
//Processing of columns starting from second column//
  for y = 2 to N
  begin
     diff1 = |column(y) - column(y-1)|
     diff2 = |L-row(y) - row(y-1)|
     if (diff1 > diff2) then column(y) = L-column(y)
  end
```

We start processing of encrypted image from second pixel $f'(1,2)$ based on previous pixel $f'(1,1)$ in case of random lines drawing and from second row based on previous row in case of horizontal/vertical lines drawing. It is not known here whether the first pixel (first row) is inverted or not during encryption. We get decrypted image g as a negative of plain image f if first pixel (first row) was inverted and we get g as a plain image if first pixel (first row) was not inverted during encryption. So we obtain two decrypted images, both have intelligible information, in which one of them has good visual perception similar to plain image. Decryption process discussed for random lines drawing can also be considered for decryption of images encrypted with any kind of drawing and pixel inversion.

5 Simulation Results and Discussions

Proposed decryption scheme for graphically encrypted images is implemented in MATLAB programming on MATLAB Platform. Scheme provides intelligible decrypted images with good visual perception. The decrypted images are same to the originals in case of encryption with horizontal/vertical lines drawing and quite intelligible also in case of encryption with random lines drawing. Proposed decryption scheme is tested for decryption of various encrypted images. As an example, decrypted images obtained for some images encrypted with horizontal/vertical lines drawing and random lines drawing are shown in *figure 2* and *figure 3* respectively. Decrypted results shown in *figure 3* are obtained for plain images (a)-(d) of *figure 2*

The error in decrypted image is measured as mean square error (MSE) which is computed as

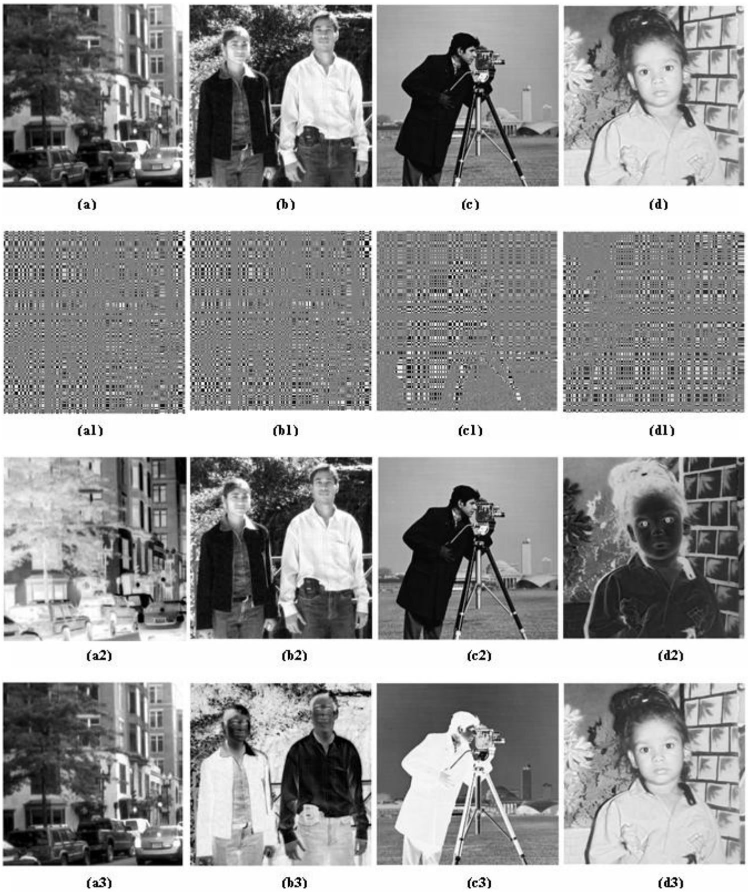


Fig. 2. Decryption of images graphically encrypted with horizontal/vertical lines: (a)-(d) plain images; (a1)-(d1) encrypted images of (a)-(d); (a2)-(d2) decrypted images; and (a3)-(d3) images as negative of (a2)-(d2)

$$MSE = \tfrac{1}{M \times N}[f(i,j) - g(i,j)]^2, 1 \leq i \leq M \text{ and } 1 \leq j \leq N.$$

Error measured as MSE in decrypted images does not reduce considerably from the error in encrypted images in case of random lines drawing encryption but it reduces to zero in case of horizontal/vertical lines drawing encryption. MSE measured in different encrypted and decrypted images is shown in *Table 1*. From *figure 2* and *Table 1*, it is clear that in case of horizontal/vertical lines drawing encryption the decrypted images which look same as to original have zero MSE and others have larger MSE even compared to MSE in encrypted images. From *figure 3* and *Table 1*, it is also clear that in case of random lines drawing encryption the decrypted images which have larger MSE even compared to MSE in encrypted images look fine whereas others have lesser MSE. Decryption scheme

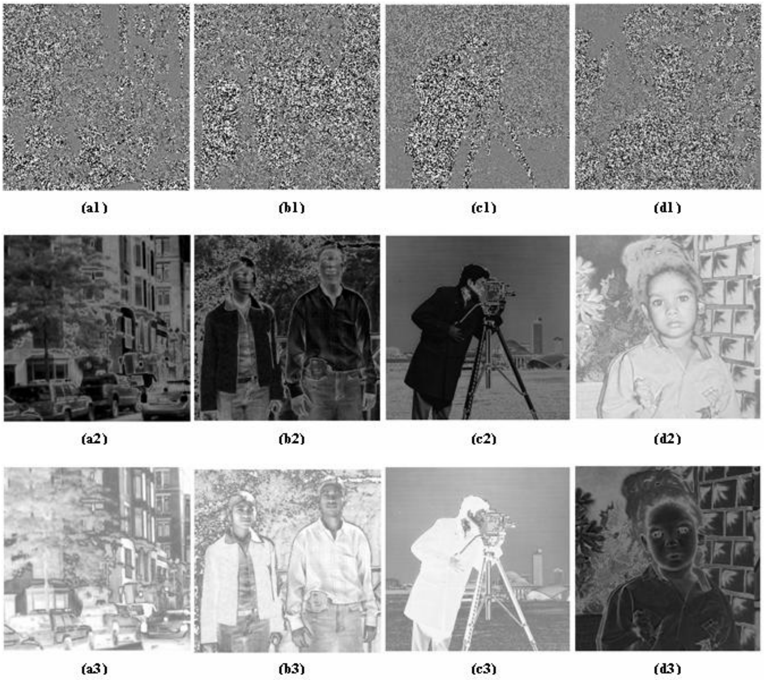


Fig. 3. Decryption of images graphically encrypted with random lines: (a1)-(d1) encrypted images of figure 2(a)-(d); (a2)-(d2) decrypted images; (a3)-(d3) images as negative of (a2)-(d2)

proposed for decryption of random lines drawing based encrypted images is also applied for decryption of horizontal/vertical lines drawing based encrypted images. The results are shown in *figure 4* where visual quality of such decrypted images is also quite intelligible. These results show that the decryption scheme proposed for decryption of encrypted images with random lines drawing is applicable also for any kind of pixel inversion based encryption to decrypt such encrypted images.

Security of graphical encryption method depends on the seed of random number generator which is used to obtain end points of lines for inverting pixels along lines drawn during encryption and decryption. In absence of key and fixed lines/pensize, we have to apply 2^{32} number of trials in exhaustive method for 32 bit seed to get the plain image. As the decryption scheme proposed is independent of key and does not require key information, the scheme decrypts encrypted image in one pass only for random lines drawing encryption case and in two passes only for horizontal/vertical lines drawing encryption case. Simulated results show that the graphical encryption method is insecure in present form. The graphical encryption can be made secure against above attacks by incorporating pixel masking, pixel substitution and pixel permutation [12-15].

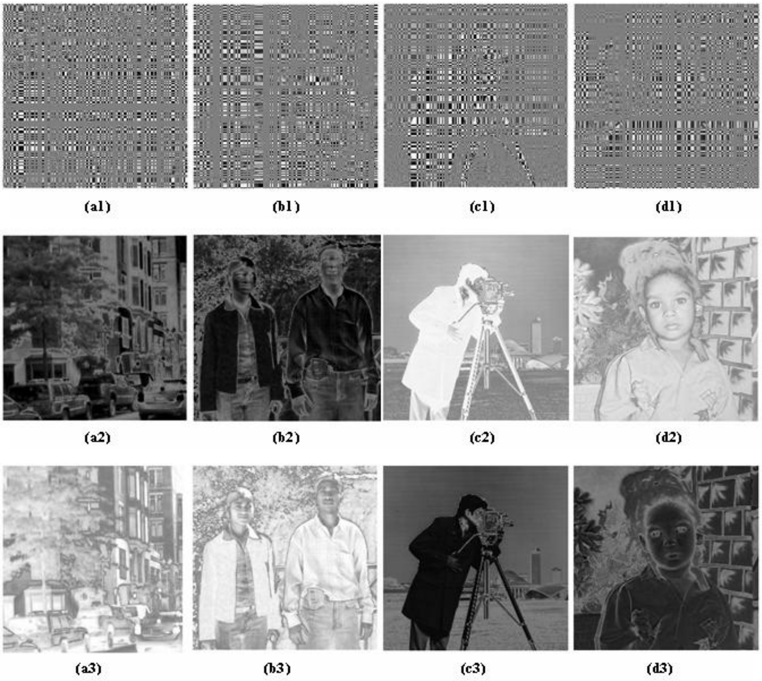


Fig. 4. Decryption of images encrypted with horizontal/vertical lines by scheme proposed for decryption of images encrypted with random lines: (a1)-(d1) images of figure 2(a)-(d) encrypted with horizontal/vertical lines; (a2)-(d2) decrypted images; and (a3)-(d3) images as negative of (a2)-(d2)

Table 1. MSE measured in different images

Plain images for simulation	MSE in case of horizontal/vertical lines drawing			MSE in case of random lines drawing		
	Encrypted images	First decrypted image	Second decrypted image	Encrypted image	First decrypted image	Second decrypted image
Street	9651	19260	0	9613	14849	4411
Children	13897	0	27866	13984	17007	10853
Cameraman	8687	0	17284	8611	13215	4069
baby	8972	17956	0	9014	1981	15975

6 Conclusion

The decryption scheme presented in this paper for decryption of graphically encrypted images is automatic and independent of key. The scheme of decryption is based on divide and conquer attack in which neighbourhood similarity between adjacent pixels or lines has been used. It has been shown in simulation results that encrypted images can be decrypted easily without the knowledge of decryption key with good intelligibility. The images encrypted with horizontal/vertical lines drawing can be decrypted as original and the images encrypted with random lines drawing can also be decrypted with good visual perception. Further, it has been shown that decryption scheme proposed for decryption of images encrypted with random lines drawing is applicable for any kind of pixel inversion based encryption to decrypt such encrypted images with good perception. Hence, graphical encryption method in present form is insecure and should not be used for security applications.

References

1. Schwartz, C.: A new graphical method for encryption of computer data. Journal of Cryptologia 15(1), 43–46 (1991)
2. Bourbakis, N.G., Alexopoulos, C.: Picture data encryption using scan patterns. Pattern Recognition 25(6), 567–581 (1882)
3. Maitra, A., Rao, Y. V. S., and Prasanna, S. R. M.: A new image encryption approach using combinational permutation techniques. International Journal of Computer Science, 19(2), 127-131, (2006)
4. Maniccam, S.S., Bourbakis, N.G.: Image and video encryption using scan patterns. Pattern Recognition 37(4), 725–737 (2004)
5. Yen, J.-C., Guo, J.-I.: A new image encryption algorithm and its VLSI architecture. In: Proceedings of IEEE Workshop Signal Processing Systems, pp. 430–437 (1999)
6. Chin, Y.-C., Wang, P.-C., Hwang, J.-J.: Cryptanalysis on Schwartz graphical encryption method. Journal of Cryptologia 17(3), 301–304 (1993)
7. Ratan, R., Saxena, P.K.: An algorithm for the restoration of distorted text documents. In: Proceedings of Intl. Conference on Computational Linguistics, Speech and Document Processing (ICCLSDP'98), pp. A38–A43 (1998)
8. Jain, A.K.: Fundamentals of digital image processing. Prentice Hall, Englewood Cliffs (1995)
9. Russ, J.C.: The image processing handbook. CRC Press, Boca Raton (1995)
10. Young, T.Y., Fu, K.S.: Handbook of pattern recognition and image processing. Academic Press, London (1986)
11. Ratan, R., Saxena, P.K.: Image processing based techniques for securing text documents. Journal of Discrete Mathematical Sciences and Cryptography 3(1-3), 113–129 (2000)
12. Fu, C., Zhu, Z.: A chaotic encryption scheme based on circular bit shift method. In: Proceedings of Intl. Conference for Young Computer Scientists, pp. 3057–3061. IEEE Computer Society, Los Alamitos (2008)
13. Menezes, A.P., Van Oorschot, Vanstone, S.: Handbook of applied cryptography. CRC Press, Boca Raton (1996)
14. Schneier, B.: Applied cryptography. John Wiley & Sons Inc., Chichester (1996)
15. Yen, J.C., Guo, J.I.: A new chaotic key based design for image encryption and decryption. In: Proceedings of IEEE International Symposium Circuits and Systems, vol. 4, pp. 49–52 (2000)

Towards Confidentiality of ID-Based Signcryption Schemes under without Random Oracle Model

Mingwu Zhang[1,3], Pengcheng Li[1], Bo Yang[1],
Hao Wang[2], and Tsuyoshi Takagi[3]

[1] Department of Computer Science and Engineering, College of Informatics,
South China Agricultural University, Guangzhou, 510642, China
csmwzhang@gmail.com

[2] School of Electronic and Information Engineering,
South China University of Technology, Guangzhou, 510641, China

[3] Faculty of Mathematics, Kyushu University, Fukuoka, 819-0395, Japan

Abstract. Signcryption, achieves confidentiality and authenticity simultaneously in an efficient manner, is a novel cryptographic primitive in a single logic step with more efficient computation and communication cost to support the security. In this paper, we analyze two identity based signcryption schemes under the without random oracles model including a signcryption scheme by Yu et al. [11] and a certificateless signcryption scheme by Liu et al. [12]. We give the chosen plaintext attacks to prove that the proposed previous schemes cannot resist on the indistinguishability on chosen ciphertext attack to support semantic secure for confidentiality.

Keywords: Confidentiality, signcryption, without random oracles, chosen plaintext attacks.

1 Introduction

Identity based cryptosystems were introduced by Shamir [1] where users uses their identities information such as name, identity number, IP address or email address as his public key, which is a good alternative for certificate-based systems from the viewpoint of efficiency and convenience. Provable security is the basic model for identity based cryptographic schemes. Most schemes [2,3] were proven secure in the random oracle model. Although the model under the random oracles is efficient and useful, it has been shown that when random oracles are instantiated with concrete hash functions, the resulting scheme may not be secure [4] where they deploy a weaker model of security for identity based encryption that they term the Selective-ID model. In the Selective-ID model, the adversary must first declare which identity it wishes to be challenged on before the global parameters are generated. Boneh and Boyen [5] improved upon this result by describing an efficient scheme that is secure in the Selective-ID model, and then they described a scheme that is fully secure without random

H. Chen et al. (Eds.): PAISI 2010, LNCS 6122, pp. 98–104, 2010.
© Springer-Verlag Berlin Heidelberg 2010

oracles [6]. However, their construction is too inefficient to be of practical use. In 2005, Waters [7] constructed an efficient and practical ID-based encryption scheme without random oracle.

Signcryption [3,10,8,9] can perform digital signature and encryption simultaneously at lower computational costs and communication overheads than sign-then-encrypt way to obtain private and authenticated communications in the open channel. Based on Waters's primary encryption model [7], Yu et al. [11] proposed the first identity based signcryption scheme without random oracles, and Liu et al. [12] proposed a certificateless ID-based signcryption scheme.

In this paper, we show that the Yu et al.'s ID-based signcryption scheme [11] is not secure under chosen plaintext attack, and also be the Liu et al.'s certificateless signcryption scheme [12]. We give the attack methods to obtain the total advantage in indistinguishability game IND-CCA2.

The remainder of the paper is organized as follows: The formal model of ID-based signcryption scheme is described in section 2; Two signcryption scheme without random oracles are given and analyzed in section 3 and 4 respectively, and conclusion is drawn in section 5.

2 Model of an ID-Based Signcryption Scheme

2.1 Formal of ID-Based Signcryption Model

-GC: Taken security parameter 1^k as input, this algorithm generates a master key MSK and the system's public parameters $params$, $(params, MSK) \leftarrow Setup(1^k)$, which also include a description of a finite message space $\mathcal{M}$ together with a description of a ciphtertext space $\mathcal{C}$.

-KG: Given an identity ID, the PKG runs this algorithm to generate the private key SID associated with ID and transmits it to ID via a secure channel.

-SC: To send a message m to Bob whose identity is ID_B, Alice with identity ID_A obtains a ciphertext σ by computing SC(m, S_{ID_A}, ID_B).

-US: After Bob receives the ciphertext σ, he computes US(σ, S_{ID_B}, ID_A) and obtains the message m or the symbol $\perp$ indicating that the ciphertext is invalid.

2.2 Security Definition

We say that a signcryption scheme is secure in the sense of indistinguishability(abbreviated by "IND"), there is no polynomial-time adversary that can learn any information about the plaintext from the signcrypted text except for its length.

Definition 1. IND-SC-CCA2*: Let SCR = (GC,KG, SC,US) be an identity based signcryption scheme. Let A^{CCA} be an attack algorithm against the indistinguishability of the scheme SCR. Consider the IND-SC-CCA2 game(SCGame) as follows:*

```
$cp \leftarrow GC(k)$
$(d_A, Q_{IDA}) \leftarrow_R \ KG(ID_A)$
$(d_B, Q_{IDB}) \leftarrow_R \ KG(ID_B)$
$(m_0, m_1) \leftarrow A^{CCA}(k, cp, find, Q_{IDA}, Q_{IDB} | SC(cp, d_A), US(cp, d_B, \cdot, \cdot))$
$\beta \leftarrow_R \{0,1\}$; $C^* \leftarrow SC(cp, d_A, Q_{IDB}, m_\beta)$
$\beta' \leftarrow A^{CCA}(k, cp, guess, Q_{IDA}, Q_{IDB}, C^* | SC(cp, d_A, \cdot, \cdot), US(cp, d_B, \cdot, \cdot))$
If $\beta' = \beta$, and $(Q_{IDA}, C^*)$ was never queried to $US(cp, d_B, \cdot, \cdot)$
Return $\top$ as succeed Else return $\bot$ as failure.
```

Note that the attacker's queries is adaptively. In other words, A^{CCA} is allowed to query (Q_{IDS}, C^*) to the unsigncryption oracle US(cp, d_B, ·,·) where unsigncryption is performed under the public key Q_{IDS} of identity ID_S which is arbitrarily chosen by A^{CCA} and is different from Q_{IDA}.

The advantage of an attacker A^{CCA} by the probability is defined as

$$Adv_{A^{CCA},SCR}^{IND-SC-CCA2}(k) = |2Pr[SCGame(k, A^{CCA}, SCR) = 1] - 1| \tag{1}$$

Definition 2. EUF-SC-CMA2*: Let SCR = (GC,KG,SC,US) be an identity based signcryption scheme. Let A^{EUF} be an attack attacker against the existing unforgeability of the scheme SCR. Consider the EUF-SC-CMA2 game(UFGame) as follows:*

```
$cp \leftarrow GC(k)$
$(d_{IDA}, Q_{IDA}) \leftarrow_R \ KG(ID_A)$
$(C^*, Q_{IDB}) \leftarrow A^{EUF}(k, cp, Q_{IDA} | SC(cp, d_A))$
Find some $d_B$ such that $(d_B, Q_{IDB}) \in \{KG(k, cp)\}$
If such $d_B$ does not exist, Return $\bot$ as failure
$m \leftarrow US(cp, d_B, Q_{IDA}, C^*)$
If $m \neq Reject$ and $(Q_{IDB}, m)$ has not been queried by $A^{EUF}$ to $SC(cp, d_A, \cdot, \cdot)$
Return $\top$ as succeed Else return $\bot$ as failure.
```

The advantage of a forger A^{EUF} can break the EUF-SC-CMA2 game is defined as

$$Adv_{A^{EUF},SCR}^{EUF-SC-CMA2} = Pr[UFGame(k, cp, SCR, A^{EUF}) - 1] \tag{2}$$

3 The Yu et al.'s Signcryption Scheme

3.1 The Yu et al.'s Scheme [11]

1. GC

1. Let (G, G_T) be bilinear groups of prime order p, $G =< g >$ and e be admissible bilinear map from $G \times G$ to G_T. H is a hash function $H : G_T \rightarrow \{0,1\}^{n_m}$, n_u and n_m are the bit strings length of identities and messages;

2. Pick a secret value $\alpha \in Z_p$, compute $g_1 = g^{\alpha}$;
3. Randomly pick $g_2, u', m' \in G$, and pick $\overrightarrow{\mathbf{u}} = (u_i) \in G^{n_u}, \overrightarrow{\mathbf{m}} = (m_i) \in G^{n_m}$;
4. Keep the master secret key $msk = g_2^{\alpha}$ and public parameter $params = (G, G_T, \mathrm{e}, g_1, g_2, u', \overrightarrow{\mathbf{u}}, m', \overrightarrow{\mathbf{m}})$.

2. KG Let $u \in \{0,1\}^{n_u}$ be an identity string of length n_u, and $u[i]$ be the i-th bit of u. Define $\mathcal{U} \subset \{1, 2, ..., n_u\}$ to be the set of indices i such that $u[i] = 1$. A user's private d_u with identity u is generated as follows.

1. Randomly pick $r \in Z_p$;
2. Compute $d_u = (d_{u1}, d_{u2}) \in G^2$ such that $d_{u1} = msk \cdot (u' \prod_{i \in \mathcal{U}} u_i)^{r_u}, \quad d_{u2} = g^{r_u}$

3. SC To send a message $M \in G_T$ to B which identity is u_B, sender A with his secret key $d_A = (d_{A1}, d_{A2})$ does

1. Randomly pick $r_m \in Z_p$;
2. Compute $\sigma_1 = \mathrm{e}(g_1, g_2)^{r_m} M \in G_T$;
3. Compute $\sigma_2 = g^{r_m}$;
4. Compute $\sigma_3 = (u' \prod_{i \in \mathcal{U}_B} u_i)^{r_m}$;
5. Compute $M' = H(M)$ where M' is an n_m-bit string and $M'[i]$ denotes the ith bit of M'. $\mathcal{M} \subset \{1, 2, ..., n_m\}$ denotes the set of i for which $M'[i]$=1. Compute $\sigma_4 = d_{A1}(m' \prod_{j \in \mathcal{M}} m_j)^{r_m}$;
6. $\sigma_5 = d_{A2}$. The ciphertext produced by A is $\sigma = (\sigma_1, \sigma_2, \sigma_3, \sigma_4, \sigma_5) \in G^4 \times G_T$.

4. US After received a ciphertext $\sigma = (\sigma_1, \sigma_2, \sigma_3, \sigma_4, \sigma_5)$, B decrypts it as follows.

1. Compute $M = \sigma_1 \frac{\mathrm{e}(\sigma_3, d_{B2})}{\mathrm{e}(\sigma_2, d_{B1})}$;
2. Compute $M' = H(M)$, and generate the corresponding set $\mathcal{M}$ for that $M'[i]$=1;
3. Accept the message M if the following equation holds:

$$\mathrm{e}(\sigma_4, g) = \mathrm{e}(g_1, g_2)\mathrm{e}(u' \prod_{i \in \mathcal{U}_A} u_i, \sigma_5)\mathrm{e}(m' \prod_{j \in \mathcal{M}} m_j, \sigma_2) \tag{3}$$

3.2 Confidential Attack by Chosen Plaintext Attack

The basic conception of confidentiality is IND-SC-CCA2 such that the attacker cannot distinguish the generated ciphertext from his submitted two messages.

In this section, we give a chosen plaintext attack to describe the Yu et al.'s scheme cannot provide IND-SC-CCA2 security.

When attacker A receives the challenged ciphertext $\sigma^* = (\sigma_1^*, \sigma_2^*, \sigma_3^*, \sigma_4^*, \sigma_5^*)$ from sender u_A that the ciphertext $\sigma*$ is a random chosen plaintext $M \in \{M_0, M_1\}$. A performs the following steps:

- First make a "wild guess" of β to be 0;
- Compute $M' = H(M_\beta)$, and generate the corresponding bits set $\mathcal{M}_\beta$ for that $M'_\beta[i]$=1;
- Compute the hash value of sender identity string u_A: $H(u_A) = u' \prod_{i \in \mathcal{U}_A} u_i$, such that $\mathcal{U}_A$ is the set satisfying $u_A[i] = 1$ $(1 \leq i \leq n_u)$;
- Check the equation $\mathrm{e}(\sigma_4^*, g) = \mathrm{e}(g_1, g_2)\mathrm{e}(H(u_A), \sigma_5^*)\mathrm{e}(m' \prod_{j \in \mathcal{M}_\beta} m_j, \sigma_2^*)$

If above equation holds, then attacker A knows that M_0 is the plaintext for the challenged ciphertext. Otherwise, A knows that M_1 is the plaintext for the challenged ciphertext. Consequently, attacker A has total advantage win the IND-SC-CCA2 game.

4 Liu et al. Certificateless Signcryption Scheme

4.1 The Liu et al.'s Scheme [12]

GC(Generator Center). Let(G, G_T) be bilinear groups where $|G| = |G_T| = p$ for some prime p and g be a generator of G. Given a pairing $e : G \times G \rightarrow G_T$ and a collision resistant hash function $H : \{0,1\}^* \rightarrow \{0,1\}^m$, the KGC chooses a randomly chosen value $\alpha \in Z_p$ and computes $g_1 = g^\alpha$. Additionally, the KGC selects three random values $g_2, u, v \in G$ and two vectors $U = (u_i)_n, V = (v_j)_m$ whose elements are chosen from G at random. The system parameters are $params = (G, G_T, e, g, g_1, g_2, u', v', U, V, H)$ and the master secret key is g_2^α.

PPKE(Partial-Private-Key-Extract). Let $u[i]$ denote the ith bit of an identity $u \in \{0,1\}^n$ and $\mathcal{U} = \{i | u[i] = 1, i = 1, ..., n\}$. The KGC picks $r \in Z_p$ uniformly and computes

$$d_u = (d_{u_1}, d_{u_2}) = (g_2^\alpha (u' \prod_{i \in \mathcal{U}} u_i)^r, g^r)$$

UKG(User-Key-Generate). An entity with an identity u chooses randomly a secret value $x_u \in Z_p$ and computes a public key $pk_u = \mathrm{e}(g_1, g_2)^{x_u}$.

PKE(Private-Key-Extract). An entity with an identity u picks $r' \in Z_p$ at random, and computes a private key

$$sk_u = (sk_{u_1}, sk_{u_2}) = (d_{u_1}^{x_u} (u' \prod_{i \in \mathcal{U}} u_i)^{r'}, d_{u_2}^{x_u} g^{r'}) = (g_2^{(\alpha x_u)} (u' \prod_{i \in \mathcal{U}} u_i)^t, g^t)$$

where $t = r x_u + r'$.

SC(Signcrypt). To send a message $M \in G_T$ to the receiver Bob with public key $pk_B = e(g_1, g_2)^{x_B}$, the sender Alice with identity A picks $r_m \in Z_p$ randomly and carries out the following steps

1. Compute $\sigma_1 = M \cdot PK_B^{r_m} = M \cdot \mathrm{e}(g_1, g_2)^{x_B r_m}$;
2. Compute $\sigma_2 = g^{r_m}$;
3. Compute $\sigma_3 = (u' \prod_{i \in \mathcal{U}_B} u_i)^{r_m}$;
4. Set $\sigma_4 = sk_{A_2}$;
5. Compute $m = H(\sigma_1, \sigma_2, \sigma_3, \sigma_4, u_B, pk_B) \in \{0,1\}^m$, where $m[j]$ denotes the jth bit of m and $\mathcal{M} = \{i | m[i] = 1, i = 1, ..., m\}$;
6. Compute $\sigma_5 = sk_{A_1}(v' \prod_{i \in \mathcal{M} v_j})^{r_m}$.

Outputs ciphertext $\sigma = (\sigma_1, \sigma_2, \sigma_3, \sigma_4, \sigma_5)$.

US(Unsigncrypt). Upon receiving a ciphertext σ, the receiver with identity R decrypts the ciphertext as follows

1. Compute $m = H(\sigma_1, \sigma_2, \sigma_3, \sigma_4, u_B, pk_B) \in \{0,1\}^m$, where $m[j]$ denotes the jth bit of m and $\mathcal{M} = \{i | m[i] = 1, i = 1, ..., m\}$;
2. Check that the equality
$$\mathrm{e}(\sigma_5, g) = pk_A \mathrm{e}(u' \prod_{i \in \mathcal{U}_A} u_i, \sigma_4) \mathrm{e}(v' \prod_{j \in \mathcal{M}} m_j, \sigma_2)) \tag{4}$$
If above equation don't hold, output $\perp$. Otherwise, compute and output
$$M = \sigma_1 \frac{e(\sigma_3, sk_{B_2})}{\mathrm{e}(\sigma_2, sk_{B_1})} \tag{5}$$

4.2 Confidentiality Analysis

We give a chosen plaintext manner to attack the Liu et al.'s certificateless signcryption scheme that cannot provide IND-SC-CCA2 security.

When attacker A receives the challenged ciphertext $\sigma^* = (\sigma_1^*, \sigma_2^*, \sigma_3^*, \sigma_4^*, \sigma_5^*)$ from sender u_A, where the ciphertext σ^* is a generated challenged random plaintext $M \in \{M_0, M_1\}$. A performs the following steps:

- A first make a "wild guess" of β to be 0, such that $M_\beta = M_0$;
- Compute $m = H(\sigma_1^*, \sigma_2^*, \sigma_3^*, \sigma_4^*, u_B, pk_B)$;
- Compute $M' = H(M_\beta)$, and generate the corresponding bits set $\mathcal{M}_\beta$ for that $M'_\beta[i]$=1;
- Compute the hash value of sender identity u_A: $H(u_A) = u' \prod_{i \in \mathcal{U}_A} u_i$, such that $\mathcal{U}_A$ is the set such that $u_A[i] = 1$ $(1 \leq i \leq n)$;
- Check the equation
$\mathrm{e}(\sigma_5^*, g) = pk_A \mathrm{e}(H(u_A), \sigma_4^*) \mathrm{e}(v' \prod_{j \in \mathcal{M}_\beta} m_j, \sigma_2^*)$

If above equation holds, A answers that M_0 is the plaintext for the challenged ciphertext σ^*. Otherwise, A knows that M_1 is the plaintext for the challenged ones. Consequently, attacker A has total advantage to guess the chosen ciphertext.

5 Conclusion

In this paper, we analyzed two id-based signcryption in the standard model including an identity based signcryption scheme and a certificateless signcryption scheme, and pointed out that the mentioned schemes cannot resist on the confidentiality attacks to support semantic secure. It is an open issue to construct efficient and semantic secure signcryption scheme and multisigncryption scheme under without random oracles.

Acknowledgment

The authors grateful thank the anonymous reviewers for their helpful comments. This work is supported by the National Natural Science Foundation of China under Grant 60773175 and 60973134, the Foundation of National Laboratory for Modern Communications under Grant 9140c1108020906, the Natural Science Foundation of Guangdong Province under Grants 9151064201000058, and the Support of JSPS Postdoctoral Research Project of Japan.

References

1. Shamir, A.: Identity-based cryptosystems and signature schemes. In: Blakely, G.R., Chaum, D. (eds.) CRYPTO 1984. LNCS, vol. 196, pp. 47–53. Springer, Heidelberg (1985)
2. Boneh, D., Franklin, M.: Identity-based encryption from the weil pairing. In: Kilian, J. (ed.) CRYPTO 2001. LNCS, vol. 2139, pp. 213–229. Springer, Heidelberg (2001)
3. Chow, S.S.M., Yiu, S.M., Hui, L.C.K., Chow, K.P.: Efficient forward and provably secure ID-based signcryption scheme with public verifiability and public ciphertext authenticity. In: Lim, J.-I., Lee, D.-H. (eds.) ICISC 2003. LNCS, vol. 2971, pp. 352–369. Springer, Heidelberg (2004)
4. Canetti, R., Goldreich, O., Halevi, S.: The random oracle methodology, revisited (preliminary version). In: Proceedings of the STOC 1998, Texas, USA, pp. 209–218 (1998)
5. Boneh, D., Boyen, X.: Efficient selective-id secure identity based encryption without random oracles. In: Cachin, C., Camenisch, J.L. (eds.) EUROCRYPT 2004. LNCS, vol. 3027, pp. 223–238. Springer, Heidelberg (2004)
6. Boneh, D., Boyen, X.: Secure identity based encryption without random oracles. In: Franklin, M. (ed.) CRYPTO 2004. LNCS, vol. 3152, pp. 443–459. Springer, Heidelberg (2004)
7. Waters, B.: Efficient identity-based encryption without random oracles. In: Cramer, R. (ed.) EUROCRYPT 2005. LNCS, vol. 3494, pp. 114–127. Springer, Heidelberg (2005)
8. Baek, J., Steinfeld, R., Zheng, Y.: Formal proofs for the security of signcryption. Journal of Cryptology 20, 203–235 (2007)
9. Li, F., Shirase, M., Takagi, T.: Identity-based hybird signcryption. In: ARES'09, pp. 534–539 (2009)
10. Wu, Q., Mu, Y., Susilo, W., Zhang, F.: Efficient Signcryption Without Random Oracles. In: Yang, L.T., Jin, H., Ma, J., Ungerer, T. (eds.) ATC 2006. LNCS, vol. 4158, pp. 449–458. Springer, Heidelberg (2006)
11. Yu, Y., Yang, B., Sun, Y., Zhu, S.: Identity based signcryption scheme without random oracles. Computer Standards & Interfaces 31(1), 56–62 (2009)
12. Liu, Z., Hu, Y., Zhang, X., Ma, H.: Certificateless signcryption scheme in the standard model. Information Sciences 180(3), 452–464 (2010)

JPEG Steganalysis Using HBCL Statistics and FR Index

Veena H. Bhat[1,3], Krishna S.[1], P. Deepa Shenoy[1], Venugopal K.R.[1],
and L.M. Patnaik[2]

[1] Department of Computer Science and Engineering,
University Visvesvaraya College of Engineering, Bangalore, India
[2] Vice Chancellor, Defense Institute of Advanced Technology, Pune, India
[3] IBS-Bangalore, India
{veena.h.bhat,krishna.somandepalli}@gmail.com,
shenoypd@yahoo.com

Abstract. This paper introduces a new statistical model for blind steganalysis of JPEG images. The functionals used to build this model are based on Huffman Bit Code Lengths (HBCL statistics) and the file size to image resolution ratio (FR Index). JPEG images spanning a wide range of resolutions were used to create a 'stego-image' database employing three embedding schemes – the advanced Least Significant Bit encoding technique, JPEG Hide-and-Seek and Model Based Steganography. Existing blind steganalysis techniques mostly involve the analyses of generalized category attacks and the higher order statistics. This work builds an effective neural network prediction model using HBCL statistics and FR Index, which are not yet explored by steganalysts. The experimental results proved to be efficient over a diverse image database and several payloads.

Keywords: Statistical Steganalysis, Steganography, Neural Networks.

1 Introduction

Covert communication, aided by advances in communication and transmission technology, has entered the digital era. Steganalysis is the art and science of detecting messages (payloads) hidden using steganography. It works with scrutinizing the carrier media for anomalies or non-ideal artifacts that are introduced by steganography. Steganalysis is usually carried out by statistically analyzing the cover media for certain ideal features that are susceptible to change on application of steganography. A steganalysis approach is said to be efficient, when the features selected for analyses are accurate, consistent and monotonic in nature.

Certain data mining techniques such as classification and artificial neural networks aid statistical analysis. Steganography has been adapted in aiding nefarious activities including violation of copyright, intellectual property rights and pornography. It has been learnt that terrorist groups use steganography to communicate with their group members, through images over the internet.

H. Chen et al. (Eds.): PAISI 2010, LNCS 6122, pp. 105–112, 2010.
© Springer-Verlag Berlin Heidelberg 2010

2 Related Work

Blind steganalysis is based on the concept of knowledge discovery - detecting the presence of the hidden data in an image with no prior knowledge about the embedding scheme. Some of earliest steganalysis techniques to effectively detect LSB embedding have been developed by Westfield [1]. Further development of LSB Steganalysis based on Chi-square analysis, was by Fridrich [2]. Though these methods are efficient they are restricted to spatial domain analysis. 'Blockiness Attack' which is a measure of discontinuity of 8X8 grid is yet another steganalysis method developed specifically for Model Based Steganography [3], and is efficient in detecting 'Outguess' and 'F5'algorithms [4], [5].

Advanced steganographic methods use efficient methods to hide data in randomized locations, which disable the self-calibration process [6], [7]. Steganalysis of 'Outguess' and 'Steghide' using neural networks has been explored in [8], however this is restricted to analyses over spatial and transform domains only.

Our work targets at active blind steganalysis, taking into consideration images of different resolutions, each of the prevalent embedding schemes based on the domains they embed in (the spatial, frequency/transform domain and adaptive) and varying payloads.

3 Tested Embedding Schemes

Steganographic techniques are classified into spatial domain, frequency domain and adaptive steganography, based on the nature of the embedding technique. Least Significant Bit (LSB) Steganography, the oldest technique in vogue, is a popular spatial domain technique, with the outstanding features of robustness, fine concealment, high steganographic-embedding capacity and easy realization. Advanced LSB which works on both sequential embedding and scattered embedding, is adopted as one of the embedding schemes, to build our stego-image database [9].

Table 1. First order histogram statistics for different embedding schemes

First order statistics	Cover	LSB	JPHS	MBS
Mean	3.053	3.053	3.064	3.054
Variance	5.564	5.564	5.581	5.563
Skewness	0.738	0.738	0.736	0.738
Kurtosis	2.005	2.005	2.003	2.005
Energy	0.239	0.239	0.237	0.239
Entropy	1.744	1.744	1.745	1.744
File-size (kb)	66.17	66.24	66.16	66.20

Table 2. Image dataset details

Basic Set	FR Index	Resized set	FR Index
75X75	0.12 – 0.59	-	-
130X130	0.09 – 0.45	100X100	0.06 – 0.42
214X214	0.09 – 0.42	200X200	0.08 – 0.34
384X256	0.08 – 0.37	300X300	0.06 – 0.32
481X321	0.06 – 0.31	400X400	0.07 – 0.47
512X512	0.03 – 0.41	500X500	0.04 – 0.52
720x480	0.07 – 0.43	600X600	0.07 – 0.37
800X600	0.03 – 0.31	700X700	0.05 – 0.46
1024X768	0.01 – 0.03	800X800	0.02 – 0.05
2048X 1536	0.004 – 0.03	900X900	0.01 – 0.05
		1KX1K	0.005 – 0.03

JPEG Hide-&-Seek (JPHS), a transform domain tool, uses least significant bit overwriting of the discrete cosine transform coefficients (DCT) used by the JPEG algorithm. Though JPHS fails to preserve the DCT histogram statistics, its statistical detectability is among the lowest based on the results reported in [10]. The JPHS stego-image database is populated using the concept proposed by Francisco Echegorri, which hides a text file in a grayscale image by disordering it into lines and columns using the Ranpermut-Encryption algorithm [11].

Model-Based Steganography, an adaptive technique, proposed by Sallee, models statistical properties of an image and preserves them during the embedding process [12]. As detailed in Table 1, it is difficult to differentiate between stego and cover images using first order histogram statistics in blind steganalysis.

4 Input Functionals

4.1 Huffman Bit Code Length Statistics

JPEG, a lossy compression format, employs sequential Huffman Encoding for data compression wherein symbols (DCT coefficients in this case) are taken and encoded with variable length codes that are assigned based on statistical probabilities. A grayscale image employs 2 Huffman tables, 1 each for AC and DC portions. When a JPEG image is embedded with a particular payload, certain non-ideal JPEG artifacts are introduced in the given image, though the enormity of this deviation varies. On analyzing the nature of the DC HBCL statistics of the images in the populated image-database, in this work we have considered Huffman bits of length two to five bits only, as these features extracted from the 6th bits onwards is negligible.

One of the scoring features of Huffman coding algorithm is its ‘unique prefix property’ that is no code is a prefix to any other code, making the codes assigned to the symbols unique. This fact further supports our choice of the HBCL statistics for our evaluation features to be efficient as the JPEG artifacts introduced by steganography on an image becomes unique and hence can be predicted using a suitable classifier or a prediction model. In our work we extracted these statistics using Huffman decoding for a JPEG image in Matlab.

4.2 FR Index – File Size to Resolution Ratio

When a raw image is compressed by JPEG compression, based on the resolution of the image, its quality and compression ratio, the resulting JPEG file takes up a particular file size, indicating the relation between the file size, resolution of an image and its quality. Thus the ratio of file size of the image in bytes to that of its resolution, is found to be unique and also within a certain range for any given resolution. This functional termed ‘FR Index’, is used as one of the inputs to build the prediction model. Table 2 shows the range of FR Index for the image resolutions considered in our stego-image database.

5 Image Database

One of the important aspects of the performance evaluation of a steganalysis technique is the dataset employed in the experiment(s) that validates the prediction model for its sensitivity and specificity. As JPEG image format is in vogue, both in the public domain and the World Wide Web, we chose to work on cover images of JPEG format. Moreover, it is harder to detect hidden data in JPEG grayscale images when compared to color images wherein steganalysis can utilize dependencies between color channels.

The entire image dataset is within the quality range of 57 to 72, approximated by JPQ-JPG Quality Estimator [13]. Table 2 illustrates details of the image database used in our performance evaluation. The Basic Set lists the different resolutions used with 100 images in each set; the Resized Set lists the set of images correspondingly resized from the Basic Set to validate against the JPEG Resizing Error.

For each tested method – Advanced LSB, JPHS and MBS, the cover grayscale JPEG images and several stego grayscale JPEG images embedded with different payloads are prepared. Thus for the entire set of 2,000 cover images, 18,000 stego images with different parameters were generated and evaluated. In order to check for consistency of the features extracted for the model and to validate this over JPEG Resizing Error, the basic dataset is resized. Thus, a set of 2000 JPEG cover images are obtained on which the consistency of the prediction model is evaluated.

Three different payloads for Advanced LSB, JPHS and MBS are tested for the 2,000 images. In case of LSB and MBS embedding schemes, a random data corresponding to the payload size is embedded, where as in JPHS a text file of the respective size is embedded after encryption. In case of MBS, we used optimized Huffman tables and the embedding rate on an average was found to be 1.0431 bits per change (bpc).

For images with resolution 100X100 and less, 340 bytes, 500 bytes and 720 bytes were the three payload sizes used; however our prediction model produced consistent results for payload sizes less than 300 bytes too. For the remaining resolutions up to that of 2048X1536, we used payloads of sizes 1024 bytes (1 KB), 2048 bytes (2KB) and 5120 bytes (5KB). Thus the prediction model is evaluated over 20,000 images.

Images of resolutions 384X256, 481X321, 1024X768 and 2048X1536 were collected from standard image datasets and photographers' archives. In order to validate the images collected as genuine, they were tested across the existing freeware steganographic softwares. However the possibility of images identified as genuine being stego-images does exists, which there-by resulted in the classifier's high false positive rate.

6 Model

The architecture of the prediction model is shown in Fig. 1. The selected cover images of various resolutions are processed through the 'steg-system' where the images are embedded with different payloads and using LSB, JPHS and MBS. These images are then processed through the functional generator where the image file-size, quality, resolution and HBCL statistics are generated. The identified inputs with their respective targets are used to train a classifier.

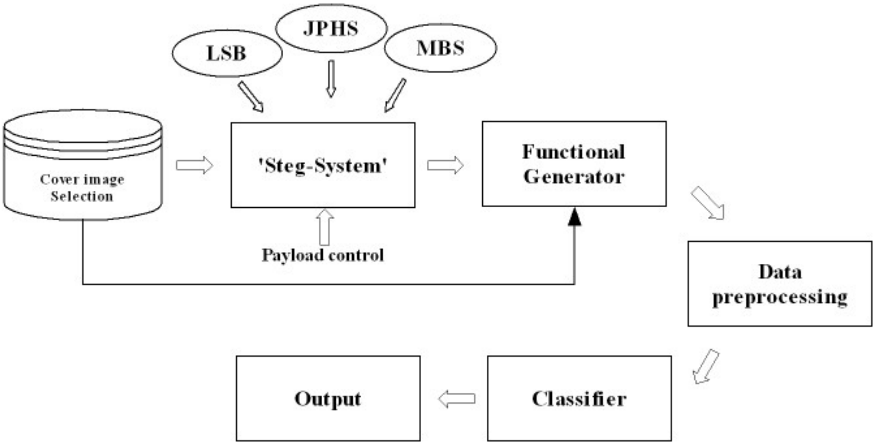


Fig. 1. Architecture of the prediction model

- Artificial Neural Network (ANN), a supervised learning data mining approach, is chosen as the classifiers as it is sensitive to non-linear input values and high on prediction efficiency. The model used for training is a Multilayer Feed Forward Neural Network with back propagation, with two hidden layers - the first hidden layer has 10 neurons and Softmax as the activation function while the second layer has 20 neurons and is activated by the Hyperbolic Tangent (TANH) function. The weights for each of the attribute are not manually assigned, instead the system is allowed to run few initial trials and accordingly assign weights for a consistent result. The variation of misclassification rate with respect to different activation functions, hidden layers and the number of neurons are as illustrated in Table 4.
- Softmax and Hyperbolic Tangent Activation functions were considered to train the neural network, the number of neurons in each hidden layer and further the number of hidden layers are chosen for consistent and efficient results over several different training and test sets.
- The predicted output of our model is required to be 0 (‘Clean’ image) or 1 (Stego-image), for this purpose it is highly desirable for the outputs to lie between zero and one, further summing to 1 so that they can be interpreted as posterior probabilities for a given categorical target variable. These constraints can be enforced on the output using a Softmax activation function given by:

$$p_i = \frac{e^{q_i}}{\sum_{j=1}^{n} e^{q_j}} \cdot \qquad (1)$$

  where p_i is the softmax output and q_i is the net input to each output unit, where i=1...c, and c the number of categories.
- Our classifier prediction model requires customization of the hidden layers due to the non linear variation of the HBCL statistics with respect to FR Index, hence Hyperbolic Tangent (TANH) activation function has been employed effectively

to perform this, however the results were the same when Tangential Sigmoid (TANSIG) activation function is employed as both TANH and TANSIG perform the same function when used in a hidden layer. However, since we use a Multilayer Perceptron, TANH is the preferred activation function. TANH function is given by

$$tanh = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} . \tag{2}$$

7 Results and Performance Analysis

A preliminary analysis was performed to test the correlation between the features extracted from the images, shown in Table 3. These features form the input parameters for the model. As the features show a non linear response in the feature space, neural networks was selected as the classifier.

Table 3. Correlation values for the input attributes selected for steganalysis

	FR Index	Quality	2_HBCL	3_HBCL	4_HBCL	5_HBCL
FR Index	1					
Quality	0.5889	1				
2_HBCL	-0.3809	- 0.0444	1			
3_HBCL	-0.0151	0.0227	-0.0622	1		
4_HBCL	0.1999	0.0666	-0.2633	-0.6701	1	
5_HBCL	-0.0120	-0.0662	-0.0304	-0.8477	0.2645	1

The neural network simulations were performed on SAS for Windows and the results were cross checked using Matlab. The Multilayer Perceptron Model was trained and evaluated for the entire training set. Separate datasets were created for each of the embedding scheme and identified payloads and were tested against the classifier.

Table 4. Performance of MLP for various activation functions in hidden layers

Activation Function	No. of hidden layers	Neurons per layer		Accuracy (%)	Misclassification rate(%)
Arc Tangent	1	20		90.673	9.328
TANH	1	20		90.672	9.638
Softmax	1	20		91.134	8.866
SoftMax+TANH	2	20	10	91.259	8.741
SoftMax+TANH	2	10	20	92.086	7.914
SoftMax+TANH	2	5	10	92.035	7.966

Table 5. Prediction accuracy of the prediction model

Model	Sensitivity	Specificity	Correctness
JPHS	67.6403	92.4616	80.9286
LSB	67.1791	97.5333	83.4345
JPHS	67.6403	92.4616	80.9286
MBS	96.3105	97.7333	97.0725
ALL	56.3036	99.3991	89.2816

The efficiency and consistency of the model is evaluated using sensitivity (true positive rate), specificity (true negative rate) and correctness, as shown in Table 5. The higher these parameters are, more efficient is the prediction model.

8 Conclusions and Future Work

This statistical model is efficient over the three embedding schemes described here, however the model is highly sensitive to MBS. The model is sensitive to payloads as low as 0.14 bpc (MBS) and 100 bytes (JPHS, LSB). The functionals that build this model are accurate, consistent and monotonic thus reducing the false alarm rate that may occur due to resizing and recompressing of a JPEG image.

The model has to be fine tuned to reduce the false positive rates. The statistical model using HBCL statistics and FR Index has to be tested on wider embedding schemes such as PQ, MMx, etc. We need to explore other prediction techniques such as Support Vector Machines (SVM) and Rough Sets on our database in order to improve the accuracy of the model. The final goal of this preliminary work, however, is to extend the fine tuned prediction model to extract the hidden message in any stego-image irrespective of the embedding scheme adopted. The model has to be evaluated over other classification techniques to reduce the false positive rate; however the reason for the high false positive rate is dependent on the features extracted and needs to be explored.

References

1. Westfeld, A., Pfitzmann, A.: Attacks on Steganographic Systems. In: Pfitzmann, A. (ed.) IH 1999. LNCS, vol. 1768, pp. 61–76. Springer, Heidelberg (2000)
2. Fridrich, J., Goljan, M., Du, R.: Reliable Detection of LSB Steganography in Grayscale and Color Images. In: Proceedings of the 2001 Workshop on Multimedia and Security – New Challenges, ACM Special Interest Group on Multimedia, pp. 27–30 (2001)
3. Christian, U., Westfeld, A.: Weakness of MB2. In: Shi, Y.Q., Kim, H.-J., Katzenbeisser, S. (eds.) IWDW 2007. LNCS, vol. 5041, pp. 127–142. Springer, Heidelberg (2008)
4. Fridrich, J., Goljan, M., Hogea, D.: Attacking the OutGuess. In: Proceedings of ACM: Special Session on Multimedia Security and Watermarking, Juan-les-Pins, France (2002)
5. Fridrich, J., Goljan, M., Hogea, D.: New Methodology for Breaking Steganographic Techniques for JPEGs. Submitted to SPIE: Electronic Imaging 2003, Security and Watermarking of Multimedia Contents. Santa Clara, California (2003)
6. Jing, D., Wei, W., Tieniu, T.: Multi-class Blind Steganalysis Based on Image Run-Length Analysis. In: Ho, A.T.S., Shi, Y.Q., Kim, H.J., Barni, M. (eds.) Digital Watermarking. LNCS, vol. 5703, pp. 199–210. Springer, Heidelberg (2009)
7. Kaushal, S., Anindya, S., Manjunath, B.S.: YASS: Yet Another Steganographic Scheme That Resists Blind Steganalysis. In: Furon, T., Cayre, F., Doërr, G., Bas, P. (eds.) IH 2007. LNCS, vol. 4567, pp. 16–31. Springer, Heidelberg (2008)
8. Zuzana, O., Jiri, H., Ivan, Z., Roman, S.: Steganography Detection by Means of Neural Network. In: Ninteenth International Conference on Database and Expert Systems Application, pp. 571–574. IEEE Computer Society, Los Alamitos (2008)

9. LSB Source code, `http://www.advancedsourcecode.com/lsbsteganography.asp`
10. Pevný, Fridrich, J.: Merging Markov and DCT features for Multi-class JPEG Steganalysis. In: Proceedings SPIE - Electronic Imaging, Security, Steganography, and Watermarking of Multimedia Contents IX, San Jose CA, vol. 6505, pp. 3–4 (2007)
11. Source code, `http://www.mathworks.co.uk/matlabcentral/fileexchange/1747-steganograph`
12. Sallee, P.: Model-based Methods for Steganography and Steganalysis. International Journal of Image Graphics (1), 167–190 (2005)
13. `http://www.softpedia.com/get/Multimedia/Graphic/Graphic-Editors/JPEG-Quality-Estima-tor.shtml`

Text Mining Technique for Chinese Written Judgment of Criminal Case

Shihchieh Chou and Tai-Ping Hsing

Dept of Information Management, National Central University, No. 300,
Jhongda Rd., Jhongli City, Taoyuan County 32001, Taiwan, R.O.C.
{scchou@mgt,92443004@cc}.ncu.edu.tw

Abstract. Text mining has become an effective tool for analyzing text documents in automated ways. Conceptually, clustering, classification and searching of legal documents to identify patterns in law corpora are of key interest since it aids law experts and police officers in their analyses. In this paper, we develop a document classification, clustering and search methodology based on neural network technology that helps law enforcement department to manage criminal written judgments more efficiently. In order to maintain a manageable number of independent Chinese keywords, we use term extraction scheme to select top-*n* keywords with the highest frequency as inputs of the Back-Propagation Network (BPN), and select seven criminal categories as target outputs of it. Related legal documents are automatically trained and tested by pre-trained neural network models. In addition, we use Self-Organizing Map (SOM) method to cluster criminal written judgments. The research shows that automatic classification and clustering modules classify and cluster legal documents with a very high accuracy. Finally, the search module which uses the previous results helps users find relevant written judgments of criminal cases.

1 Introduction

As the development of information technology, digital content and documents have a significant raise, people cannot use manual ways to find the necessary information. As a result, text mining techniques are implemented by enterprises and organizations to manage their information and knowledge more effectively. In this research, we focus on legal documents such as written judgments of criminal cases. Written documents of criminal cases describe the facts of crime, as well as which legal provisions regulate these crimes, and how the judge passed the last sentence of this crime. For example, when a criminal trial is conducting, the judge may want to find the similar cases that occurred before to help him conduct a correct judgment, and avoid a misjudgment. In addition, when the police investigate a criminal case, they will be able to discover the modus operandi of similar cases to dig out more clues. The objective of this research is to develop an effective methodology to automatically classify, cluster the written judgments, and identify the similar cases.

A prototype system is implemented and tested by using the written judgments of criminal cases which is taken from the written judgment retrieving system in the

H. Chen et al. (Eds.): PAISI 2010, LNCS 6122, pp. 113–125, 2010.
© Springer-Verlag Berlin Heidelberg 2010

Judicial Yuan of Republic of China. The implementation of our research is described in the following steps: First, the significant Chinese keywords of written judgments are abstracted, and their frequencies are computed by the Chinese word segment system the Academia Sinica developed, which is a free Chinese word segment system and widely used in Taiwan. Here, we developed a keyword database based on this word segment system. Second, we use term extraction scheme to select the top-100 keywords with the highest frequency as the input vectors of the back-propagation network, and the seven criminal categories as target output vectors of it. Third, the neural network model is trained by using the 140 sample documents. The trained model is assessed until it reaches a satisfactory level of accuracy. After the network model is trained, we use 70 validation documents to verify the precision of the classification. Fourth, we use 140 training documents as same as the classification model to train clustering model of Self-Organization Map (SOM). The final step is to combine the classification and clustering models for an automated search to identify the similar written judgments.

2 Literature Survey

Legal documents are generally presented in the form of natural language text –a combination of segments. Some are necessary, and the others are optional, arranged in a fixed or partially fixed order. A segment which often presents a specific topic, such as a criminal written judgment, may be described in the dispositif, corpus delicti, or trial grounds.

Archivists of legal documents traditionally depend upon legal thesauri and classification structures. Legal documents most commonly used in retrieval systems are statutes (legislation) and court decisions (cases). Because the vocabulary of legal texts is diverse and many terms have a specific semantic meaning, retrieval of legal decisions is more complicated. This makes retrieval of cases without human intervention (e.g., in indexing the cases) an extraordinarily difficult task.

Text mining is one of the most important techniques of knowledge management systems. It can extract sensitive information from text corpora. It is a highly valuable tool for analyzing legal documents in automated ways. Law experts have a strong need for identifying relevant patterns and legal cases in law corpora. Schweighofer (1999) has an automated text analysis of international law. Moens (2001) gives an overview of the state of the art of these innovative techniques and their potential for legal text retrieval. Conrad et al. (2005) has conducted research on clustering of primary and second law documents as well as actual law firm data.

The judiciary or law firms in Taiwan have implemented a number of retrieving systems in statutes law and the written judgments of legal cases, for example, Judicial Yuan has already deployed a legal information retrieval system which provides the ordinary people or legal experts to search the latest regulations, judicial interpretations, and the written judgments of court decisions. These systems can provide the keyword search in full-text of law or legal cases, or a search according to a specific time, sentence number, or sentence cause of court decisions. However, these systems provide too much information, so the users need to spend a lot of time and efforts to filter. Intelligent retrieval systems of Chinese legal document are still lacking, and Chinese word segment and syntax analysis are more difficult than English. Hence, the main purpose of our study is to use text mining and neural network technology to implement an intelligent legal information retrieval system.

3 System Methodology

This section depicts the details of the methodoligies for term extraction model, vector space model, document classification and clustering. First, a document content extraction model is built to represent the document content with a vector consisting of keyword frequencies. Second, a document classification model based on the Back-Propagation Network (BPN) approach is developed. Third, a document clustering model based on the Self-Organization Map (SOM). Finally, a document search model is implemented by combining the results of a trained BPN and SOM.

3.1 Term Extraction and Frequency

Term recognition is typically the first step in text document for information extraction. The stopping, stemming and splitting processes are used to segment sentences (Selamat and Omatu, 2004). Stopping is the process of removing repetitive and low-meaning words. Stemming is the procedure of reducing words to their original roots (Lovins, 1968) and segmenting is the process of splitting a sentence into segments or individual words separated by blanks. After stopping, stemming, and segmenting, the term-weight in the document is calculated. Term frequency (TF) and inverse document frequency (IDF) are the two parameters to measure the weight of a term within text, Low TF and IDF terms are often removed from the indexing of a collection (Salton and Buckley, 1988), and the term-weighting schemes can be expressed as:

$$w_{jk} = tf_{jk} \times idf_j \tag{1}$$

where $w_{jk}$is the weight of term j in document k, $tf_{jk}$is the number of term j that appears in document k, and $idf_j$is the inverse document frequency of term j as derived in equation (2):

$$idf_j = log_2\left(\frac{n}{df_j}\right) \tag{2}$$

where n is the total number of documents in the target set, and $df_j$is the number of documents in which the index term j appears. When the idf_jvalue increases, the term j representing specific documents becomes more significant as proposed by Horng and Yeh (2000). Finally, the top ranked terms with high w_{jk} values are identified as the terms for the given document k.

3.2 Vector Space Model (VSM)

The VSM can evaluate the degree of similarity between documents and used for automatic document classification and clustering. There are three key steps where terms are first extracted from the document text, then the weights of the indexed terms are calculated to improve the document retrieval accuracy, and the documents are ranked with respect to a similarity measure (Raghavan and Wong, 1986). VSM is a multi-dimensional vector where each feature of a document is a dimension. For instance, term frequency (TF) and inverted document frequency (IDF) are two factors of a text document. After the vector of a text document is derived, a cosine function is applied to measure the similarity between two documents (Tam, Santoso, and Setiono, 2002). For example, the vector of document *X* is represented by $X = (x_1, x_2, \cdots, x_n)$, where x_i represents the *i*th feature of document *X*. Likewise, *Y* $=(y_1, y_2, \cdots, y_n)$ represents the vector of document *Y*. This similarity between X and *Y* can be calculated as

$$\cos(X, Y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \times \sum_{i=1}^{n} y_i^2}} \tag{3}$$

3.3 Automated Document Classification

Document classification is a model which predicts the pre-defined class of document when the other attributes are given. Artificial Neural Networks (ANN) are frequently used for document classification and learn by using training data to adjust the weights between connecting nodes. The Back-Propagation Network (BPN) algorithm is the most widely used model among artificial neural networks, which is a multi-layered network first proposed by Rumelhart, Hinton, and Williams (1986). The back-propagation network learning consists of two phases performed through the different layers: a forward pass and a backward pass. The following sections describe these two phases.

Forward pass

In the forward pass, a training sample (input data vector) is applied to the input nodes of the network, and its effect propagates through the network layer by layer. The input of every node is calculated. Then, the output from the activation functions of the nodes is derived and is passed to the next layer where processing continues until reaching the final output layer. The final outputs are the actual response of the network. The net input from input layer to the hidden layer node *j* is calculated as the following:

$$\text{net}_j^h = \sum_{i \in \text{previous layer}} w_{ij}^h x_i + b_j \tag{4}$$

where w_{ij}^h is the weight of connection between input layer node *i* and hidden layer node *j*, x_i is the input of node *i*, and b_j is the bias associated with node *j*. The output of node *j* is H_j as a certain function of net_j^h value by using the following equation:

$$H_j = f\left(net_j^h\right) \tag{5}$$

The activation function, $f(x) = \frac{1}{1+e^{-x}} + c$, where *c* is a constant (e.g., $c = 0$ or $c = 5$) and the net input from hidden layer to output layer node *k* is computed as

$$net_k^0 = \sum_j w_{jk}^0 \, H_j \tag{6}$$

where $w_{jk}^0$is the weight of connection between hidden layer node j and output layer node k. Finally, we can determine the output of the neural network.

$$O_k = \mathrm{g}(net_k^0) = \mathrm{g}\left(\sum_j w_{jk}^0 H_j\right) \tag{7}$$

where g(x) is the activation function of node k. The error of the network is the following equation:

$$E = \frac{1}{2}\sum_k (T_k - O_k)^2 \tag{8}$$

where T_k is the desired output of the training data.

Backward pass

The backward pass starts at out layer, passing the error signal (the difference between the computed and the real output value) leftward through the network, layer by layer, and recursively computing the local gradient δ_k^0 for each neuron. Since E is defined as the function of O_k, and $O_k$is the function of w_{jk}^0, the weight adjustment between output layer and hidden layer Δw_{jk}^0 can be expressed as

$$\begin{aligned}\Delta w_{jk}^0 &= -\eta \frac{\partial E}{\partial w_{jk}^0} = -\eta \frac{\partial E}{\partial O_k}\frac{\partial O_k}{\partial w_{jk}^0} \\ &= \eta (T_k - O_k)\mathrm{g}'(\mathrm{net}_k^0) H_j = \eta \delta_k^0 H_j\end{aligned} \tag{9}$$

where η is the learning rate and $\delta_k^0 = (T_k - O_k)\mathrm{g}'(\mathrm{net}_k^0)$. For the neuron located at the output layer, δ_k^0 is equal to error signal of that neuron multiplied by the first derivative of its nonlinearity represented in the activation function. Based on local gradients δ_k^0, it is straightforward to compute Δw_{jk}^0 for each connection to the output node. Given the δ_k^0 values for all neurons in the output layer, we use them in hidden layer to compute modified local gradients δ_j^h , and to correct Δw_{ij}^h for input connections for this layer. Since $O_k$is a function of H_j, the function of w_{ij}^h is the weight adjustment between hidden layer and input layer:

$$\begin{aligned}\Delta w_{ij}^h &= -\eta \frac{\partial E}{\partial w_{ij}^h} = -\eta \frac{\partial E}{\partial H_k}\frac{\partial H_j}{\partial w_{ij}^h} = -\eta \frac{\partial E}{\partial O_k}\frac{\partial O_k}{\partial H_j}\frac{\partial H_j}{\partial w_{ij}^h} \\ &= \eta (T_k - O_k)\mathrm{g}'(\mathrm{net}_k^0) w_{jk}^0 f'(net_j^h) H_j \\ &= \eta \textstyle\sum_k \delta_k^0 w_{jk}^0 f'(net_j^h)\, H_j = \eta \delta_j^h H_j\end{aligned} \tag{10}$$

where $\delta_j^h = f'(net_j^h) \sum_k \delta_k^0 \, w_{jk}^0$.

Therefore, the weight adjustment is depicted as

$$\Delta w_{ij} = \eta \delta_j f(net_i) \tag{11}$$

where δ_j is the output error of layer j, net_i is the input of layer i.

The backward procedure is repeated until all layers are covered and all weight factors in the network are modified. Then, the backward-propagation algorithm continues with a new training sample. When there are no more training samples, the first iteration of the learning process finishes. With the same samples, it is possible to go through hundreds of iterations until average error energy for the given iteration is small enough to stop the algorithm.

3.4 Automated Document Cluster

Document clustering divides a set of documents into groups without using pre-defined classes, and splits many documents into groups using a measure of similarity, so that the documents in one group are similar and documents belonging to different groups are different. The Self-Organization Map (SOM) algorithm is also the most widely used clustering model among artificial neural networks, which is an unsupervised learning network first proposed by Kohonen (1997). The SOM algorithm assigns a set of high-dimensional patterns as a vector into a two-dimensional map of neurons according to the similarities among vectors. Similar patterns will map to the same or nearby neurons after the training process. When a pattern is presented, one of the output neurons is selected as a "winner", and the weights connecting the patterns to that neuron will be strengthened, and the neurons in neighborhood of the winning neuron are also strengthened their weights (although not as much). Once learning is completed, "similar" patterns will map into the same or neighboring neuron. In this way, similar patterns can be clustered in a group.

Let $x_i \in \Re^N$, $1 \le i \le M$, x_i be the encoded vector of the i^{th} document in the corpus, where N is the number of indexed terms and M is the number of the documents. These vectors are applied to training inputs in the SOM network. The network consists of a regular grid of neurons each of which has N synapses. Let $w_j = \{w_{jn} | 1 \le n \le N\}, 1 \le j \le J$, w_j be the synaptic weight vector of the j^{th} neuron in the network, where J is the number of neurons in the network. SOM algorithm can be expressed as:

Step 1. Randomly select a training vector x_i from the corpus.
Step 2. Find the neuron j with synaptic weight vector w_j which is the closest to x_i, i.e.

$$\left\| x_i - w_j \right\| = \min_k \left\| x_i - w_k \right\| \tag{12}$$

Step 3. For each neuron l in the neighborhood of neuron j, update its synaptic weights by

$$w_l^{new} = w_l^{old} + \alpha(t)(x_i - w_l^{old}) \tag{13}$$

where $a(t)$ is the training gain at time stamp t.

Step 4. Increase time stamp *t*. If *t* reaches the preset maximum training time *T*, halt the training process; otherwise decrease *a*(*t*) and the neighborhood size, goto Step1.

when *T* is large enough so that every vector may be applied as training input for certain times, the training process stops. The training gain and neighborhood size both decrease when *t* increases.

4 System Analysis and Design

The criminal written judgments sourced from the written judgment retrieving system of the Judicial Yuan in Taiwan are applied to our study. Our system is implemented in the following sub-sections.

4.1 Data Preparation

We download the 210 criminal written judgments sourced from the written judgment retrieving system of the Judicial Yuan in Taiwan as training and testing samples. Seven criminal categories including homicide, sex crimes, drug related crimes, corruption, computer related crimes, robbery, and fraud are selected as research targets. These crimes are easier committed by a number of specific criminals in a specific way. Law experts and the police have a strong interesting to search similar cases. After downloading these samples, we have to remove the HTML tags from these legal documents to implement the following processes.

4.2 Keyword Recognition

The significant keywords of the written judgments are abstracted, and their frequencies are computed from Chinese word segment system developed by Taiwan’s Academia Sinica. This system is a free Chinese word segment system and widely used in Taiwan. We build a keyword database based on this system to save developing time and efforts. In addition, the TF-IDF values of keywords are computed, and extracted the top-20 keywords with the highest TF-IDF values as a representation of the given document.

4.3 Document Classification

The extracted keywords of all documents by the previous step are joined together into a union set which has 2604 keywords totally, we compute the frequencies of these keywords in all documents, and then we select the top-100 with the highest frequencies as the input vectors of the back-propagation network, and assign seven criminal categories as target output vectors of it. The neural network model is trained using the 140 sample documents. The trained model is assessed until it reaches a satisfactory level of accuracy. After the network model is trained, we use 70 sample documents to verify the precision of the classification.

Here, we also implement a word segment process in the specific provisions of seven criminal categories applied as outputs of BPN, and then we invite two law experts who teach criminal law courses in the university to check these segmented words, they select 251 keywords with the legal meanings. We think these selected keywords should be more significant than other ones, hence, when the keywords in the samples belong to one of 251 keywords, their TF-IDF values and frequencies would be doubled to increase their weights. We want to know whether the weighting method can increase the precision of classification.

Besides, we want to realize whether fewer segments where the keywords are extracted can achieve a close classification precision which is achieved by using the whole written judgment. So, we will only use both segments of dispositif and corpus delicti as the representation of the whole written judgment, because the dispositif abstracts the applied articles and measure punishment, and the corpus delicti illustrates the whole criminal process, we think that these keywords extracted from both segments should have a more significant weight and be a representative of this written judgment. Hence, we will compare the difference of the classification precisions between both segments and all segments as the inputs of BPN.

In order to answer the previous questions, four keyword extracting models are applied to BPN in our study, they can be expressed as:

Table 1. Four keyword extracting models as input vectors of BPN

	The unweighted input vectors	The weighted input vectors
All segments in written judgment	Model 1	Model 2
Both segments in written judgment	Model 3	Model 4

1. Model 1: the training and test samples use all segments of the written judgment, and the unweighted vectors are used as the inputs of BPN.
2. Model 2: the training and test samples use all segments of the written judgment, and the weighted vectors are used as the inputs of BPN.
3. Model 3: the training and test samples use both segments of the written judgment, and the weighted vectors are used as the inputs of BPN.
4. Model 3: the training and test samples use both segments of the written judgment, and the unweighted vectors are used as the inputs of BPN.

We construct and train BPN model for classification. There are five parameters for the network:

1. Number of layers: The network has three layers, one input layer, one hidden layer, and one output layer.
2. Number of input nodes: The input nodes correspond to the top-100 keywords with highest frequencies in each document, so the number of input nodes is 100.
3. Number of output nodes: The output nodes correspond to seven criminal categories, so the number of output nodes is seven.
4. Number of hidden layer nodes: The number of hidden nodes is 26, it meets the experience rules of most researchers, i.e. where *n* is the nodes of input layer, *m* is

the nodes of output layer, and s is the nodes of hidden layer, and can be expressed as $s = \sqrt[2]{n \times m}$

5. Activation function: The activation functions of input layer, hidden layer and output layer are set as LOG-Sigmoid.

4.4 Document Clustering

Most of text mining systems apply either classification or clustering module. Our study hopes this system can not only provide classification results, but also clustering ones. Because the classification is a kind of supervised learning algorithm, we select training samples which were classified by the judge. A legal case might be the concurrence or several criminalities. In general, the judge might evaluate which crime will have the most severe penalty to decide which class this criminal case should be classify, and record it in the written judgment. So the classifications of our samples are based on the results of measure punishment in a legal case by the judge. However, a legal case associates a lot of complexity factors, for example, a criminal behavior can result in a combined punishment or lapping of legal provisions. Two different categories of crime may have a similar modus operandi or acts. Hence, we want to use clustering technique with an unsupervised learning characteristic to identify the likely similar legal cases. We use 140 training samples as same as the classification model to train clustering model of Self-Organization Map (SOM).

The desired output of SOM network is a two-dimensional map of 144 neurons in 12×12 grid format. Each neuron in the map contains 100 synapses. The initial training gain is set to 0.4 and the maximal training time is set to 500. These settings are also determined experimentally. We had tried different gain values ranged from 0.1 to 1.0 and various training time setting ranged from 50 to 500. We simply adopt the setting which achieves the most satisfying result.

4.5 Document Search

When searching for a legal case, a user assigns a set of related keywords as a query which represents the target. The system imports the 0/1 vector into the trained neural networks. Following the models for document classification and clustering, the outputs are calculated separately. The system joins both outputs of the classification and clustering model to recommend the similar cases ordered by similarities between the query and the recommended documents.

5 Results

5.1 Classification Results

To evaluate the effectiveness of the proposed methodology, 140 training legal documents were imported to derive the neural network model with acceptable error. In this paper, the classification precision is measured (Salton, 1973) using the following equation:

$$p_i = \frac{A_i}{A_i + B_i} \tag{14}$$

where P_i is the precision of category i, A_i the number of legal documents that are classified to category i correctly, and B_i the number of legal documents that are classified to category i incorrectly.

As shown in Table 2, 70 legal documents are used to test the system performance. The testing precision of the classification in Model 1 is 94%. Model 2 is 96%, Model 3 is 67%, and Model 4 is 70%. This result shows the weighted scheme with expert knowledge can raise a higher precision and increase about 2% to 3%, but it isn't very significant. Besides, Model 3 is a lower precision, and there is a bid gap between Model 1 and Model 3. This result also shows that all segments of the written judgment as input vectors have a much higher precision than the both segments of it. We try to know the reason why Model 3 is much lower than Model 1. We think the frequency and TF-IDF value of each keyword in Model 3 is much lower than Model 1. It will result that the generated representative keywords of each document in Model 3 can't be effectively represented the characteristics of the original written judgment.

Moreover, we find the computer related crimes and drug related crimes classes have a very high precision in any Model. After investigating the representative keywords of all written judgments in the both classes, we find that high frequency keywords of each document have a relatively high consistency in their own classes. In addition, corruption class have a lower precision in any Model, when we check the representative keywords of all written judgments in this class, we find that the high frequency keywords among documents have a higher inconsistency. Besides, we also find the homicide, fraud, robbery, and corruption classes have a sensitive precision change with the number of the segment.

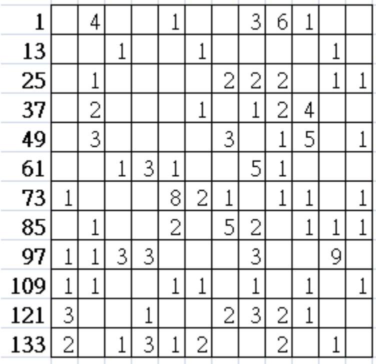
Table 2. The classification precisions of each class and total samples for 4 classification models

	Model 1	Model 2	Model 3	Model 4
Number of test documents	70	70	70	70
Number of homicide class classified correctly	9	10	4	4
Number of fraud class classified correctly	10	10	7	7
Number of computer-related crimes class classified correctly	9	9	10	10
Number of drug related crime class classified correctly	10	10	9	9
Number of robbery class classified correctly	10	10	5	7
Number of corruption class classified correctly	9	8	4	5
Number of sex crimes class classified correctly	9	10	8	7
Number of legal documents classified correctly	66	67	47	49
Number of legal documents classified incorrectly	4	3	23	21
Precision	94%	96%	67%	70%

5.2 Clustering Results

Because Model 2 has a highest precision, we apply the weighted input vectors of it as inputs of SOM, and the clustering result shows as Fig 1. Each grid in the map represents a neuron. Starting from neuron 1 in the upper left corner of the map, the neuron index increases row by row to the lower-right corner.

We can observe that documents labeled to the same cluster are similar in context. For example, the cluster of neuron 2 which contains 4 documents related to homicide, the cluster of neuron 58 which contains 5 documents related to fraud, the cluster of neuron 77 which contains 8 documents related to drug related crimes, and the cluster of neuron 107 which contains 9 documents related to corruption. Moreover, neuron 8 and 9, which locate closely in the map, and all documents related to drug related crimes. Neuron 26, 38 and 50, which locate closely in the map, and all documents related to rubbery. Neuron 99 and 100, which locate closely in the map, and all documents related to sex crimes.



1		4			1			3	6	1		
13			1			1					1	
25		1					2	2	2		1	1
37		2				1		1	2	4		
49		3					3		1	5		1
61			1	3	1			5	1			
73	1				8	2	1		1	1		1
85		1			2		5	2		1	1	1
97	1	1	3	3				3			9	
109	1	1			1	1		1		1		1
121	3			1			2	3	2	1		
133	2		1	3	1	2			2		1	

Fig. 1. The document cluster map for 140 training samples. We only show the number of document associated with the same neuron. The starting neuron index of each row is shown on the left of the map.

5.3 A Case Study for Searching Model

As describe in Section 4.5, our system implements the search function. A user assigns manually a set of related keywords as a query which represents the target someone wants to search, and these keywords would be the input vectors in our system. We describe a scenario to express our searching function. For example, we select the following nine keywords {攻擊 (attack), 兇器 (a tool or weapon for criminal purpose), 死 (death), 被害人 (the victim), 機車 (motorbike), 頭部 (head), 殺人 (homicide), 攜帶 (carry), 告訴人 (Prosecutor)} as the input vectors of BPN and SOM. The classification module classifies the query to the homicide category.

In addition, the clustering module assigns the query to the cluster of neuron 38. We select the documents inside the assigned node and its neighboring nodes (the distance equals 1 from them to neuron 38) as the recommended cases. In our example, the documents inside neuron 38 are Document No. 7 and 19 which belong to homicide class, and the documents inside the neighboring nodes are Document No. 9, 16, 88, 92, and 97 (Document No. 9 and 16 are also classified to homicide class in the classification module, and Document No. 88, 92, and 97 are classified to robbery class in it). Hence, we join the classification and clustering results to recommend the similar documents ordered by similarities as shown in Table 3.

We find that the documents which similarities are over 0.5 have the victims whose fatal wounds are on the head, or the criminal cases have a motorbike to leave the scene of the crime. However, Document No. 92, 88, and 97 belong to the rubbery class, and have a very low similarities between the query and the representative keywords of these documents. The higher frequencies of keywords in Document No. 92 include 機車 (motorbike) and 兇器 (a tool or weapon for criminal purpose), and it has a higher similarity than Document No. 88 and 97, and it is even higher than Document No. 9 and 7 taken from the classification module. We also find the similarities of Document No. 88 and 97 are approximate zero, and it results from the vectors of the query which scarcely exist in the representative keywords of these two documents.

Table 3. The recommended results for the searching vectors

Rank	Document No.	Similarity	Class	Rank	Document No.	Similarity	Class
1	5	0.72279	homicide	13	1	0.44031	homicide
2	20	0.69027	homicide	14	10	0.42407	homicide
3	3	0.64285	homicide	15	15	0.41704	homicide
4	2	0.62812	homicide	16	11	0.29600	homicide
5	18	0.62681	homicide	17	13	0.19893	homicide
6	14	0.60465	homicide	18	6	0.19626	homicide
7	19	0.57577	homicide	19	92	0.15910	robbery
8	17	0.55162	homicide	20	9	0.12774	homicide
9	4	0.51299	homicide	21	7	0.06904	homicide
10	16	0.50757	homicide	22	88	0.00000	robbery
11	12	0.49512	homicide	23	97	0.00000	robbery
12	8	0.46071	homicide				

6 Conclusion

In this paper, we present a methodology for classifying and clustering criminal written judgments automatically using back-propagation network (BPN) and self-organization map (SOM). The criminal written judgments taken from the written judgment retrieving system of the Judicial Yuan in Taiwan are applied to our study. After keyword extraction, frequency and TF-IDF calculation, the top-100 keywords of our data set with the highest frequencies represent the given written judgments. Furthermore, we propose four models for classifying the written judgments. According to section 5.1, Model 2 that uses all segments of the written judgment and the weighted vectors as the input of BPN has a highest precision to classify the training samples. In addition, we demonstrate Kohonen's self-organization map to organize the training written judgments onto two-dimension map, and find the resulting map is very meaningful and assists in searching potentially relevant documents. Finally, the search model which combines the advantages of classification and clustering method determines the likely similar written judgments to recommend and effectively narrows the searching ranges for law experts under time constraints. The extensions of this study will investigate how many keywords should be selected and how much the weights of the weighted keywords should be given without a lot of time-consuming calculation to obtain a highest classification precision. We will also investigate how many numbers of neurons in SOM should be determined such that a better clustering can be achieved. In addition, more classification and clustering methods will be compared the performances with BPN and SOM, and it will enhance the research value in our system. Our future research will combine the neural network technique with other approaches such as Genetic Algorithm (GA) and other ontology-based feature extraction methods to improve the availability and accuracy of the current approach.

References

1. Antonie, M.-L., Zaiane, O.R.: Text document categorization by term association. In: Proceedings of IEEE international conference on data mining, pp. 19–26 (2002)
2. Conrad, J., Al-Kofahi, K., Zhao, Y., Karypis, G.: Effective Document Clustering for Large Heterogeneous Law Firm Collections. In: 10th International Conference on Artificial Intelligence and Law (ICAIL), pp. 177–187 (2005)
3. Horng, J.T., Yeh, C.C.: Applying genetic algorithms to query optimization in document retrieval. Information Processing & Management 36, 737–759 (2000)
4. Kohonen, T.: Self-organizing maps. Springer, Heidelberg (1997)
5. Lovins, J.B.: Development of a stemming algorithm. Mechanical Translation and Computational Linguistics 11, 22–31 (1968)
6. Moens, M.F.: Innovative techniques for legal text retrieval. Artificial Intelligence and Law 9, 29–57 (2001)
7. Raghavan, V.V., Wong, S.K.M.: A critical analysis of vector space model for information retrieval. Journal of the American Society for Information Science 37(5), 279–287 (1986)
8. Rumelhart, D., Hinton, G., Williams, R.: Learning internal representations by error propagation. In: Parallel distributed processing, vol. 1, MIT Press, Cambridge (1986)
9. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing & Management 24(5), 513–523 (1988)
10. Salton, G.: Recent Studies in Automatic Text Analysis and Document Retrieval. Journal of the ACM 20(2), 258–278 (1973)
11. Schweighofer, E.: Legal Knowledge Representation, Automatic Text Analysis in Public International and European Law. In: Kluwer Law International, Law and Electronic Commerce, The Hague, vol. 7 (1999)
12. Selamat, A., Omatu, S.: Web page feature selection and classification using neural networks. Information Sciences 158, 69–88 (2004)
13. Tam, V., Santoso, A., Setiono, R.: A comparative study of centroid-based, neighborhood-based and statistical approaches for effective document categorization. In: Proceedings of the 16th international conference on pattern recognition, vol. 4, pp. 235–238 (2002)

Exploration of Document Relation Quality with Consideration of Term Representation Basis, Term Weighting and Association Measure

Nichnan Kittiphattanabawon, Thanaruk Theeramunkong, and Ekawit Nantajeewarawat

Sirindhorn International Institute of Technology, Thammasat University, Thailand
knichcha@wu.ac.th, {thanaruk,ekawit}@siit.tu.ac.th

Abstract. Tracking and relating news articles from several sources can play against misinformation from deceptive news stories since single source can not judge whether the information is a truth or not. Preventing misinformation in a computer system is an interesting research in intelligence and security informatics. For this task, association rule mining has been recently applied due to its performance and scalability. This paper presents an exploration on how term representation basis, term weighting and association measure affect the quality of relations discovered among news articles from several sources. Twenty four combinations initiated by two term representation bases, four term weightings, and three association measures are explored with their results compared to human judgement. A number of evaluations are conducted to compare each combination's performance to the others' with regard to top-k ranks. The experimental results indicate that a combination of bigram (BG), term frequency with inverse document frequency (TFIDF) and confidence (CONF), as well as a combination of BG, TFIDF and conviction (CONV), achieves the best performance to find the related documents by placing them in upper ranks with 0.41% rank-order mismatch on top-50 mined relations. However, a combination of unigram (UG), TFIDF and lift (LIFT) performs the best by locating irrelevant relations in lower ranks (top-1100) with rank-order mismatch of 9.63 %.

Keywords: Document Relation, Association Rule Mining, News Relation.

1 Introduction

The explosion of the Internet makes it easier for broadcasting news to a large volume of readers. Most of the readers prefer to read news stories from several publishers in order to avoid bias from a single source of information. News portals are the most popular alternatives for the readers because several news portals facilitate news content access by providing linkages among news articles on the web, releasing readers from directly browsing through news publishers' home pages. Reading news stories which come from several sources from the

H. Chen et al. (Eds.): PAISI 2010, LNCS 6122, pp. 126–139, 2010.
© Springer-Verlag Berlin Heidelberg 2010

services of news portals can prevent the readers from false information since misinformation can make the readers in changing their perception as an attack directed at the mind of the readers of a computer system. Such an attack was defined as cognitive hacking in intelligence and security informatics research [1]. For the success of systems for news provider services, an appropriate organization of news contents is a major requirement. News portals usually organize news into some kinds of relationship structures, e.g., group news articles by category, by recency, or by popularity, summarize news contents, and create relations between news articles. Currently most of these functions require manual arrangement processes. Towards automated content organization, while classification techniques can be applied to assign a category label to each document based on a number of criteria, such as text genre, text style, and users' interest [2,3,4]. Some of them can be adopted for classifying news articles [5,6]. By the classification method, it requires users to provide a number of predefined classes and a large number of training examples. However, the classification approach requires users to provide a number of predefined classes and a large number of training examples. Releasing from these requirements, clustering can be used to group documents according to their similar characteristics [7,8]. As a more complicated application, a multidocument summarization can be performed to obtain a shorter description from a cluster of news describing similar events [9]. For the past several years, event-based topics of news stories has been investigated by Topic Detection and Tracking (TDT) research [10,11]. Event clustering and first story detection are two main problems in TDT. Normally, by the way of event clustering, news stories that include several events can be grouped into a number of clusters, each of which is about a single news topic. On the other hand, the task of first story detection is to identify whether a news story includes new events which are never seen. Recently, an association rule mining approach [12] has been applied for discovering document relations in scientific research publications due to its performance and scalability [13]. In Thai language, even there have been several works towards extraction of information on online document, most of them still have limitation in finding document relations. As an early work on relation discovery in multiple Thai documents, Kittiphattanabawon and Theeramunkong [14] have proposed a method based on association rule mining to find the relations among Thai news documents. The work gave a preliminary exploration on the performance of support-confidence and support-conviction frameworks under limited environment of top-k ranking evaluation.

In this paper, besides support-confidence and support-conviction, a support-lift framework is investigated and compared with human judgment in a more general environment of up-to top-1100 ranking evaluation. Towards optimal settings, twenty four combinations generated from two term representation bases, four term weightings and three association measures are examined to find suitable combinations for discovering meaningful relations among news articles. In Sect.2, news relation generation is described under the formation of association rules. The factors for discovering news associations are then presented in Sect.3. The generalized association measures are also defined in this section. Section 4 presents evaluation methods including a description of types of news relations, a construction of evaluation dataset and criteria for evaluation. A number of experimental results and discussion are given in Sect. 5. Finally, a conclusion and future works are made in Sect.6.

2 Association Rule Mining for Discovering Relations among News Articles

Association rule mining (ARM) is well-known as a process to find frequent patterns in the form of rules from a database. Recently ARM or its derivatives has been applied in finding relations among documents [13,14]. By encoding documents as items, and terms in the documents as transactions, we mine a set of frequent patterns, each of which is in the form of a set of documents sharing common terms more than a threshold, called support. Thereafter, as a further step, a set of frequent rules can be found based on these frequent patterns with another threshold, namely confidence. In this work, in order to work with non-binary data, we adopt the generalized support and generalized confidence in [13], and the generalized conviction in [14] as association measures. A formulation of the ARM task on news article relation discovery can be summarized as follows. Assume that $I = \{i_1, i_2, ..., i_m\}$ is a set of m news articles (items), T=$\{t_1, t_2, ..., t_n\}$ is a set of n terms (transactions), a news itemset X=$\{x_1, x_2, ..., x_k\}$ is a set of k news articles, and a news itemset Y=$\{y_1, y_2, ..., y_l\}$ is a set of l news articles. As an alternative to confidence and conviction, a measure called lift is introduced in this work. Conventionally, the lift of an association rule $X \rightarrow Y$ is defined as $conf(X \rightarrow Y)/supp(Y)$, where $conf(X \rightarrow Y)$ is the confidence value of the rule $X \rightarrow Y$ and $supp(Y)$ is the support value of Y. The generalized support of X ($sup(X)$), the generalized confidence of $X \rightarrow Y$ ($conf(X \rightarrow Y)$), the generalized conviction of $X \rightarrow Y$ ($conv(X \rightarrow Y)$), and the generalized lift of $X \rightarrow Y$ ($lift(X \rightarrow Y)$) are shown in Table1, where $w(i_a, t_b)$ represents a weight of a term t_b in a news articles i_a and $Z = \{z_1, z_2, ..., z_{k+l}\} \subset I$ with $k + l$ news articles since they are the co-occurrence of terms in both k news articles in the X and l news articles in the Y. By this method, the discovered relations are in the form of "$X \rightarrow Y$", where X as well as Y is a set of news articles. This rule represents that the content overlap among the news articles in the X has a relationship with the content overlap among the news articles in the Y. As a special case of one single antecedent and one single consequent, the rule can be interpreted that the news article in the X relates to the news article in the Y. Among efficient algorithms such as Apriori [15], CHARM [16,17] and FP-Tree [18], in this work we select FP-Tree since it is the most efficient mining algorithm that can generate conventional frequent itemsets, not closed frequent itemsets.

Table 1. Definitions of association measures: (a) generalized support, (b) generalized confidence, (c) generalized conviction, and (d) generalized lift

(a) $sup(X) = \frac{\sum_{b=1}^{n} min_{a=1}^{k} w(x_a,t_b)}{\sum_{b=1}^{n} max_{a=1}^{m} w(i_a,t_b)}$	(b) $conf(X \rightarrow Y) = \frac{\sum_{b=1}^{n} min_{a=1}^{k+l} w(z_a,t_b)}{\sum_{b=1}^{n} min_{a=1}^{k} w(x_a,t_b)}$
(c) $conv(X \rightarrow Y) = \frac{1-\frac{\sum_{b=1}^{n} min_{a=1}^{l} w(y_a,t_b)}{\sum_{b=1}^{n} max_{a=1}^{m} w(i_a,t_b)}}{1-\frac{\sum_{b=1}^{n} min_{a=1}^{k+l} w(z_a,t_b)}{\sum_{b=1}^{n} min_{a=1}^{k} w(x_a,t_b)}}$	(d) $lift(X \rightarrow Y) = \frac{\frac{\sum_{b=1}^{n} min_{a=1}^{k+l} w(z_a,t_b)}{\sum_{b=1}^{n} min_{a=1}^{k} w(x_a,t_b)}}{\frac{\sum_{b=1}^{n} min_{a=1}^{l} w(y_a,t_b)}{\sum_{b=1}^{n} max_{a=1}^{m} w(i_a,t_b)}}$

3 Term Representation Basis, Term Weighting and Association Measure

In general, the results from the mining process can differ according to setting factors in the process. In this paper, to find an appropriate environment in discovering the news relations, we explore three main factors, (1) term representation basis, (2) term weighting, and (3) association measure. For the term representation basis, unigram (UG) or bigram (BG) are investigated as the term representation for the content of news documents. Intuitively, UG may be not sufficient for representing the content of a news document since there exists term ambiguity in the context. As an alternative, BG considers two neighboring terms as a unit in order to to handle compound words and then partially solve the ambiguity of words. For term weighting, binary term frequency weighting (BF), term frequency weighting (TF) and their modification with inverse document frequency weighting (BFIDF, TFIDF) are explored. BF simply indicates the existence or non-existence of a term in a news document while TF indicates the frequency of a term in the document. IDF is often used in complementary with TF, to promote a rare term which occurs in very few documents, as an important word. Although it can be calculated as the total number of documents in the collection (N) divided by the number of documents containing the term (DF), it is usually used in the logarithm scale. BFIDF and TFIDF of the i-th terms are defined as $BF_i \times \log(N/DF_i)$ and $TF_i \times \log(N/DF_i)$ respectively. To measure the appropriateness of relations, quantitative measure is another factor, which needs to be carefully selected. In this work, to find a suitable measure, we consider confidence (CONF), conviction (CONV) and lift (LIFT) as association measures. For CONF, it is a well-known rule measure for ARM approach. As for CONV and LIFT, they can result in more interesting relations [19,20], CONV and LIFT are investigated to improve the association of news articles in our work, as shown in the previous section.

4 Evaluation Methodology

In this section, we describe an evaluation methodology to investigate the potential combinatorial factors to make a judgement on types of news relations.

4.1 Types of News Relations

Most tasks about document relations judged stories to be either relevant or non-relevant, that is, two classes ("yes" and "no") were considered. In this work, three main types of news relations are classified based on the relevance of news events: (1) "completely related" (CR), (2) "somehow related" (SH) and (3) "unrelated" (UR) [14]. The CR relation is detected when two new articles are about an exactly same event. Such a CR relation is always found because every news reporter tries to report daily important events. The same event, therefore, is often published by many publishers in the same time. However, the CR relation may be presented in either different headlines or different writing styles. As a result, this type is often retrieved among different news publishers whose publishing times are the same or quite close to the same. For the SH relation, it is a kind of relation which has only somewhat closely related. The events in both news articles may have similar topics. connect together forming a sequential time series of events. or contain same contents in some parts. However, the contents in news documents whose relationship is SH type are not mentioned exactly the same story. The relation of UR is defined as a relationship of having absolutely nothing related between news articles. In other words, It could be considered as a non-relevant story.

4.2 Evaluation Dataset

As there is no standard dataset for news relations in Thai available as a benchmark for assessing performance of our approach, we construct our own dataset based on an evaluation of human. The dataset is selected approximately 1,100 news relations mined from 811 Thai news articles of three news online sources (Dailynews (313 articles), Komchadluek (207 articles), and Manager online (291 articles)) [1] during August 14-31, 2007, consisting of three categories (politics, economics, and crime). Each set is comprised of the news relations with their relation types defined in previous section. For the evaluation of human, three assessors who admire reading news provide their judgments on predefined relation types (CR, SH, and UR). After they have been instructed for making a decision about how related two news articles should be identified, they make their decision by comparing the contents in both news articles and then assign only one type for their relation. Note that, every news relation is judged by all three assessors. If there are different opinions on the association, the best judgment is given by voting. However, voting may not be able to guarantee a majority for such an agreement. To decide the answer, an iteration process is performed by asking the assessors to repeat their considerations until the final decision is made. To this end, our dataset contains records of news relations along with their relation types determined by human judgment (65 relations of CR, 571 relations of SH, and 496 relations of UR). Details for evaluation of these types will be described in next section.

[1] `www.dailynews.co.th, www.komchadluek.com, and www.manager.co.th`

4.3 Evaluation Criterion

The quality of twenty four combinations in discovering news relations is evaluated by comparing the results generated by each of them to those from human judgments. The evaluation method is applied from a paired-wise comparison technique [21] since the paired-wise comparison has been applicably used for counting the mismatches between rankings. For each combination, an evaluation is proceeded by creating a ranking list of relations ordered by its association measure, mapping the resultants from the human judgments to each of these relations in the list, calculating a mismatch score between both of them, and comparing the quality among twenty four combinations with a criterion, so-called rank-order mismatch (ROM) shown in Eq. (1). The ROM value is a calculation of dividing a mismatch score ($M(A, B)$) with the mismatch score of the worst case which all news relations in one method (A) are arranged in the reverse order compared to the other method (B). In addition, the ROM value expresses, according to the assessment of assessors, the mismatches of relation types of our method to the human suggestions. Note that the ROM value is a value in $\{0, 100\}$. If all news relations are found corresponding to the human suggestions, ROM value is equal to 0. $M(A, B)$, the mismatch score, indicates the number of rank mismatches between two methods, say A and B, which rank a set of N objects, as shown in Eq. (2), where $r_A(k)$ and $r_B(k)$ are the respective rank of the k-th objects based on method A and B respectively. The mismatch score expresses the importance of method A whether corresponding to method B or not. As for our work, since A is the machine ranking method and B is the human ranking method, the ROM(A, B) is denoted by ROM$_h(A)$. This equation implies, therefore, relation type mismatches between our method and the evaluation method from human.

$$ROM(A, B) = \frac{2 \times M(A, B)}{N(N-1)} \times 100 \tag{1}$$

$$M(A, B) = \sum_{i=1}^{N} \sum_{j=i+1}^{N} |\delta(r_A(i), r_A(j)) - \delta(r_B(i), r_B(j))| \tag{2}$$

A mismatch function, $\delta(a, b)$, returns 1 when a less than b, otherwise 0. Such a function indicates that a relation in an upper rank (a) which has a score lower than one in a lower rank (b) presents in a mismatch order. As stated above, the constructed rank order is arranged by the association measure. It is important to recognize that CONF and CONV are directional while LIFT is not. The direction of rules obtained by LIFT is not taken into account, i.e., $lift(X \rightarrow Y)$ is equal to $lift(Y \rightarrow X)$ but $conf(X \rightarrow Y)$ is not equal to $conf(Y \rightarrow X)$, and also $conv(X \rightarrow Y)$ is different from $conv(Y \rightarrow X)$. Through our work with three types of news relations, we do not account for the direction of the rules because it does not perceive meaningful differences on the types. CONF and

CONF will be treated to be undirectional by $min()$ function, as presented in following equations.

$$conf(X,Y) = min(conf(X \rightarrow Y), conf(Y \rightarrow X)) \quad (3)$$

$$conv(X,Y) = min(conv(X \rightarrow Y), conv(Y \rightarrow X)) \quad (4)$$

The reason why we use the $min()$ function is that, the smaller value is make sense to the human judgments because the assessors disregard the direction of news relations. For example, in an evident of vastly different occurrence frequencies between two news articles, if the relation $news1 \rightarrow news2$ has very high confidence of 90% and $news2 \rightarrow news1$ has very low confidence of 10%, the judgments of assessors will be made on the UR relation rather than the CR relation because the contents of both news articles are definitely dissimilar.

5 Experiments

5.1 Experimental Setting

To examine how three factors affect the quality of discovered news relations, three experiments are performed using our evaluation dataset. In the first experiment, the effect of each single factor on the relation quality is focused. This experiment includes three comparative studies (UG vs. BG, BF vs. BFIDF vs. TF vs. TFIDF, and CONF vs. CONV vs. LIFT). Here, any pair of possible alternatives for each factor is compared by calculating the difference of their ROM values, i.e., subtraction of the ROM value (Eq. (1)) of an alternative with that of the other alternative. If the ROM value of the method A, $\text{ROM}_h(A)$, is higher than that of the method B, $\text{ROM}_h(B)$, the ROM difference between A and B becomes positive, that is, the method A has more mismatches than the method B. In other words, the method B provides more similar results to human answers. In the second experiment, for each of twenty four methods, we visualize the ratio of CR, SH, and UR relations for each association measure with respect to top-k intervals in order to investigate whether CR relations can be located at higher ranks followed by SH, and UR can be placed at lower ranks, or not. The third experiment targets the exploration of the detailed performance (ROM values) of the combinations which perform very well in top-k ranks.

5.2 Experimental Results

Paired Comparative Studies

Term Representation Bases. Figure 1 (a)–(c) show the ROM differences between UG and BG ($\text{ROM}_h(UG)$-$\text{ROM}_h(BG)$) for CONF, CONV, and LIFT respectively. In the figures, the bar graphs are plotted with respect to top-k ranks. As one observation, almost top-k cases give the number of more positive values than negative values, except cases of LIFT (Fig. 1 (c)), after the top-500 mined relations. The results show that BG outperforms UG in almost cases except the cases of LIFT under top-500 ranking. These results suggest that using either BG with CONF or BG with CONV is effective in all ranks, and using BG with LIFT is effective in upper rank (<500). For lower rank (>500), applying UG with LIFT appears to be effective.

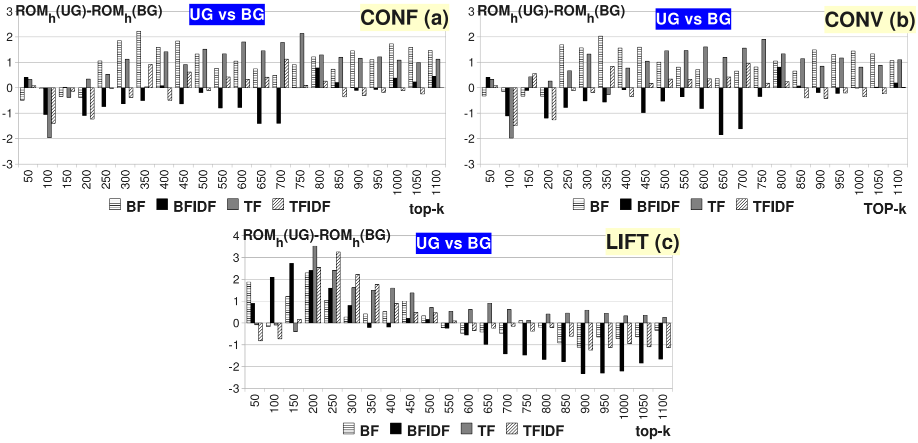


Fig. 1. ROM differences between term representation bases: UG vs. BG in the cases of CONF (a), CONV (b), and LIFT (c)

Term Weightings. Similar to previous experiment, except that BF, TF, BFIDF, and TFIDF are taken into account instead of UG and BG. We can see that, from Fig. 2 (a) and (d), in almost cases, BFIDF gives lower ROM values than BF and TF respectively, which BF presents higher ROM values than TF, as shown in Fig. 2 (b). However, in Fig. 2 (e), BFIDF outputs higher ROM values than TFIDF in almost cases. Therefore, TFIDF also outperforms BF and TF, as seen in Fig. 2 (c) and (f), in almost cases. These results suggest that TFIDF appears to be the most effective, since TFIDF gives the lowest ROM values.

Association Measures. Like previous experiments, the comparisons between association measures are depicted by ROM differences, as shown in Fig. 3 (a)–(c). Figure 3 (a) shows that most of bars are on the positive side of the graph. They indicate that CONV produces lower ROM values than CONF. In both Fig. 3 (b) and (c), the bar graphs appear to be characterized into two groups, upper ranks (< 300) and lower ranks (> 300). With upper ranks, CONF and CONV present higher ROM values than LIFT, while LIFT outputs lower ROM values in lower ranks. However, CONV outperforms CONF as reported above. These suggest that using CONV is effective for the relations placed in upper ranks, but using LIFT is more effective when applied to the relations in lower ranks. For more details, in next experiment, we will investigate types of relations in each rank to determine how effective different combinations produce different relation types.

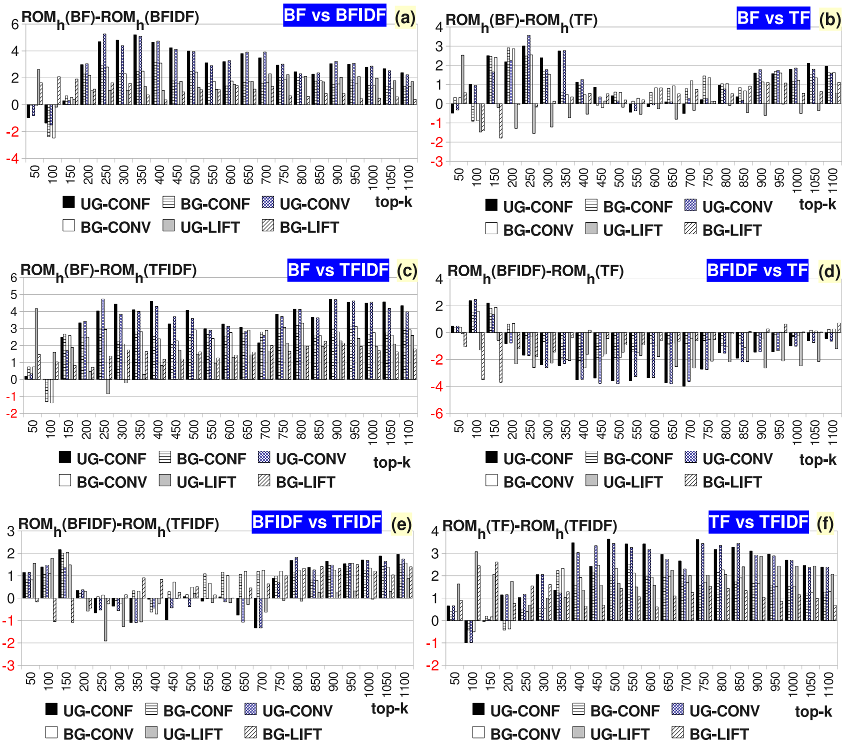


Fig. 2. ROM differences between term weightings: (a) BF vs. BFIDF, (b) BF vs. TF, (c) BF vs. TFIDF, (d) BFIDF vs. TF, (e) BFIDF vs. TFIDF, and (f) TF vs. TFIDF

Rank Analysis on News Relation Types. The analysis on the number of three relation types under different twenty four combinations and different top-k rank intervals is investigated separately by three association measures (CONF, CONV, and LIFT), as shown in Fig. 4 (a)–(c). Trends of these three graphs can be discussed into three different groups. Such three groups correspond to three types of relations (CR, SH, and UR). In the graphs, the CR and UR relations are represented by the curves with triangle and square symbols respectively. For SH relations, they are plotted by solid lines. Detailed descriptions of them are given here: (1) CR relations are located at upper ranks (say 50-100), (2) SH relations are located next to CR at middle ranks (say 100-350), and (3) UR relations are positioned at lower ranks (say $>$ 350). Obtained results can be stated that, in upper ranks ,the relations are judged to either CR or SH without UR relations while no CR relations are available in lower ranks. When observing gaps between the CR lines and the SH lines in the leftmost position (1-50 interval), the gaps, in the cases of CONF and CONV (Fig. 4 (a) and (b)), are quite large. On the contrary, the cases of LIFT in Fig. 4 (c) trigger smaller gaps between these two types. For the rightmost location (1051-1100 interval), CONF and CONV produce narrow gaps while LIFT causes big gaps between

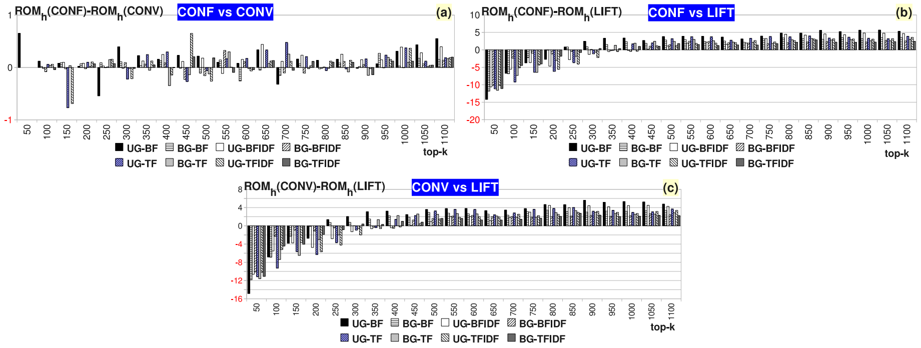


Fig. 3. ROM differences between association measures: (a) CONF vs. CONV, (b) CONF vs. LIFT, and (c) CONV vs. LIFT

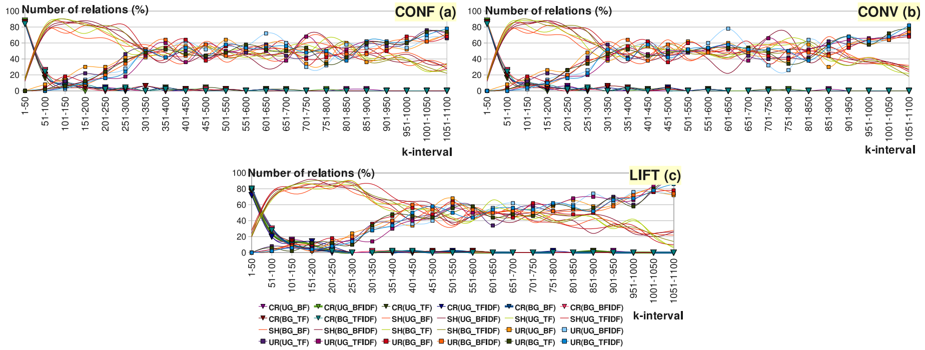


Fig. 4. Percentages of the number of relations in the cases of CONF (a), CONV (b), and LIFT (c)

UR and SH relations. We can point out that CONF and CONV are useful for predicting the CR relations, whereas LIFT is fine for separating the SH types from the UR types. The reason is that, for CONF and CONV, the number of SH relations in the leftmost side are significantly different from those of CR relations and are slightly different from those of UR relations in the rightmost side, and vice versa for LIFT.

Analysis on Combinations of Three Factors. In the final experiment, the detailed performance on the combinations of three factors is examined. Table 2 shows ROM values of twenty four methods in each top-k rank in order to observe the best combination for using in discovering news relations. For the top-50 mined relations, the combination of BG, TFIDF, and CONF (BG-TFIDF-CONF), as well as BG, TFIDF, and CONV (BG-TFIDF-CONV), appears to be the most effective since it has the lowest ROM values (0.41%). The method with UG, TFIDF, and LIFT (UG-TFIDF-LIFT) performs the best on the top-1100 rankings by giving the lowest ROM values (9.63%). Such ROM values involve

Table 2. ROM values (ROM_h) of twenty four combinations of three factors

Combinations	top-*k*											
	50	100	200	300	400	500	600	700	800	900	1000	1100
UG-BF-CONF	0.65	4.08	7.97	10.35	11.41	12.79	13.47	14.27	16.87	17.98	18.15	17.55
UG-TF-CONF	1.14	3.07	5.78	7.95	10.29	12.37	13.63	14.78	15.91	16.38	16.36	15.60
UG-BFIDF-CONF	1.63	5.45	4.98	**5.54**	6.77	8.80	10.26	10.79	14.42	14.92	15.36	15.17
UG-TFIDF-CONF	0.49	4.06	4.64	5.90	6.82	8.73	10.21	12.12	12.74	13.27	13.66	13.21
UG-BF-CONV	0.82	3.96	7.95	9.95	11.25	12.57	13.39	14.59	16.74	17.99	17.84	17.00
UG-TF-CONV	1.14	**3.01**	5.68	8.17	9.99	12.43	13.46	14.30	15.97	16.22	15.98	15.41
UG-BFIDF-CONV	1.63	5.47	4.90	5.56	6.52	8.62	10.11	10.67	14.44	14.77	14.97	14.77
UG-TFIDF-CONV	0.49	4.00	**4.53**	6.12	6.96	8.99	10.27	12.00	12.62	13.30	13.29	13.02
UG-BF-LIFT	14.86	10.81	10.69	7.90	8.02	8.97	9.58	11.12	12.06	12.38	12.51	12.22
UG-TF-LIFT	12.33	12.28	11.97	9.11	8.56	9.15	9.85	11.46	12.14	12.99	13.01	11.70
UG-BFIDF-LIFT	12.24	10.99	9.65	6.85	6.95	7.69	**8.07**	**8.83**	**9.96**	10.35	**10.53**	10.50
UG-TFIDF-LIFT	10.69	9.21	10.23	8.11	7.21	7.49	8.28	9.45	10.09	**10.12**	10.58	**9.63**
BG-BF-CONF	1.14	4.12	8.35	8.49	9.82	11.47	12.42	13.79	15.66	16.52	16.42	16.09
BG-TF-CONF	0.82	5.03	5.44	6.83	8.88	10.86	11.83	13.00	14.62	15.22	15.27	14.47
BG-BFIDF-CONF	1.22	6.51	6.08	6.17	6.69	8.99	11.04	12.19	13.64	15.02	14.98	14.72
BG-TFIDF-CONF	**0.41**	5.45	5.87	6.28	7.31	8.84	9.88	11.00	12.48	13.58	13.76	13.21
BG-BF-CONV	1.14	4.10	8.28	8.39	9.70	11.57	12.68	13.94	15.69	16.51	16.40	15.93
BG-TF-CONV	0.82	4.99	5.42	6.85	9.22	10.98	11.85	12.75	14.64	15.37	15.17	14.31
BG-BFIDF-CONV	1.22	6.59	6.10	6.08	6.60	9.15	10.93	12.29	13.63	14.97	14.99	14.58
BG-TFIDF-CONV	**0.41**	5.49	5.80	6.30	7.31	8.65	9.92	11.05	12.38	13.71	13.64	13.01
BG-BF-LIFT	12.98	10.97	8.40	7.63	7.50	8.64	10.06	11.60	12.25	13.50	13.22	12.56
BG-TF-LIFT	12.41	12.38	8.45	7.50	6.96	8.45	9.23	10.85	11.73	12.40	12.68	11.45
BG-BFIDF-LIFT	11.35	8.89	7.24	6.05	7.15	7.53	8.63	10.24	11.63	12.68	12.74	12.16
BG-TFIDF-LIFT	11.51	9.94	7.69	5.90	**6.32**	**7.03**	8.63	9.61	10.30	11.36	11.53	10.77

the relation types placing in each rank as investigated in previous experiment (Fig. 4). From Fig. 4, mostly, the types of news relations on top-50 ranks are CRs, while those on top-1100 rankings are URs. The results obtained heretofore can be separated into two cases, (1) the upper ranks, and (2) the lower ranks. In the case of the upper ranks, the best combination is the combination of either BG-TFIDF-CONF or BG-TFIDF-CONV. Another case is the lower ranks that the UG-TFIDF-LIFT combination performs better. The results are in consensus with the three studies of the first experiment. For almost cases, in Fig. 1, BG combined with either CONV or CONF gives better effectiveness than UG, while UG combined with LIFT is more effective when working in the lower rank. Moreover, the results in Fig. 2 suggest that TFIDF is the most effective term weighting. However, CONF and CONV appear to be candidate measures for our method. Possible reasons why these three combinations are the best are given as following. First, BG works well on the related relations at higher ranks because BG can effectively solve the context ambiguity. At lower ranks, the relations are not rather related, therefore the terms counted by the pattern of BG are rarer than those checked by the form of UG. Second, for the case that a news relation (say $news1 \rightarrow news2$) whose document sizes are very different (assume that document size of $news2$ is greater than document size of $news1$), when applying LIFT measure to the relation which is CR, the strength of the relation will be affected by $news2$, due to computing with the equation of LIFT ($lift(X \rightarrow Y) = P(X \cap Y)/P(X)P(Y)$). Thus, LIFT may push the CR relation to the lower ranks which are inappropriate order, but CONF and CONV do not effect them. Consequently, using CONF or CONV for finding semantic relations is more effective than using LIFT. Conversely, when using LIFT measure on the relation which is UR, because the LIFT equation also uses the weight of *news*2 for measuring its strength, *news*2 has great influence in distinguishing unrelated relation from related relation. This is why the UR relations are located to appropriate orders which are the lower ranks. As a result, applying LIFT on the lower ranks appears to be effectiveness.

5.3 Discussion

Once the experimental results obtained by our methods are established, the discussion and error analysis are then given in this section. Our proposed method still misclassified some instances. Some relations assigned by human as CRs are identified by low value of association measures. Causes of this error can be concluded as two cases corresponding to the document size. The first case is when size of two related news documents (*news*1 and *news*2) is very different. In this case, we found out that a news article (*news*1) is a summary of another news article (*news*2), agreeing with the document size ratio which shows that two news articles have very different size. Therefore, the measure value is low because the content overlap between two news articles is not frequent, This leads the proposed method to select the SH relation instead of the CR relation. We plan to overcome this problem in our future works by assigning more weight to headlines of news articles since they comprise a number of the same words. Furthermore, the publishing times of news articles will be considered because the CR relations are rather published in the same time or quite close to the same time. The second case why the relations assigned by human as CRs are identified by low value of association measures is when the document size ratio is not quite different. The main errors of this case are listed as follows. First, although the relations are completely related, two news articles can be written in the different styles due to different publishers who may use various words for representing same meaning words. A synonym detection approach can help for correcting this error. Second, different publishers deliver dissimilar contents for the exactly same event because they may write a news story with either unequal facts or contrasting details. For unequal facts, we plan to give the importance on the news headlines and news contents positioned at the first paragraph of news documents, by weighting them more heavily since the publishers rather bring out the similar contents on these positions. With contrasting details, their different data should be removed before using the proposed method. We plan to project a news difference problem as another research topic. Much more interesting in our discovery is the support-lift framework in performing well in the lower ranks. This is our further study on future works for analyzing on a hybrid method since it considers the number of qualitative criteria.

6 Conclusions

In this paper, we have investigated the effects of two term representation bases, four term weighting, and three association measure to discover relations among news articles. Totally twenty four combinations are explored. To evaluate the quality of discovered news associations, top-k ranked relations are analyzed by rank-order mismatch. By comparing the results to the human judgements rendered by the vote of three assessors, the experimental results show that the ROM values under the combination of BG-TFIDF-CONF, as well as BG-TFIDF-CONV, is suitable to achieve finding related relations with 0.41% ROM on top-50 mined relations while UG-TFIDF-LIFT performs well, up to top-1100, with ROM of 9.63%. These results suggest that, for relevant relations like completely related relations, the combination of term frequency with inverse document frequency together with bigram, and either confidence or conviction is applicable, and that together with unigram and lift works well in distinguishing unrelated relations from related relations. For somehow relations, they are still gray areas which need further analysis. As future works, we plan to work on various features by considering more factors and more criteria on each factor for enhancing the quality of document relations. Furthermore, the direction of news relations and deeper analysis on the characteristics of relations will be examined.

Acknowledgements

This work was granted by Strategic Scholarships Fellowships Frontier Research Networks 2006 from the Commission on Higher Education and Thailand Research Fund under project number BRG50800013.

References

1. Thompson, P., Cybenko, G., Giani, A.: Cognitive Hacking, ch. 19. Book of Economics of Information Security, pp. 255–287. Springer, US (2004)
2. Ferizis, G., Bailey, P.: Towards practical genre classification of web documents. In: Proc. 15th international conference on World Wide Web, pp. 1013–1014. ACM, New York (2006)
3. Gamon, M.: Linguistic correlates of style: authorship classification with deep linguistic analysis features. In: Proc. Coling 2004, Geneva, Switzerland, COLING, August 23-27, pp. 611–617 (2004)
4. Carreira, R., Crato, J.M., Gonçalves, D., Jorge, J.A.: Evaluating adaptive user profiles for news classification. In: Proc. 9th international conference on Intelligent user interfaces, pp. 206–212. ACM, New York (2004)
5. Antonellis, I., Bouras, C., Poulopoulos, V.: Personalized news categorization through scalable text classification. In: Zhou, X., Li, J., Shen, H.T., Kitsuregawa, M., Zhang, Y. (eds.) APWeb 2006. LNCS, vol. 3841, pp. 391–401. Springer, Heidelberg (2006)
6. Mengle, S., Goharian, N., Platt, A.: Discovering relationships among categories using misclassification information. In: Proc. 2008 ACM symposium on Applied computing, pp. 932–937. ACM, New York (2008)
7. Zhang, N., Watanabe, T., Matsuzaki, D., Koga, H.: A novel document analysis method using compressibility vector. In: Proc. the First International Symposium on Data, Privacy, and E-Commerce, November 2007, pp. 38–40 (2007)

8. Weixin, T., Fuxi, Z.: Text document clustering based on the modifying relations. In: Proc. 2008 International Conf. on Computer Science and Software Engineering, December 2008, vol. 1, pp. 256–259 (2008)
9. Lin, F., Liang, C.: Storyline-based summarization for news topic retrospection. Decision Support Systems 45(3), 473–490 (2008)
10. Allan, J., Carbonell, J., Doddington, G., Yamron, J., Yang, Y.: Topic detection and tracking pilot study final report. In: Proc. the DARPA Broadcast News Transcription and Understanding Workshop, pp. 194–218 (1998)
11. Papka, R., Allan, J.: Topic Detection and Tracking: Event Clustering as a Basis for First Story Detection, ch. 4. Book of Advances Information Retrieval: Recent Research from the CIIR, pp. 96–126. Kluwer Academic Publishers, Dordrecht (2006)
12. Kotsiantis, S., Kanellopoulos, D.: Association rules mining: A recent overview. International Transactions on Computer Science and Engineering 32(1), 71–82 (2006)
13. Sriphaew, K., Theeramunkong, T.: Quality evaluation for document relation discovery using citation information. IEICE Trans. Inf. Syst. E90-D(8), 1225–1234 (2007)
14. Kittiphattanabawon, N., Theeramunkong, T.: Relation discovery from thai news articles using association rule mining. In: Chen, H., Yang, C.C., Chau, M., Li, S.-H. (eds.) PAISI 2009. LNCS, vol. 5477, pp. 118–129. Springer, Heidelberg (2009)
15. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: Proc. the 20th International Conf. on Very Large Data Bases, San Francisco, CA, USA, pp. 487–499. Morgan Kaufmann Publishers Inc., San Francisco (1994)
16. Zaki, M.J., Hsiao, C.J.: Charm: An efficient algorithm for closed association rule mining. Technical report, Computer Science, Rensselaer Polytechnic Institute (1999)
17. Zaki, M.J., Hsiao, C.J.: Efficient algorithms for mining closed itemsets and their lattice structure. IEEE Trans. on Knowl. and Data Eng. 17(4), 462–478 (2005)
18. Han, J., Pei, J., Yin, Y., Mao, R.: Mining frequent patterns without candidate generation: A frequent-pattern tree approach. Data Min. Knowl. Discov. 8(1), 53–87 (2004)
19. Lallich, S., Teytaud, O., Prudhomme, E.: Association rule interestingness: Measure and statistical validation. In: Quality Measures in Data Mining. Studies in Computational Intelligence, vol. 43, pp. 251–275. Springer, Heidelberg (2007)
20. Azevedo, P.J., Jorge, A.M.: Comparing rule measures for predictive association rules. In: Kok, J.N., Koronacki, J., Lopez de Mantaras, R., Matwin, S., Mladenič, D., Skowron, A. (eds.) ECML 2007. LNCS (LNAI), vol. 4701, pp. 510–517. Springer, Heidelberg (2007)
21. David, H.: The Method of Paired Comparisons. Oxford University Press, Oxford (1988)

Identifying Controversial Issues and Their Sub-topics in News Articles

Yoonjung Choi, Yuchul Jung, and Sung-Hyon Myaeng

Korea Advanced Institute of Science and Technology
355 Gwahak-ro Yuseong-gu Daejeon, 305-701, South Korea
{choiyj35,enthusia77,myaeng}@kaist.ac.kr

Abstract. We tackle the problem of automatically detecting controversial issues and their subtopics from news articles. We define a *controversial issue* as a concept that invokes conflicting sentiments or views and a *subtopic* as a reason or factor that gives a particular sentiment or view to the issue. Conforming to the definitions, we propose a controversial issue detection method that considers the magnitude of sentiment information and the difference between the amounts of two different polarities. For subtopic identification, candidate phrases are generated and checked for containing five different features, some of which attempts to capture the relationship between the identified issue phrase and the candidate subtopic phrase. Through an experiment and analysis using the MPQA corpus consisting of news articles, we found that the proposed method is promising for both of the tasks although many additional research issues remain to be tapped in the future.

Keywords: Controversial issue detection, Subtopic identification, Sentiment analysis.

1 Introduction

People tend to be naturally interested in popular stories such as controversial issues and social events that invoke sentiment, which are usually found in news articles. Distinguished from opinions that have been a target for analysis in the natural language processing community[1], sentiment is more broadly defined to include one's judgment or evaluation, affective state, or intended emotional communication [Wikipedia].

Sentiment analysis, which determines the sentimental attitude of a speaker or writer, is known to be important for governments, companies, and individuals since governments should listen to public opinions to improve their services, companies need to scrutinize reviews to advance their products, and individuals want to know others' thinking or feeling about interesting subjects such as products, movies, and

[1] The term *sentiment analysis* has been used for the meaning of *opinion analysis* in the literature. In this paper, we define sentiment analysis to be broader and include opinion analysis.

H. Chen et al. (Eds.): PAISI 2010, LNCS 6122, pp. 140–153, 2010.
© Springer-Verlag Berlin Heidelberg 2010

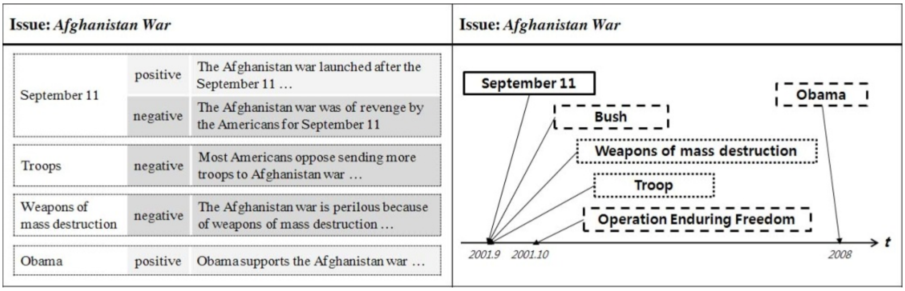
issues [3,4]. As such, many researchers have studied sentiment analysis and opinion analysis using such data as product reviews, blogs, and news articles, obtaining reasonable performances for the tasks of identifying subjective sentences or documents, determining their polarity values, and finding the holder of the sentiment or opinion found in a sentence [3,5,10].

While the past research has focused on the three tasks mentioned above, where sentiment is the focus, the main thrust of this paper is to identify controversial issues (or topics) in news articles and their reasons or subtopics for the controversy conveyed in the issues. That is, we focus on detecting controversial issues and their subtopics by means of sentiment information. We assume that a controversial issue receives sentiment of various sorts (e.g. positive vs. negative feelings, pros vs. cons, or rightness vs. wrongness in their judgments). A related subtopic often times invokes sentiment or serves as the reason why people feel or express particular sentiment. In short, our goal is to identify main topics/issues, which are controversial, and their subtopics by means of sentiment information conveyed in text segments.

Figure 1 shows an example using the topic "*Afghanistan War,*" one of the most controversial issues that appeared in newspapers for a while. Given sentiment-embedding sentences related to the topic, such as "*Obama supports the Afghanistan war*" and "*The Afghanistan war is perilous because of weapons of mass destruction,*" we can identify subtopics such as *"Obama"* and *"weapons of mass destruction"*. By showing the subtopics on a time line as in the Figure 1, we can provide a sentiment-time summary of the subtopics for the issue of *Afghanistan war.*

The notion of topics in relation to sentiment analysis has been explored in the past. Several studies exploited topic relevance to analyze the sentiment [4, 6]. However, there was little explicit effort to find topics because topics were usually given for sentiment analysis. Even if topics were extracted as in [2, 4], they were different from the issues we deal with in this paper. A recent study [1] proposed a topic-sentiment mixture model to extract subtopics and sentiments, but the scope is limited in that they used product reviews. Compared to news articles, product reviews contain explicit opinions with obvious clues (e.g., "like", "good") and related subtopics that are usually pre-defined features given by the manufacturers (e.g. "battery"). Our work is unique in that we attempt to detect sentiment-invoking issues using news articles and then their related subtopics that serve as reasons or focal points for different views.

As a way to detect candidates for controversial issues, we first compute the weight of each noun phrase and verb phrase utilizing query generation methods [9]. We then determine whether the selected candidates are controversial or not by observing how many sentiment clues are included within pertinent sentences. If a candidate is not controversial enough, we choose another one until finding the proper one. For identification of related topics, phrases in the sentences near the identified issue are examined to determine whether each of them belongs to the sub-topic category or not. Features used for this classification function include collocation information between the candidate phrase and the sentiment clues and the degree to which the candidate is correlated with the issue.



Issue: *Afghanistan War*

September 11	positive	The Afghanistan war launched after the September 11 ...
	negative	The Afghanistan war was of revenge by the Americans for September 11
Troops	negative	Most Americans oppose sending more troops to Afghanistan war ...
Weapons of mass destruction	negative	The Afghanistan war is perilous because of weapons of mass destruction ...
Obama	positive	Obama supports the Afghanistan war ...

Fig. 1. A summary of the sentiment-generating subtopics for an issue "*Afghanistan War*"

2 Definition

We begin with some definitions of key concepts to be used in our models and analysis. While some of them are our own, other are borrowed from the existing literature and extended.

Controversial Issue: It is operationally defined to be a piece of text that lends itself to a query for a search engine and invokes conflicting sentiment or views. The criteria ensure its topical coherency and controversy aspects, respectively. While it can be any arbitrarily long text, we confine it to a noun or verb phrase in this paper. For example, "*Afghanistan war*" would be a controversial issue.

Subtopic: A subtopic is defined to be an entity or concept that is meaningfully associated and subordinated with a controversial issue. It plays the role of making the associated issue controversial and thus serves as a reason or a factor for a particular view or sentiment about the issue. Given a controversial issue, it is natural that related subtopics change over time. In this paper, the unit of a subtopic is confined to be a noun phrase. For example, "*troops*" can be a subtopic associated with "*Afghanistan war*". Note that a person who expresses an opinion on (e.g. *should send more*) or takes an action (e.g. *support*) for a topic can be a subtopic in that opinions or actions of different polarity (e.g. *should reduce* or *disapprove*) may exist.

Sentiment: It refers to an affective state invoked by a piece of text that describes an issue or topic involving real-world entities or abstract concepts such as events and policies. Such an affective state can be associated with evaluation or judgment, approval/disapproval, and feelings about an issue or a topic. Readers of a news article may feel some sentiment even from factual statements based on their own evaluations. Sentiment has three polarities: positive, neutral, and negative. For example, not only do subjective opinions about "*Afghanistan war*" contain sentiment but also a report on a death of a soldier and civilian generates sentiment.

Sentiment Clue: A sentiment clue is a word or phrase that causes positive or negative sentiment in a sentence [4]. Sentiment clues may be domain-dependent or independent. Some domain-independent clues are found in a lexicon such as SentiWordNet [7]. For example, "*sad*" or "*death*" can be sentiment clues for the topic "*troops*" under "*Afghanistan war*". However, the sentiment from a word like "*long*" or "*increase*" may depend on the domain, depending on what is described.

Sentiment Holder: It refers to the source of sentiment [8], which can be explicit or implicit, depending on whether a sentiment-revealing sentence contains the entity who expresses the sentiment. When a factual statement invokes sentiment, the author is assumed to be the implicit sentiment holder. For example, "*Obama*" in "*Obama supports Afghanistan war*" is the sentiment holder since the sentiment clue "*support*" generates positive sentiment. In this paper, a sentiment holder can be a subtopic as explained above.

3 The Method

Our proposed method consists of two parts. The first one is for issue detection with which controversial issues are identified from news articles. We measure the "controversiness" of a phrase by its topical importance and sentiment gap it incurs. The other part is for extraction of subtopics that are related to the detected issue

3.1 Controversial Issue Detection

According to the definition in Section 2, a controversial issue would be often found in the form of a search query that retrieves a set of documents containing positive or negative sentiment towards the query topic. Therefore, the first step for identifying a controversial issue is to guess a potential query. We do this by adopting a known-item query generation method in [9] and extending it to generate controversial issues. A known-item query can be generated from the document known to be relevant based on its probabilistic model. Given a document, the algorithm selects a query length *l* and repeats *l* times the process of selecting a term based on its generation probability from the document model. Our algorithm follows the same flow:

```
1. Initialize issue term set: $issue\_terms = \{\}$
2. Repeat $l$ times: /* $l$ is an empirical value */
    i. Select a term $t_i$ with probability $p(t_i \mid \theta_{it})$
    ii. Add $t_i$ to the issue term
```

but uses a different method for the probability calculation to suit our need and extends it further to consider phrases. The issue term model can be expressed as in Equation (1) where θ_{it} is the model of the issue term. The probability is estimated based on the mixture of a topic model and sentiment model because a term in an issue may contain sentiment in it (e.g. "*war*").

$$p(t_i \mid \theta_{it}) = \lambda \cdot p(t_i \mid \theta_{topic}) + (1-\lambda) \cdot p(t_i \mid \theta_{senti}) \tag{1}$$

where λ is parameter whose range is $0 \le \lambda \le 1$.

The topic model can be estimated as:

$$p(t_i \mid \theta_{topic}) = \frac{n(t_i, D)}{\sum_t n(t, D)} \tag{2}$$

where *D* is the set of news documents, and $n(t_i, D)$ is the number of occurrences of t_i in *D*. The sentiment model is based on calculation of sentiment score of each term:

$$scr(t) = MAX[scr(t \mid POS), scr(t \mid NEG)] \tag{3}$$

$$p(t_i \mid \theta_{senti}) = \frac{scr(t_i)}{\sum_t scr(t)} \tag{4}$$

where $scr(t \mid POS)$ and $scr(t \mid NEG)$ are a positive sentiment score and a negative score of the term *t*, respectively, which are provided by SentiWordNet[2].

In order to handle a phrase *ph*, we consider the issue terms contained in it and calculate its score, average probability:

$$w(t_i) = \begin{cases} p(t_i \mid \theta_{it}) & if\ t_i \in issue\_term, \\ 0 & otherwise \end{cases} \tag{5}$$

$$score(ph) = \frac{\sum_{t_i \in ph} w(t_i)}{\mid ph \mid} \tag{6}$$

where $\mid ph \mid$ is the number of terms of the phrase.

While the algorithm selects phrases with one or more issue terms, there is no guarantee that they are controversial enough. An additional step is to check the degree of controversy using the contextual information. We first compute the score for positive and negative sentiment for a phrase (Equation (7)) and then determine if it is sufficiently controversial not only by the sum of the magnitude of positive and negative sentiments but also the difference between them (Equation (8)). A topic may be of great importance for people but may not be controversial if its sentiment has one polarity (e.g. "*swine flu*" is only negative).

$$scr_{POS} = \sum_{t \in \theta_{ph}} scr(t \mid POS),\ scr_{NEG} = \sum_{t \in \theta_{ph}} scr(t \mid NEG) \tag{7}$$

$$controversial(ph) = \begin{cases} 1 & if\ scr_{POS} + scr_{NEG} \geq \delta, \mid scr_{POS} - scr_{NEG} \mid \leq \gamma \\ 0 & otherwise \end{cases} \tag{8}$$

where $\theta_{ph} = \{sentence \mid ph \in sentence\}$, and δ and γ are empirical values.

3.2 Subtopic Extraction

As we defined in Section 2, every noun phrase in the issue document (i.e., a news article containing one of the detected issues) is eligible as subtopic candidates. After

[2] SentiWordNet, `http://sentiwordnet.isti.cnr.it`

extracting all atomic noun phrases using a parser[3] from each document containing at least one detected controversial issue, we generate a statistical classifier to select a subtopic based on collocation information. A subtopic would co-occur not only with a detected issue but also with sentiment clues, which we use as classification features. In order to incorporate the features, we opt for the linear regression model that works well even with a small amount of training data. The five types of features we consider are explained below: two basic features and three statistical features.

One of the two basic features is whether or not the title section in a document includes a given candidate; a candidate is likely to be a subtopic if it appears on the title section in a document. The other is where in the sentence the candidate appears as it is likely to be a subtopic if it appears on the subject or object position in a sentence. In the case of "*Afghanistan war*", for example, words like "*Bush*", and "*Troop*" would appear on the title of a news article. In addition, they would occupy the position of the subject or object in a sentence.

In addition, three statistical features are used in our method: contextual similarity measured with issue likelihood and subtopic likelihood distribution models, subtopic likelihood computed through a sentiment model obtained for the issue, and direct correlation between an issue and a subtopic in terms of their co-occurrence in the same sentences. The first statistical feature considers the extent to which an issue and a candidate subtopic emerge from the same context whereas the second measures how strongly a subtopic is related with the sentiment in the article. The third feature measures how close the subtopic is to the issue.

We calculate contextual similarity between an issue and a noun phrase (NP) subtopic candidate through KL-divergence that calculates the difference between two probability distributions. The NP and issue models are constructed out of the word frequencies of the sentences containing the term, after removing stop words.

$$KL(\theta_{NP} \,||\, \theta_{Issue}) = \sum\nolimits_{t_i \in \theta_{Issue \cup NP}} p(t_i \,|\, \theta_{NP}) \cdot \log \frac{p(t_i \,|\, \theta_{NP})}{p(t_i \,|\, \theta_{Issue})} \tag{9}$$

where $\theta_{NP} = \{sentence \,|\, NP \in sentence\}$, $\theta_{Issue} = \{sentence \,|\, Issue \in sentence\}$, and $\theta_{Issue \cup NP} = \{sentence \,|\, (Issue \in sentence) \cup (NP \in sentence)\}$. Furthermore, the probabilities are estimated as follows:

$$p(t_i \,|\, \theta_{NP}) = \frac{n(t_i, \theta_{NP})}{|\, \theta_{NP} \,|} \tag{10}$$

where $|\,\theta_{NP}\,|$ is the number of sentences, and $n(t_i, \theta_{NP})$ is the number of sentences which contains a term t_i in θ_{NP}. $p(t_i \,|\, \theta_{Issue})$ can be estimated in a similar way. KL divergence is 0 if and only if two models (i.e., NP and Issue model) are the same. In this paper, we take its inverse and normalize it to make the similarity value range between 0 and 1.

[3] We used Stanford Parser in
http://nlp.stanford.edu/software/lex-parser.shtml

To compute the relatedness of a subtopic candidate with the sentiment expressed for the issue, we estimate the probability of a candidate given an issue as follows:

$$P_{senti}(NP \mid Issue) = \frac{\sum_{t_i \in \theta_{NP \cap Issue}} scr(t_i)}{\sum_{t \in \theta_{Issue}} scr(t)} \tag{11}$$

where $\theta_{Issue \cap NP} = \{sentence \mid (Issue \in sentence) \cap (NP \in sentence)\}$.

While the contextual similarity (the first statistical feature) measures the extent to which an issue and a subtopic share the same context, they may occur in different sentences. In order to give a higher weight to a subtopic that occurs together with an issue in the same sentence, we measure their sentential correlation by counting the number of times they appear in the same sentence as follows:

$$Cor(S_{Issue}, S_{NP}) = \frac{\mid S_{Issue} \cap S_{NP} \mid}{\sqrt{\mid S_{Issue} \mid} \cdot \sqrt{\mid S_{NP} \mid}} \tag{12}$$

where $\mid . \mid$ is the number of sentences in a set.

4 Experiment

4.1 Experimental Setup

We used the MPQA corpus that contains news articles from 187 different foreign and U.S. news sources, ranging June 2001 to May 2002, some of which are classified into 10 different topics [12]. The articles were collected based on information retrieval system search results and manual analysis. Since they contain sentiment information, the 10 topics can be regarded as controversial issues. It contains 355 documents with topic information and 8,955 sentences. Table 1 lists the topics and the number of documents in each topic.

To build a gold standard for evaluation of the subtopic extraction method, a portion of the corpus was manually analyzed and the subtopics for the issues were tagged. Five topics, AE, GB, HR, KP, and ZI, were chosen randomly because of the limited human resources. Since polarized sentences are marked in the MPQA corpus, we easily obtained 2,002 (36.84%) polarized sentences among 5,434 in the set of 204 documents in the chosen topics. Three annotators picked subtopics from the polarized sentences. When there are conflicts, they were resolved by majority voting. As a sentence may have more than one subtopic instance, a total of 2,723 subtopics (1.36 subtopics / sentence) were identified by one or more annotators. Among those, a total of 1,947 subtopics (71.6%) were agreed by two or three annotators. After collapsing the redundant subtopic instances (i.e. the same phrases), the average number of unique subtopics per issue was 17.8.

Table 1. The topic list and the number of documents about each topic in the MPQA corpus

Topic	# documents
President Bush's 2002 State of the Union Address – axis of evil (AE)	30
U.S. holding prisoners in Guantanamo Bay (GB)	55
Reaction to U.S. State Department report on human rights (HR)	30
Ratification of Kyoto Protocol (KP)	43
2002 president election in Zimbabwe (ZI)	46
Israeli settlements (IS)	25
Space missions of various countries – space station (SM)	30
Relations between Taiwan and China (TC)	28
Presidential coup in Venezuela (VE)	37
Economic collapse in Argentina (AR)	31

4.2 Experimental Results

Controversial Issue Detection. It is not easy to determine how appropriate a detected issue is. The goal of this part of the experiment is not to quantify any performance but to obtain a qualitative assessment of the proposed issue detection method. Issue phrases were detected based on the algorithm with two thresholds δ and γ set to 250.0 and 50.0, respectively (As we mentioned, two thresholds are empirical values).

Table 2 shows top-three phrases detected as controversial issues for the ten topic categories, together with the number of detected phrases over the thresholds. The underlined parts indicate they appear in the original topic category names in the corpus as in Table 1. Although the algorithm does not consider the location of phrases, those in the human generated topic titles were captured as controversial issues. Even in the case of IS (Israeli settlement), where none of the system-generated issues appear in the human-generated titles, our perusal of the related articles indicates that the main issue centres on "*occupied Palestinian territory*" for which "*Palestinian*" and "*Palestinian state*" seem appropriate. We feel that not only are the issues close to the topics but also the granularity of the detected issues is appropriate by and large.

Table 2. Sentiment issues extracted from MPQA corpus

Topic	Count	Detected sentiment issues		
AE	5	axis of evil	president Bush	Iran Iraq and North Korea
GB	4	prisoner	Guantanamo bay	Guantanamo detainees
HR	3	state department report	human rights	the human rights report
KP	4	Kyoto protocol	greenhouse	climate change
ZI	4	Mugabe	Zimbabwe election	presidential election
IS	3	Palestinian	Israel	recognize Palestinian state
SM	3	space station	Russian space officials	space activities
TC	4	Taiwan	China	Taiwan relations
VE	3	president Hugo Chavez	Venezuela	Chavez government
AR	4	Argentine government	economy	help Argentina

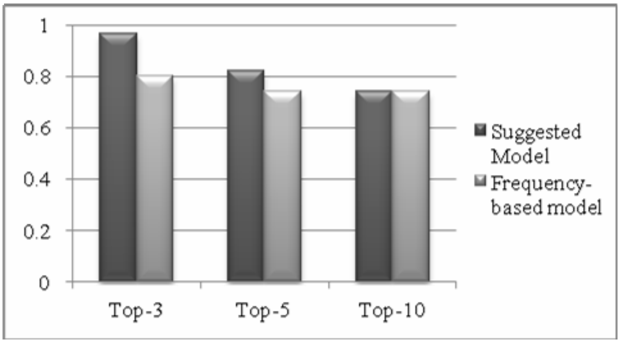


Fig. 2. Comparison of the controversial issue detection model with the frequency-based model

We adopted a user evaluation to quantify the performance. We chose top-ten phrases detected as controversial issues for each topic and provided these to three users with random-ordering. Then, users judged whether a given phrases is appropriate for a controversial issue or not. We considered phrases correct when they were agreed by two or three users. We compared the ranking generated by the proposed model with that of the frequency-based model which ranks the phrases in their frequency (Figure 2). We experimented in three cases: top-three phrases, top-five phrases, and top-ten phrases. The precision is calculated as follow:

$$precision = \frac{the\ number\ of\ correct\ phrases}{the\ total\ number\ of\ phrases} \qquad (13)$$

In the top-ten phrases case, they showed the same result because the set of the phrases is identical and only rank lists are different. Although the difference between two models is small in the top-five phrases case (the precision of our suggested model is 0.83 while that of the frequency-based model is 0.74), our suggested model outperforms the frequency-based model, and the precision is nearly 1.0 in the top-three phrases case (the precision of our suggested model is 0.9667 while the frequency-based model is 0.8).

However, our model includes some problems. Most of the detected phrases are noun phrases while verb phases would be equally useful as issues. Our analysis shows that the frequency-based algorithm prefers noun phrases because noun phrases are repeated more often than verb phrases in the same article. This is the reason why "*recognize Palestinian state*" for IS and "*help Argentina*" in AR were ranked low in the list. Taking this phenomenon into account is left for future work.

Subtopic Extraction. The subtopic extraction method was evaluated with a manually annotated gold-standard. Its performance was measured in terms of precision and recall, which counts how many of the extracted subtopics were found in the gold-standard and how many of the annotated subtopics were detected by the method, respectively.

The precision and recall results are shown in Table 3 for the five topic areas chosen for this part of the experiment. It was not possible to make comparisons with other methods, simply because we do not know any other previous approach to subtopic identification for a given issue. The precision and recall values are not high in general. There are several weaknesses in the proposed algorithm. Most notably, its inability to handle anaphora that is found very often in news paper articles gave many incorrect subtopics. Since we considered all noun phrases in a sentence as candidates of subtopics, anaphoric phrases, such as "*these three counties*", "*such policy*", and "*his point of view*", also are considered as candidates. For example, for the incorrectly detected subtopic "*these three countries*", the correct answer in the gold standard is the list of actual country names such as "*Iran, Iraq, and North Korea*", i.e., the noun phrase which the anaphor refers to. Furthermore, semantically identical expressions that differ from each other at the surface-level are treated as different phrases. For example, "*Iran, Iraq and DPRK*" and "*Iran, Iraq, and North Korea*" are treated as different phrases. Another example is "*mass destruction weapons*" and "*weapons of mass destruction.*" Since the expressions in news articles are very diverse, we found there are many cases involved with this problem. Another weakness is associated with reliance on frequency in the features. Some of the important subtopics do not actually occur very often in news articles. Finally, incongruent candidates, such as "*all efforts*" and "*a few countries*", also are problem because we considered all noun phrases as candidates.

Table 3. Precision and recall of our proposed model

Topic	Precision	Recall	F1-score
AE	0.4170	0.7778	0.5429
GB	0.4000	0.6842	0.5049
HR	0.3167	0.7778	0.4501
KP	0.3167	0.8333	0.4590
ZI	0.4000	0.7500	0.5217
Total	**0.3700**	**0.7640**	**0.4986**

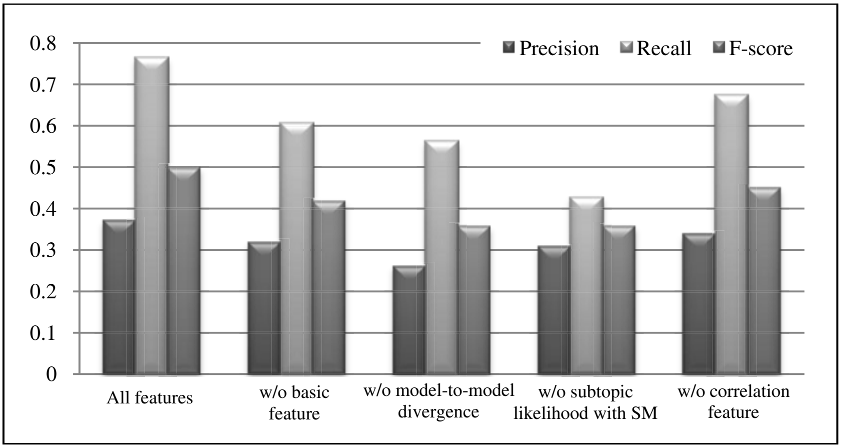


Fig. 3. The effect of ignoring one feature at a time

In order to see relative importance of the five features, we measure changes of classification performance caused by excluding one feature at a time. As in Figure 3, the subtopic likelihood with respect to the sentiment model (SM) and the model-to-model divergence turn out to be the most important since the performance was decreased most significantly. Without associating sentiment information, the subtopics were mostly terms with high frequency and named entities (e.g., "*Bush*" and "*U.S.*"). Without the contextual similarity measured by KL-divergence, many of the extracted subtopics are not directly related to the issue. Although the basic features are simple, their effect cannot be disregarded. The correlation feature contributes the least but also affect the performance in a non-trivial way.

Application. We envision that the proposed method can be used to build a system that detects a controversial issue (or alternatively received a user query for it) and organizes the subtopic detection result as in Figure 4, where the identified subtopics are shown on the left together with the polarity, the actual text, and the date it appears first. This will allow users to quickly understand what subtopics exist, what sentiment has been expressed, and what the examples are.

Axis of Evil			
A spokesman of the EU Commission	NEG	A spokesman of the EU Commission, at a press conference, expressed concerns over the Bush gaffe labeling North Korea, Iran, and Iraq as the axis of evil and said that the EUs high representatives do not agree to such policy.	2002.03.14
the Finnish and B elgian	NEG	On 17 February, the Finnish and Belgian Foreign Ministers also expressed opposition to the so-called axis of evil remark and the US plan to launch strikes against Iraq.	2002.03.15
a new threat	NEG	In relation to Bushs axis of evil remarks, the German Foreign Minister also said, Allies are not satellites, and the French Foreign Minister caustically criticized that the United States unilateral, simplistic worldview poses a new threat to the world.	2002.03.15
weapons of mass destruction	POS	A day after President Bushs threat to crush these countries and use all means to prevent them from developing weapons of mass destruction, Secretary Powell said before the Senate Foreign Relations Committee that characterizing these counties as the axis of evil does not mean that his government intends to invade them. Another goal is to prevent countries that support terrorism from threatening the United States …	2002.02.01

Fig. 4. An example result of a controversial issue and subtopics

5 Related Work

To the best of our knowledge, there has been no attempt to identify controversial issues and their related subtopics that serve as reasons for different sentiments from news papers. The closest work we know is found in Qiaozhu Mei et al. [1]. It proposes a topic-sentiment mixture model to extract topics and their sentiments from blog articles. They used a mixture of multinomials; background topic model, subtopic model, positive sentiment model, and negative sentiment model. To evaluate the topic extraction model, they used only two data sets which are constructed by submitting queries such as "*ipod*" and "*da vinci code*". Since a query is a product like a movie or a book and the blog articles are review data, its subtopics are related to its attribute. Aside from the fact that they only deal with unigrams where as we deal with phrase, their approach is not to detect issues and their subtopics but to find products and their attributes, which is much simpler problem.

Another is Ba-Quy Vyong et al. [11] work. They suggested Controversy Rank model in Wikipedia. They thought a dispute in an article is more controversial, so this model utilized the controversy level of disputes which can be derived from the articles' edit histories. The CR Models defined the article controversy score (an article is controversial when it has a lots disputes among less contributor) and the contributor controversy score (a contributor is controversial when he/she is engaged in a lots disputes in less articles). However, this model can only apply to Wikipedia because general news articles have no edit histories and no contributor except writer.

There was an attempt to extract a sentiment topic, which is similar to a controversial issue. Kerstin Denecke et al. [6] focused on discovering a main topic and identifying sentiment at sentence level from blogs. They detected a topic by applying the Latent Dirichlet Allocation algorithm, and identified its sentiment using SentiWordNet. Although their output is a topic and its sentiment at sentence-level, to be accurate, we cannot say the definition of a topic in this study is same as our definition of a controversial issue because they do not use the sentiment model when a main topic is detected.

Soo-Min Kim and Eduard Hovy [10] tried to extract an opinion topic utilizing FrameNet from news articles. If all sentences are simple and FrameNet covers lots of opinion words, this method would work very well because an opinion topic may be semantically related to an opinion word. However, the method cannot capture a sentiment topic if there are complicated semantic relations. To resolve this problem, it is necessary to analyze text in a complex manner. Besides, it can be applied to only sentence-level topic detection.

The work in [4] extracts a sentiment topic exploiting co-occurrence information with sentiment clues. It computes cosine similarity between a candidate (noun phrase) and sentiment clues to identify a sentiment topic. However, the goal is to generate domain specific sentiment clues for the sentiment classification by using topics as a vehicle, not to identify a sentiment topic of issues. Besides, the boundary of sentiment topics is confined to a sentence.

6 Conclusion and Future Work

This paper tackles two problems: controversial issue detection and their subtopic extraction. For issue detection, we identify noun or verb phrases as candidate issues using a mixture of topical and sentiment models. To compute the degree of controversy, we measure the amount of both positive and negative sentiment and the difference between them. For subtopic extraction, we generate noun phrases as candidates and calculate their feature scores for classification. We use two positional features and three statistical features. The statistical features are: contextual similarity between the issue and a subtopic candidate, relatedness of a subtopic to sentiment, and the degree of physical vicinity between the issue and the candidate phrases. The experimental result shows that the proposed method is reasonable as the first attempt in extracting controversial issues and their related subtopics.

Our qualitative analysis of the result generates a number of possible extensions to the current method. Besides adding more meaningful features for classification and improvements in probability estimates, which are always possible, we need to investigate on issue and subtopic boundary detection for the right level of granularity, beyond the current approach of using verb and noun phrases. Another important issue is to deal with paraphrases both for the purpose of identifying both issues and subtopics but also for evaluations.

Acknowledgments. This work was financially supported by the MKE (The Ministry of Knowledge Economy), Korea, under the ITRC (Information Technology Research Center) support program supervised by the NIPA (National IT Industry Promotion Agency" (NIPA-2010-C1090-1011-0008).

References

1. Mei, Q., Ling, X., Wondra, M., Su, H., Zhai, C.X.: Topic Sentiment Mixture: Modeling Facets and Opinions in Weblogs. In: The 16th International Conference on World Wide Web (WWW), pp. 171–180. ACM, Canada (2007)
2. Denecke, K., Taytsarau, M., Palpanas, T., Brosowski, M.: Topic-related Sentiment Analysis for Discovering Contradicting Opinions in Weblogs. Technical Report, University of Trento (2009)
3. Ku, L.-W., Liang, Y.-T., Chen, H.-H.: Opinion Extraction, Summarization and Tracking in News and Blog Corpora. In: American Association for Artificial Intelligence-Spring Symposium on Computational Approaches to Analyzing Weblogs (AAAI-CAAW), pp. 100–107 (2006)
4. Choi, Y., Kim, Y., Myaeng, S.-H.: Domain-specific Sentiment Analysis using Contextual Feature Generation. In: 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion Measurement (TSA), pp. 37–44. ACM, Hong Kong (2009)
5. Zhuang, L., Jing, F., Zhu, X.-Y.: Movie Review Mining and Summarization. In: 15th ACM International Conference on Information and Knowledge Management (CIKM), pp. 44–50. ACM, USA (2006)
6. Zhang, M., Ye, X.: A Generation Model to Unify Topic Relevance and Lexicon-based Sentiment for Opinion Retrieval. In: 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 411–418. ACM, Singapore (2008)
7. Esuli, A., Sebastiani, F.: SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. In: 5th Conference on Language Resources and Evaluation (LREC), pp. 417–422 (2006)

8. Stoyanow, V., Cardie, C.: Annotating Topics of Opinions. In: 6th International Conference on Language Resources and Evaluation (LREC), pp. 3213–3217 (2007)
9. Azzopardi, L., de Rijke, M., Balog, K.: Building Simulated Queries for Known-Item Topic: An Analysis using Six European Languages. In: 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 455–462. ACM, Amsterdam (2007)
10. Kim, S.-M., Hovy, E.: Extracting Opinions, Opinion Holder, and Topics Expressed Online News Media Text. In: The Workshop on Sentiment and Subjectivity in Text, pp. 1–8. Association for Computational Linguistics, Sydney (2006)
11. Vuong, B.-Q., Lim, E.-P., Sum, A., Le, M.-T., Lauw, H.W., Chang, K.: On Ranking Controversies in Wikipedia: Models and Evaluation. In: 1st ACM International Conference on Web Search and Data Mining (WSDM), pp. 171–182. ACM, USA (2008)
12. MPQA corpus, http://www.cs.pitt.edu/mpqa

Efficient Privacy Preserving K-Means Clustering

Maneesh Upmanyu, Anoop M. Namboodiri, Kannan Srinathan, and C.V. Jawahar

International Institute of Information Technology, Hyderabad
{upmanyu@research.,anoop@,srinathan@,jawahar@}iiit.ac.in

Abstract. This paper introduces an efficient privacy-preserving protocol for distributed K-means clustering over an arbitrary partitioned data, shared among N parties. Clustering is one of the fundamental algorithms used in the field of data mining. Advances in data acquisition methodologies have resulted in collection and storage of vast quantities of user's personal data. For mutual benefit, organizations tend to share their data for analytical purposes, thus raising privacy concerns for the users. Over the years, numerous attempts have been made to introduce privacy and security at the expense of massive additional communication costs. The approaches suggested in the literature make use of the cryptographic protocols such as *Secure Multiparty Computation (SMC)* and/or *homomorphic encryption schemes* like Paillier's encryption. Methods using such schemes have proven communication overheads. And in practice are found to be slower by a factor of more than 10^6. In light of the practical limitations posed by privacy using the traditional approaches, we explore a paradigm shift to side-step the expensive protocols of SMC. In this work, we use the paradigm of *secret sharing*, which allows the data to be divided into multiple shares and processed separately at different servers. Using the paradigm of secret sharing, allows us to design a provably-secure, cloud computing based solution which has negligible communication overhead compared to SMC and is hence over a million times faster than similar SMC based protocols.

Keywords: K-Means, Privacy, Security, Secret Sharing.

1 Introduction

K-means clustering [1] [2] is one of the most widely used techniques for statistical data analysis. Researchers use cluster analysis to partition the general population of consumers into market segments and to better understand the relationships between different groups of consumers/potential customers. However the collected data may contain sensitive or private information, thus heightening the privacy concerns [3] [4]. The privacy and secrecy considerations can prohibit the organizations from sharing their sensitive data with each other. The solution should not just be provably secure i.e. it leaks no additional useful information, but should also minimize the additional overheads in terms of communication and computation costs required to introduce privacy. Addressing the problem requires many practical challenges to overcome before a possible wide-scale deployment. Solutions were sketched to extract knowledge by making the participating parties to compute common functions, without having to actually reveal their individual data to any other party [5] [6].Vaidya *et al.* [7] summarize the state of

H. Chen et al. (Eds.): PAISI 2010, LNCS 6122, pp. 154–166, 2010.

© Springer-Verlag Berlin Heidelberg 2010

art methods available for privacy preserving data mining. More detailed reviews of the previous work can be found in Verykios *et al.* [8].

Previous solutions can be primarily categorized as, *i)* those using *Data Perturbation techniques*, and *ii)* those employing *Secure Multiparty Computation (SMC)*. The first category of approaches introduces noise and data transformations to achieve partial privacy [9] [10] [11]. The clustering is then done of the noisy version of the data, resulting in approximately correct clusters [5] [12]. Such approaches compromise privacy for practicality, however the key advantage is the negligible communication overhead needed by such approaches.

The second category of approaches aims to achieve complete privacy. This is done using the well known cryptographic protocol of SMC [13]. SMC facilitates a group of people, each with its own private data, to perform some common computation task on the aggregate of their data. SMC ensures that, in the process, no personal information of data is revealed to any one [14]. However, the SMC based protocols are found to be extremely computationally expensive [13]. In other words, an operation which requires a single round of communication in a non-secure implementation, would require hundreds of thousands of rounds of communication (depending on the domain size) to achieve the same operation in a secure implementation using SMC. For data mining applications, the sheer volume of the data involved makes the protocol infeasible in terms of the communication cost. For example, Vaidya *et al.* [15], İnan *et al.* [16] and Wright *et al.* [17] use SMC as a subroutine to propose privacy preserving clustering. However, the huge computational costs makes these solution of limted practical interest.

Another set of proposed approaches uses the semantically secure additive or multiplicative homomorphic encryption schemes [18] [19]. In such a protocol, one party encrypts its data using its public key, and share the encrypted data with the other party for computation. Interactive protocols are then designed to carry out the clustering algorithm [20] [10]. The overheads of encryption and the communication costs needed to carry out clustering limits the scope of such algorithms. Interaction can be reduced with the usage of a doubly homomorphic scheme [21]. However, the only known doubly homomorphic scheme is the one recently proposed by Craig Gentry [22] and would most likely lead to a computationally intensive theoretical solution.

In this work, we achieve the security at the level of SMC while keeping the communication costs to a level similar to that of the first category. We achieve this using the paradigm of the *Secret Sharing*[23] [24] over a mesh of processing servers. Our solution is first of its type, and is both efficient and mathematically simple. In the process we also side-step the communication bottlenecks posed by the usage of SMC and asymmetric encryption schemes. Our proposed solution is not only computationally efficient but also secure independent of whether or not $P \neq NP$. We however do assume the servers to be non-colluding and having the ability to generate random numbers.

We address the scenario of N parties, sharing an arbitrary partitioned data [17], wishing to privately collaborate for doing cluster analysis on their aggregated data. In our setting, the attribute names form the public information. Each of the entity is either completely owned by one of the users, or the attributes are shared among the N users, where the share of some users can also be ϕ. If a record is ‘completely’ owned by anyone, then its existence remains hidden from other users. If any of the attributes for an

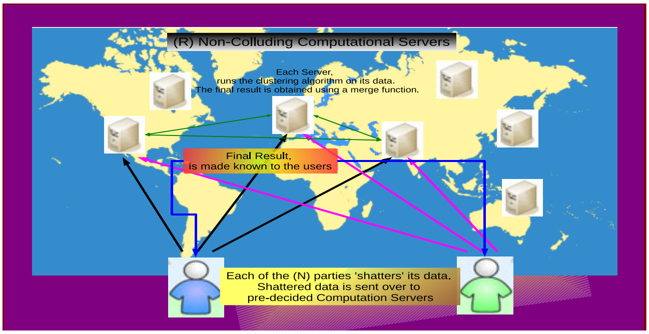

(R) Non-Colluding Computational Servers
Each Server,
runs the clustering algorithm on its data.
The final result is obtained using a merge function.
Final Result,
is made known to the users
Each of the (N) parties 'shatters' its data.
Shattered data is sent over to
pre-decided Computation Servers

Fig. 1. Sample Mesh of servers. Each of the N users shatters their private data (Sec: 2) and sends over the shares to the pre-selected R servers for computation. The final result is obtained by merging (Sec: 2) the outputs of the computational servers. In above example, N is 2 and R is 3.

entity are with more than one user, then a weighted average of the attribute values is considered for the computational purpose. Entities are indexed using a mutually agreed upon indexing scheme. The indexing scheme addresses the two concerns of *i)* hiding the entity's identity from the servers, and *ii)* a common index for accessing the vertically partitioned data. We now look at the architecture of our proposed solution.

We propose a 'cloud computing' based solution that utilizes the services of R, $(R > 2)$, non-colluding servers. Each of the N users, is required to compute the R secret shares of its private data using a *shatter function* (see the algorithm, defined in Sec: 2). Each share is then sent over to a specific server for processing. Note that the shatter function ensures that the computed secret shares on its own reveal no information about the original private data. The cloud of employed servers, now runs the K-means algorithm using just the secret shares. The protocol ensures that none of the users/servers have sufficient information to reconstruct the original data, thus ensuring privacy. As shown, later in the paper, the shatter function that we choose allows efficient computations using just the shares. That is, unlike SMC, the number of rounds of communication to implement an operation on secret shares is equivalent to that required in a non-secure implementation of the same operation. The advantage of this is that it significantly reduces the communication costs over the similar SMC based protocols, thus making privacy preserving clustering practical. Figure 1 shows a pictorial description of the proposed architecture, while the algorithm is discussed in detail in Sec: 3.

2 The Building Blocks of Security

We use the paradigm of Secret Sharing (SS) to achieve privacy and efficiency. Secret Sharing (SS) [25] [26] [23] refers to the methods for distributing a secret among a group of servers, each of which is allocated a share of the secret. The secret can be reconstructed only when the shares are combined together; on their own, they have no meaningful information. In our problem setting, we ask each of the collaborating users to compute the secret shares of their private data, and send them over to the processing servers. The processing servers then privately collaborate (without reconstructing the actual data) to run the K-means algorithm over the secret shares. *Note that*, not all SS methods allows computation on the secret shares. In order to achieve this, we adopt the Chinese Remainder Theorem (CRT) based secret sharing schemes [23] [27].

However, in the SS schemes of Asmuth *et al.* [23], and Goldreich *et al.* [27], the size (the number of bits) to represent each share is greater than the size of the original data. In other words, for R servers, using these schemes results in a minimum of R fold storage increase. Data expansion is important since it results in cost overheads in terms of storage and interaction among the servers. It becomes even more critical for applications such as data mining that deals with voluminous data.

Understanding the similar limitations, Upmanyu *et al.* [24] recently proposed an efficient method to do privacy preserving surveillance on videos (voluminous data). In this work, we extend their method and propose secure protocols to privately carry our collaborative clustering. The data to be clustered using K-means can be thought of as points in a D dimensional Cartesian space. The data is bounded, i.e. it has a fixed range, and its scale invariant, i.e. even if we scale the axis, the cluster assignment will still be the same. These two are the required desirable properties of the data, that are sufficient for one to adopt the secret sharing scheme as proposed by Upmanyu *et al.* in [24]. We therefore, adopt their *Shatter* (to compute the secret shares) and *Merge* (to reconstruct the secret) functions for the Cartesian data and design a communication and computationally efficient solution to achieve privacy preserving K-means clustering.

Our proposed solution can be summarized as a three step protocol, 1) each user computes the secret shares of his private data, 2) shares are then sent over to a cloud of servers and clustering is privately carried out over the shares, and 3) the users reconstructs the cluster assignment and the cluster centers using the *Merge* function. Before we jump into describing the K-means protocol in Sec: 3, for the sake of completeness we briefly describe the *Shatter* and *Merge* functions as defined in [24]. We also provide an outline of the analysis of the computational and communication overheads and the privacy achieved in each of the sub-steps of the protocol. For those interested, the detailed analysis can be done in a manner similar to in [24]. The *Shatter* and *Merge* functions as defined in [24] are as follows:

Shatter Function $\phi(x)$ **-** *Compute and store the secret shares of the private data :* is defined as the one that splits the data x into R parts, x_1, x_2, ..., x_R, such that each share, x_i, by itself does not reveal any information about x. The participating users pre-decide a set of R primes $P_1, \cdots, P_R$ and a scale factor S. The *Shatter function* is defined as:

$$x_i = \phi(x, P_i) = (x \cdot S + \eta) \bmod P_i, \tag{1}$$

where x_i is i^{th} secret share, and η is an independent random number for each secret x, such that $0 \leq \eta \leq S/2$. The secret share x_i is stored with the i^{th} server and on its own gives little meaningful information of x.

In our scenario, each user can shatter his data (each attribute of a record is shattered independently, η is random for each attribute) and sends over the shares to the specific servers for storage. The size of each share is given by $log(P_i)$ per attribute.

Merge Function $\mu()$ **-** *Reconstruct the secret :* given, $x_i = \phi(x, P_i)$ for different prime P_is, the secret x can be recovered using *CRT* [28] by solving a system of congruence. The *merge function* $\mu()$ is defined as:

$$x = \mu(x_i, P_i) = \frac{CRT(x_i, P_i)}{S} \tag{2}$$

CRT recovers $(x \cdot S + \eta)$, which is appropriately scaled down (integer division by the scale factor) to get the actual value of x. Note that η, which was randomly chosen for each attribute value is not used for recovering the secret. The CRT hence forms our recovery transformation $\mu()$. In our scenario, $\mu()$ is used for reconstructing the cluster centers as computed by the clustering algorithm.

3 The Proposed Algorithm

Following notations are used for describing the protocol. Let L be the number of entities, each made up of D attributes. K be the number of clusters required, and $\boldsymbol{C}_i$, $1 \leq i \leq K$, denotes the cluster locations. The data is arbitrary partitioned among N users. R $(R > 2)$ is the number of computation servers employed. Each server is associated with a unique prime P_i, therefore the number of primes is also R. Each entity is represented in a D dimensional space. The common distance metrics; such a Euclidean, Manhattan or Minkowski; are used for finding the distances. To explain the algorithm we will consider a Euclidean space. As the final output of the privacy-preserving K-means (PPKM) algorithm, each user learns the cluster assignment of the entities owned by them, i.e. which of their entities belong to each clusters. If agreed upon, the location of the K-clusters is also revealed to the users.

The complete protocol can be divided into two phases. The *first phase* deals with *i)* choosing the appropriate primes and the scale factor, *ii)* shattering the data, and *iii)* secure aggregation of the data at the servers. The *second phase* of the protocol deals with the clustering algorithm on the aggregate of the shattered data available with the R computational servers. The basic algorithm follows directly from the standard K-means algorithm [29], which consists of three steps, *i)* Initialization, *ii)* Lloyd Step, and *iii)* Stopping Criterion. The complete protocol is as follows:

3.1 Phase One: Secure Storage

The first step is the selection of an appropriate residue number system (RNS) [24] for secure storage. We extend the *analytical method* [24] to compute the parameters required for ensuring the security and privacy in our problem setting. For a value of R we select $P_1, \cdots, P_R$, such that their product, P, is larger than any intermediate value we have to represent in our algorithm. This range can be easily computed from the range of values we expect in the computations. Scaling the axis and translating the origin of an Euclidean space does not change the final cluster assignment. Hence we represent negative numbers with an implicit sign [30], i.e. $-x \equiv 2M - x$. Floating point data is taken care of by appropriately scaling the dataset to retain a certain decimal precision.

Let $[-U, U]$ be the range of numbers we expect in the computations on secret shares. We choose P_j's such that $P = \prod_{j=1}^{R} P_j \geq 2U$. Typically, one could just choose the smallest of the R consecutive primes satisfying the above property. For complete obfuscation of the data, the scaling factor chosen should be higher than the largest prime [24]. We now analytically choose the optimal set of parameters for our problem setting.

Parameter Selection: Let $[-M, M]$ be the attributes domain. Then the points can be represented in a $D - dimensional\ Euclidean\ space$, $\mathbb{R}^D_{2M}$. Let W_1 be the square of the maximum possible Euclidean distance between two points, i.e. the distance between the two extreme points, thus we get $W_1 = 4M^2D$. Also let W_2 be the maximum sum of the coordinates we can get for a cluster (needed for computing the cluster's mean). This is easily computable as $W_2 = 2ML$ (entire database belong to a single cluster). Let W be the upper range of number we expect in K-means, therefore we have $W = max(W_1, W_2)$. Let us now assume, S to be the required scale factor to get complete privacy. The input data is scaled using this factor. This can be viewed as scaling the axis of the Euclidean space by S, i.e. a point x in the old coordinate system is mapped to $S{\cdot}x$ in the new scaled space. Therefore, we get $U = max(W_1{\cdot}S^2, W_2{\cdot}S)$. The primes now need to be chosen such that:

$$S \geq \max_j P_j, \text{ and } P \geq 2U. \tag{3}$$

Simplifying the above, we find that if:

$$S \approx (2W)^{\frac{1}{R-2}} \tag{4}$$

then the individual servers will have little meaningful information [24].

Each of the N-parties uses the *shatter function* (Eqn: 1), to compute secret shares of their respective data. The shares are then sent over to the servers for processing. Note that we make no assumptions on how the attributes of various data points are partitioned among the N-parties. If $\mathcal{D}$ is the (virtual) database arbitrarily shared among the N parties. Each server j basically then stores the shatter of $\mathcal{D}$ w.r.t. P_j.

Privacy: Each server stores only the shattered share of the data. As long as the servers do not collude, little meaningful information of the entities is learned by any of the servers. This follows directly from the security of the shattering scheme [24]. In this entire phase the only information learned is of how the data is actually being partitioned among the users, i.e., for each entity which all attributes are being held by which user. However we note that, in practice this information gain is not significant, and known a prior [15]. The indexing scheme employed ensures that the identity of the entity remains unknown to the servers.

3.2 Phase Two: Secure K-Means

At the end of the phase one, each computation server stores the secret shares (w.r.t. prime P_j) of the database $\mathcal{D}$. Since the scaling factor S was kept positive, the distance comparison in the original space will be equivalent to distance comparison in the new scaled space. Thus, the cluster assignment of the entities in the scaled space would be identical to what we would have expected in the original space. The final cluster locations are obtained from the cluster centers that are learned in the transformed space after appropriately scaling down and removing the introduced randomness.

Our algorithm will follow the same iterative structure as that of the standard K-means algorithm [29]. The objective is to cluster the data (available as secret shares), without leaking any information to any of the servers. RNS being doubly homomorphic, the operations of addition and multiplication can be independently carried out at each server. However division and comparison (both used in K-means) are difficult to do privately in the RNS. We overcome these difficulties by designing communicationlly efficient, privacy preserving protocols for them over one round of communication.

We now give a step by step description of the protocol used for phase two. *Note here*, that the N users are oblivious of algorithm and the data involved in phase two. The contribution of this paper is not to improve upon the K-means algorithm as such but to propose an efficient protocol to privately carry out the clustering.

Step one: Initialization Let $\boldsymbol{C}_1$, $\boldsymbol{C}_2$, $\cdots$, $\boldsymbol{C}_K$ be the K cluster centers, where each C_k is a D dimensional vector. The clusters are initialized as the K entities from the database $\mathcal{D}$ chosen in a pseudo-random fashion. Since, we want to keep the actual cluster locations also private, we thus store only their secret share components. i.e. for a cluster location $\boldsymbol{C}_k$, $1 \leq k \leq K$, the computational server j, $1 \leq j \leq R$, stores the vector $\boldsymbol{C}_{kj}$, where, $\boldsymbol{C}_{kj}$ is the secret share of $\boldsymbol{C}_k$ w.r.t. P_j.

The servers commonly choose the indices of K entities as the initial cluster centers. The secret shares of the chosen K entities, present with the servers, are used as the secret shares of the initial cluster centers C_k. That is, at server j, C_{kj} initialized to the secret share of the chosen entity. The pseudo-code of the algorithm is given in Algo: 1.

Privacy: Servers do not learn any additional information of the data. The initialization is done, directly using the secret shares. This is done independently at each server, thus resulting in zero computation and communication overheads over TTP.

Algorithm 1. PPKM: Initialization

```
for each cluster, $k$ = 1 to K do
    Choose a random entity index $l$, $l \leq L$
    We want to initialize $\boldsymbol{C}_k = \boldsymbol{X}_l$, where $\boldsymbol{X}_l$ be the $D$ dimensional vector of entity $l$.
    for each server, $j$ = 1 to R do
        Let $\boldsymbol{X}_{lj}$ be the data corresponding to entity $l$ available with the server. We know $\boldsymbol{X}_{lj}$
        is shatter of $\boldsymbol{X}_l$ with mod $P_j$, and was stored with the server during phase one.
        Initialize, $\boldsymbol{C}_{kj}$ to $\boldsymbol{X}_{lj}$, where $C_{kj}$ is the shatter share of $C_k$ with mod $P_j$.
    end for
end for
```

Step two: Lloyd Step In an attempt to minimize the objective function, each iteration reclassifies and recomputes the new cluster locations. The algorithm terminates when it detects *'no change'* (defined by the termination criterion) in the cluster locations. Every iteration can be represented as a sequence of three steps as described below.

i) Finding Closest Cluster Centers: As stated before, since the scaling factor was set to a positive number, finding the closest point is equivalent to finding the one with the minimum of the distances squared in the scaled space. Thus, for every data entity $\boldsymbol{X}_l$, $1 \leq l \leq L$, we find the square of the Euclidean distance to each of the cluster centers $\boldsymbol{C}_k$. The distance square between two D dimensional vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$, is defined as

$$\sum_{d=1}^{D}(X_d^2 + Y_d^2 + 2.X_d.Y_d) \tag{5}$$

which is a set of additions and multiplications. Now, RNS being doubly homomorphic, the above equation can be directly computed using the secret shares. Hence, every server can independently compute the respective secret shares of the distances between the L data points and the K cluster locations. For every data point $\boldsymbol{X}_l$, let $\boldsymbol{T}_l$ be the K length vector, whose share T_{lk} denotes the distance square between data point $\boldsymbol{X}_l$ and cluster center $\boldsymbol{C}_k$. The task is to, without actually reconstructing, compute T_{lk} from the shatter shares of $\boldsymbol{X}_l$ and to assign the point $\boldsymbol{X}_l$ to a closest cluster k.

T_{lk} is represented in the RNS such that T_{lkj} denotes the secret share of T_{lk} (w.r.t. P_j) available at server j. Now, each of the server j can use the Eqn: 5 to compute the share (T_{lkj}) using its locally available secret shares of $\boldsymbol{X}_{lj}$ and $\boldsymbol{C}_{kj}$.

Next, for each data point l, we need to find the cluster k such that T_{lk} is minimum. This would require reconstructing and comparing T_{lk}'s. However, to maintain privacy, the actual distances, T_{lk}'s should be kept private. We overcome this dilemma by applying a clever permutation and randomization scheme. $\boldsymbol{T}_{lk}$ is secured by applying another layer of randomization on the secret shares before sending them over for comparison to another untrusted server (thresholder). Finding the minimum of the K numbers is an $O(K)$ algorithm, i.e. the current minimum has to be compared against the next potential candidate. We next describe the protocol to find the minimum of two numbers, Z_1 and Z_2. This can then be repeated $K-1$ times to find the minimum of K numbers.

Finding the minimum: $(Z_1 - Z_2) \leq 0$ implies $Z_1 \leq Z_2$ else otherwise. In-order to check for this, at each server, we can compute the difference $Z_{1j} - Z_{2j}$ and send over the difference shares to an untrusted server for reconstruction and comparison. However, this naive approach reveals to the thresholder the distance between the two data points. We secure this by randomizing the secret shares of the differences before sending it over for comparison. We can even keep the random number itself unknown to any of the servers by the following protocol.

Each of the R servers chooses a random number r_i and sends over $r_i \ mod \ P_j$ to server j. Thus, each server j, has $\sum_{i=1}^{i=R} r_i \ \% \ P_j$ or $r \ \% \ P_j$, where $r = \sum_1^R r_i$ (Algo 2: steps 5-12). The servers uses this to randomize its share of difference. The randomized difference shares are then sent over to an un-trusted server who reconstructs the randomized difference and returns the comparison against zero for finding the minimum of the two. The smaller number is then compared against the next potential candidate. After a series of $K-1$ comparisons a data point is confidently and privately assigned

Algorithm 2. Find Minimum of K Numbers Protocol

```
1: Let $Z_1$, $Z_2$, ... $Z_K$ be the $K$ numbers we want to find minimum of
2: R is the number of computational servers, each knowing $Z_{kj}$, for $1 \leq k \leq K$ and $1 \leq j \leq R$, where $Z_{kj}$ is the shatter share of $Z_k$ with mod $P_j$. Note that the actual value of $Z_k$ is kept secret from all the servers.
3: Initialize $minIndex$ = 1
4: for every index, k = 2 to K do
5:     for every server, j = 1 to R do
6:         Select a positive random number $r_j$ and share the modulo of $r_j$ with every other server (step 7).
7:         for every other server: i = 1 to R do
8:             Send $r_{ji}$ = $r_j$ mod $P_i$ to the server $i$.
9:         end for
10:    end for
11:    for every server, j = 1 to R do
12:        Let $r'_j$ be the summation of the $R$ random numbers received at each server $j$.
13:        Compute the difference of the secret shares of $Z_{minIndex}$ and $Z_k$. Randomize the difference by multiplying with $r'_j$.
14:        The randomized difference share is sent over to the thresholder.
15:    end for
16:    Thresholder applies the merge function to obtain $R'.(Z_{minIndex} - Z_k)$, where $R'$ is the summation of R positive random numbers $r_j$. The randomized difference is compared with 0 and the result sent back to the servers.
17:    if Threshold Result $> 0$ then
18:        minIndex = k
19:    end if
20:    For next iteration, the role of the thresholder is switched to another pseudo-randomly chosen server.
21: end for
22: Return $min\_index$
```

to a nearest cluster center. Note that the communication costs can further be reduced by choosing the random numbers offline, i.e. when the systems are idle. Each server maintains the list of the secret shares of the random numbers, r's used in the final protocol.

Correctness: Consider a point $\boldsymbol{X}$, for which we want to find which is closer $\boldsymbol{Y}$ or $\boldsymbol{Z}$. Let the points be *shattered* with scale S and randomization $\boldsymbol{a}$, $\boldsymbol{b}$ and $\boldsymbol{c}$ respectively. Thus, we have:

$$(X_1, X_2, \cdots, X_D) \rightarrow (S \cdot X_1 + a_1, \cdots, S \cdot X_D + a_D) \tag{6}$$

$$(Y_1, Y_2, \cdots, Y_D) \rightarrow (S \cdot Y_1 + b_1, \cdots, S \cdot Y_D + b_D) \tag{7}$$

$$(Z_1, Z_2, \cdots, Z_D) \rightarrow (S \cdot Z_1 + c_1, \cdots, S \cdot Z_D + c_D) \tag{8}$$

Let us assume Y is closer than Z, then following holds:

$$\sum (X_i - Y_i)^2 \leq \sum (X_i - Z_i)^2 \tag{9}$$

Using the secret shares, the corresponding distances in the scaled space are computed as:

$$Dist_1 = \sum (S(X_i - Y_i) + (a_i - b_i))^2 \tag{10}$$

$$Dist_2 = \sum (S(X_i - Z_i) + (a_i - c_i))^2 \tag{11}$$

Given that Eqn: 9 holds, the protocol is correct if $Dist_1 \leq Dist_2$. From the constraints given in Sec: 3.1, we know $0 \leq a_i, b_i, c_i \leq S/2$, thus we get $-S/2 \leq (a_i - b_i) \leq S/2$.

$$\sum (S(X_i - Y_i - 1/2))^2 \leq Dist_1 \leq \sum (S(X_i - Y_i + 1/2))^2 \tag{12}$$

$$\sum (S(X_i - Z_i - 1/2))^2 \leq Dist_2 \leq \sum (S(X_i - Z_i + 1/2))^2 \tag{13}$$

Thus, the protocol satisfies correctness if Eqn: 14 is true whenever Eqn: 9 is true.

$$\sum (S(X_i - Y_i + 1/2))^2 \leq \sum (S(X_i - Z_i - 1/2))^2 \tag{14}$$

This will hold if the Cartesian System is designed so as to nullify the effect of the additional $\pm 1/2$ in Eqn: 14. This is achieved by having the step-size in the Cartesian system as 2, i.e. the data is scaled by 2 before choosing the parameters (Sec: 3.1).

Privacy: The protocol is secure against both the GCD and factorization based attacks. The servers are made to jointly choose the randomization, which is different for every threshold operation. This ensures security against the factorization based attacks. The role of the thresholder is also switched among the R servers in an random order, thus ensuring security against the GCD based attacks.

ii) Updating Cluster Locations: Once each of the L data points has been assigned to one of the K clusters, the next step is to recompute the cluster locations. For every cluster k, the cluster center is updated to the center of mass of the newly assigned points to the cluster. Thus, the new coordinate of the cluster k is a (weighted) mean of the corresponding coordinates of the n_k points assigned to the cluster k. Let n_k be the number of data points assigned to cluster k. For any cluster k, each server stores the secret shares of the data points. Each server j, can thus independently compute the sum (Sum_{kdj}) using the secret shares of the n_k data points. The updated cluster location is then obtained by dividing the sum of co-ordinates by n_k. However as we know that the generic division is not defined in the RNS, therefore we cannot directly divide the sum's shares. Furthermore, so as to maintain complete privacy, we will like to keep the updated cluster locations unknown from all the servers. Therefore, an interactive protocol, similar to the one used for thresholding is employed for the job. We now describe the *privacy-preserving division protocol (PPDP).*

PPDP: Consider a number X, secret shares of which are stored at the R servers. The task is to privately divide X by n, such that the secret X and the quotient $q = \lfloor \frac{X}{n} \rfloor$ is kept private from all of the servers. At the end of the protocol, all that the server j gets is the secret share of q w.r.t. P_j. PPDP is achieved through a single round of interaction, and the secret data, X, is secured using a permutation and a randomization method.

Just as in previous protocol (Algo: 2, steps 5-12), the R servers jointly computes two random numbers r and r', such that server j knows only the shares of them. Each server now randomizes its share of X according to Eqn: 15, before sending it over to an un-trusted server. As in the previous protocol, this server is switched among the R servers in a pseudo-permutation fashion. The randomized shares are then reconstructed using the *merge* function to compute X' (Eqn: 15).

Division is then performed to compute the randomized quotient q', as given by Eqn: 17, where q is the actual quotient that we wish to compute (Eqn: 16). We next compute the secret shares of q' and sends them over to the specific servers for de-randomization. Each server computes its share of quotient, q_j, from q'_j using Eqn: 18. The secret share of the cluster center is then updated to the computed share of the quotient. The pseudo-code of the protocol is given in algorithm 3.

Algorithm 3. Privacy Preserving Division Protocol (PPDP)

```
$R$ computational servers, stores i) $X_j$ = shatter of $X$ with mod $P_j$, ii) n
Randomly select $r$, $r'$, in the manner similar to as described in steps(5-12) of algorithm 2.
Let at each server $j$, $r_j$, $r'_j$ be the shatter shares of the two chosen random numbers $r$ and $r'$.
for each server, j = 1 to R do
    Compute $X'_j = r_j \cdot (X_j + r'_j \cdot n)\ mod\ P_j$
    Send $X'_j$ to the thresholder (switched among servers in a pseudo random order).
end for
Thresholder uses the merge function to compute $X'$
Compute $q' = \lfloor \frac{X'}{n} \rfloor$
Send over the $q'_j$ to server $j$, where $q'$ is the shatter share of $q'$ with mod $P_j$.
for each server, j = 1 to R do
    De-randomize the received quotient to get $q_j = (q'_j * r_j^{-1} - r'_j)\ mod\ P_j$
end for
Now, $q_j$ is the required shatter share of the quotient, $q$, with prime $P_j$.
```

$$X \rightarrow X' = r \cdot (X + r' \cdot n) \tag{15}$$

$$q = \frac{X}{n} \tag{16}$$

$$q' = \frac{X'}{n} = r \cdot (q + r') \tag{17}$$

$$q_j = (q'_j * r_j^{-1} - r'_j)\ mod\ P_j \tag{18}$$

Privacy: The PPDP method provides high level of privacy for the secret data. The randomization parameters r and r' are jointly chosen and remains unknown to all. The randomization of the secret data, X, is itself done using the secret shares. The randomization function (Eqn: 15) is designed so as to safeguard against the potential attacks such as factorization and GCD based. In the entire process, no additional meaningful information is leaked to any one. The method not only provides provable privacy but is also efficient with communication cost limited to one round of interaction.

iii) Checking Termination Criterion: At the end of every iteration, we check for the closeness of the new clusters. The 'closeness' is defined as *i)* minimizing the total energy of the clusters, the energy of a cluster k is given as $E_k = \sum_1^{n_k}(\|\boldsymbol{x}_l - \boldsymbol{c}_l\|)$, *ii)* the new clusters locations are close to the old ones. i.e $\sum_1^K(\|\boldsymbol{c}_k - \boldsymbol{c}'_k\|)$, or iii) the number of points making transition across clusters is small.

If the closeness is below the threshold, then we go to step three otherwise continue with next iteration. Any of these definitions can be privately implemented using the approaches like already described.

Step three: Knowledge Revelation. At the termination of the Lloyd step, the cluster centers are stored as the secret shares at the R serves. The cluster assignment of the anonymized entities is also available. To learn the cluster locations, the servers are made to collude under legal agreements. The identity of the entities is known only to the data owner, and hence he is the only one who learns the final cluster assignment. The cluster locations can be revealed, only if agreed upon.

Analysis. We have proposed a provably secure protocol, the proofs of which are similar to those in [24]. The computation overhead at each server is limited to the randomization. The communication overhead is due to one round of interaction to simulate division and comparison operation. However, this overhead is negligible when compared to SMC. No numerical comparisons are provided, due to *a)* space constrains, and *b)* theoretical efficiency, an operation taking hundreds of rounds of communication in SMC is do able using zero or at max one round of interaction in our protocol.

4 Conclusion

We propose a novel 'cloud computing' based solution using the paradigm of Secret Sharing to privately cluster an arbitrary partitioned data among N users. Traditional approaches uses primitives such as SMC or PKC, thus compromising the efficiency and in return provide very high level of privacy which is usually an overkill in practice. The paper contributes a different approach to solve the problem. We show that privacy need not be always at the cost of efficiency. We exploit the properties of the data and the problem to circumvent the limitations faced by traditional methods (that are general-purpose). Our solution does not demand any trust among the servers or users. Security is based on the standard assumptions of honest-but-curious, non-colluding servers having ability to generate random numbers. As expected, the protocol is costly compared to the one with zero-security. However, the additional communication costs are kept to a minimum (one round) and are negligible compared to those of SMC. With the RNS being doubly homomorphic, the paradigm of shattering and merging is generic and has potential to extend over to even more diverse data mining applications.

References

1. Duda, R., Hart, P.: Pattern Classification and Scene Analysis. John Wiley and Sons, Chichester (1973)
2. Fukunaga, K.: Introduction to Statistical Pattern Recognition. Academic Press, London (1990)
3. Cranor, L.F.: Internet privacy. Commun. ACM 42(2), 28–38 (1999)

4. Turow, J.: Americans and online privacy: The system is broken. Technical Report (2003)
5. Agrawal, R., Srikant, R.: Privacy-preserving data mining. SIGMOD 29(2), 439–450 (2000)
6. Lindell, Y., Pinkas, B.: Privacy preserving data mining. In: Bellare, M. (ed.) CRYPTO 2000. LNCS, vol. 1880, pp. 36–54. Springer, Heidelberg (2000)
7. Vaidya, J., Clifton, C.: Privacy-preserving data mining: why, how & when. Security & Privacy, 19–27 (2004)
8. Verykios, V.S., Bertino, E., Fovino, I.N., Provenza, L.P., Saygin, Y., Theodoridis, Y.: State-of-the-art in privacy preserving data mining. SIGMOD Rec. 33(1), 50–57 (2004)
9. Kargupta, H., Datta, S., Wang, Q., Sivakumar, K.: On the privacy preserving properties of random data perturbation techniques. In: ICDM, pp. 99–106 (2003)
10. Bunn, P., Ostrovsky, R.: Secure two-party k-means clustering. In: CCS, pp. 486–497 (2007)
11. Liu, K., Giannella, C., Kargupta, H.: A Survey of Attack Techniques on Privacy-Preserving Data Perturbation Methods. Privacy-Preserving Data Mining 34(15), 359–381 (2008)
12. Oliveira, S.R.M.: Privacy preserving clustering by data transformation. In: 18th Brazilian Symposium on Databases, pp. 304–318 (2003)
13. Goldreich, O.: The Foundations of Cryptography, vol. 2. Cambridge Univ. Press, Cambridge (2004)
14. Lindell, Y., Pinkas, B.: Secure multiparty computation for privacy-preserving data mining. Cryptology ePrint Archive, Report 2008/197 (2008)
15. Vaidya, J., Clifton, C.: Privacy-preserving k-means clustering over vertically partitioned data. In: KDD (2003)
16. Inan, A., Kaya, S.V., Saygin, Y., Savas, E., Hintoglu, A.A., Levi, A.: Privacy preserving clustering on horizontally partitioned data. Data Knowl. Eng. 63(3), 646–666 (2007)
17. Jagannathan, G., Wright, R.N.: Privacy-preserving distributed k-means clustering over arbitrarily partitioned data. In: KDD, pp. 593–599 (2005)
18. Upmanyu, M., Namboodiri, A.M., Srinathan, K., Jawahar, C.V.: Blind authentication: A secure crypto-biometric verification protocol. IEEE-Transactions on Information Forensics and Security, TIFS (to appear, 2010)
19. Orlandi, C., Piva, A., Barni, M.: Oblivious neural network computing via homomorphic encryption. In: EURASIP, pp. 1–10 (2007)
20. Jha, S., Kruger, L., Mcdaniel, P.: Privacy preserving clustering. In: di Vimercati, S.d.C., Syverson, P.F., Gollmann, D. (eds.) ESORICS 2005. LNCS, vol. 3679, pp. 397–417. Springer, Heidelberg (2005)
21. Rappe, D.: Homomorphic cryptosystems and their applications. Ph.D. dissertation, University of Dortmund (2004)
22. Gentry, C.: Fully homomorphic encryption using ideal lattices. In: STOC, pp. 169–178 (2009)
23. Asmuth, C., Bloom, J.: A modular approach to key safeguarding. IEEE Transactions on Information Theory 29, 208–210 (1983)
24. Upmanyu, M., Namboodiri, A.M., Srinathan, K., Jawahar, C.V.: Efficient privacy preserving video surveillance. In: International Conference on Computer Vision, ICCV (2009)
25. Shamir, A.: How to share a secret. ACM Communications 22(11), 612–613 (1979)
26. Beimel, A., Chor, B.: Universally ideal secret sharing schemes. In: Brickell, E.F. (ed.) CRYPTO 1992. LNCS, vol. 740, pp. 183–195. Springer, Heidelberg (1993)
27. Goldreich, O., Ron, D., Sudan, M.: Chinese remaindering with errors. IEEE Transactions on Information Theory 46 (2000)
28. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein., C.: The chinese remainder theorem. In: Introduction to Algorithms, pp. 873–876. MIT Press, McGraw-Hill (2001)
29. Mitchell, T.: Machine Learning. McGraw-Hill, New York (1997)
30. Ulman, Z.: Sign detection and implicit-explicit conversion of numbers in residue arithmetic. IEEE Transactions on Computers C-32 (1983)

A Fuzzy Threshold Based Modified Clustering Algorithm for Natural Data Exploration

Binu Thomas[1] and G. Raju[2]

[1] Lecturer,Marian College,
Dept. of Computer Applications
Kuttiikkanam, Kerala, India
binumarian@rediffmail.com
[2] Reader, Department of Information Technoogy,
Kannur University, Kannur, Kerala, India
kurupgraju@rediffmail.com

Abstract. Traditional supervised clustering methods require the user to provide the number of clusters before we start any data exploration. The data engineer also has to select the initial cluster seeds. In c-means clustering method, the performance efficiency of the algorithm depends mainly on the initial selection of number of clusters and cluster seeds. With the real world data, the initial selection of cluster count and centroids becomes a tedious task. In this paper we propose a modified clustering algorithm which works on the principles of fuzzy clustering. The method we propose is using a modified form of popular fuzzy c-means algorithm for membership calculation. The algorithm begins on the assumption that all the data points are initial centroids. . The clusters are continuously merged based on a threshold value until we get the optimum number of clusters. The algorithm is also capable of detecting the outliers The algorithm is tested with the data for Gross National Happiness (GNH) program of Bhutan and found to be highly efficient in segmenting natural data sets.

Keywords: Clustering, data mining, fuzzy c-means, fuzzy clustering, unsupervised clustering.

1 Introduction

The conventional clustering algorithms like k-means and fuzzy c-means require the number of clusters as an initial parameter for the expected performance[4]. In fuzzy c-means algorithm, the efficiency mainly depends on the selection of initial cluster seeds[2]. The intent of data mining and clustering techniques is to reveal hidden and previously unknown patterns from the collection of data. For an acceptable level of performance, the clustering algorithms have to deal with concerns like the number of clusters in the data, the uneven distribution of the data points, the initialization of the clustering algorithm, the large difference of cluster's sizes, the shape of the clusters, etc[5]. Among this, determining the optimal number of clusters and initial selection cluster seeds remain as a challenge in clustering.

H. Chen et al. (Eds.): PAISI 2010, LNCS 6122, pp. 167–172, 2010.
© Springer-Verlag Berlin Heidelberg 2010

This paper provides an overview of the supervised fuzzy clustering techniques and its limitations. Finally we propose an unsupervised fuzzy based clustering algorithm which is very efficient in handling natural data. Section 2 explains the basic principles of fuzzy clustering and its limitations. Section 3 Introduces the new modified algorithm. In section 4 we demonstrate the concepts presented in the paper. Section 5 concludes the paper.

2 C-Means Fuzzy Clustering Algorithm

Fuzzy c-means clustering involves two processes: the calculation of cluster centers and the assignment of points to these centers using a form of Euclidian distance[5]. This process is repeated until the cluster centers stabilize[3]. The algorithm is similar to k-means clustering in many ways and it also requires the initial cluster count (k) as a parameter. But it assigns a membership value to the data items for the clusters within a range of 0 to 1 [2]. So it incorporates fuzzy set's[10] concepts of partial membership and forms overlapping clusters to support it[1]. The algorithm needs a fuzzification parameter *m* in the range [1,n] which determines the degree of fuzziness in the clusters[9].

2.1 Modification to the C-Means Algorithm

The fuzzy c-means approach to clustering suffers from several constrains that affect the performance. We suggested a slight modification to the c-means fuzzy membership calculation[6] and we recommended a new expression (expr.1) for calculating the fuzzy memberships. The modified c-means fuzzy clustering method we proposed was used in devising a new unsupervised fuzzy clustering algorithm. The new algorithm we suggested starts with two initial clusters and uses two threshold values[7]. The algorithm uses the threshold values for generating new clusters and to merge the existing clusters to reach at an optimum number of clusters. So the algorithm works in two steps, the first step is used for creating new clusters and the second step is to merge these clusters to reach at an optimum number of clusters

$$\boxed{\mu_j(x_i)=\frac{n}{2}*\frac{\left(\frac{1}{d_{ji}}\right)^{\frac{1}{m-1}}}{\sum_{i=1}^{n}\left(\frac{1}{d_{ji}}\right)^{\frac{1}{m-1}}}} \tag{1}$$

where

$\mu_j(x_i)$: is the membership of x_i in the j^{th} cluster
d_{ji} : is the distance of x_i in cluster c_j
m : is the fuzzification parameter
n : is the number of data points

3 The New Modified Algorithm

Now we propose a modified algorithm[7], which can converge to the optimum number of clusters in a single step and it requires only one threshold value. The algorithm doesn’t require the user to provide the number of clusters as an initial parameter and it doesn't also require the user to initialize the cluster centers based on the general distribution of data. The algorithm works in a single step and initially it assumes that all the data points are cluster centers The method then uses one threshold value and it is the cluster center membership threshold (β). It is used to delete(merge) a cluster if it has a membership value greater than this in any of the other existing clusters. If a cluster center has a membership value greater than β with any other existing cluster center then it means that it is strongly associated with another cluster and one of the cluster centers can be deleted(both the clusters can be merged). The experiments showed that initializing β to .5(half of maximum fuzzy membership) produces desired outputs with natural data.

Pseudo code of new unsupervised clustering algorithm is given below:

```
initialize m=fuzzification parameter
initialize n=the number of data points
initialize C1=$x_1$,C2=$x_2$,---,Cn=xn
    //(initialize all the points as centroids)
initialize p=n
initialize $\beta$=.5
For i=1 to p
    Update $\mu_p(x_i)$ for each data points applying (expr 1)
    Find the sum of distances for all data points
    in $C_p$
    Sort cluster centers based on sum of distances
For each cluster center $C_i$ in the descending order of
    sum of distances
        If $\mu$(Ci)>= $\beta$ in any of the remaining cluster
            centers
     Delete $C_i$ and the corresponding membership values
     Update cluster center indexes
     p=p-1
```

After starting with n initial centroids, the fuzzy memberships of all the points in the other clusters are found. The sums of the distance to all the points from the cluster centers are also calculated. It is found that, a centroid situated at the center of a group of points will have minimum sum of distance to other data points. A centroid which is away from a group of points will have maximum sum of distance. The centroids are selected in the descending order of sum of distances for deletion. So the new centroids which are away from the groups of points(clusters) are considered for deletion first. Such centroids are deleted if it has a fuzzy membership of at least β in any other existing clusters. Otherwise the point is treated as an outlier point. When we continue cluster deletion process like this, only the new centroids situated at the middle of clusters and the extreme outlier points will be remaining. The algorithm ends by finding the natural cluster centers and extreme outliers in the dataset.

4 Illustration

To study the exact behavior of the algorithm we tested it with a natural dataset which was collected for Bhutan's Gross National Happiness(GNH) program.

Happiness and satisfaction are directly related with a community's ability to meet their basic needs and these are important factors in safeguarding their physical health. The unique concept of Bhutan's Gross National Happiness(GNH) depends on nine factors like health ecosystem, emotional well being, preservation of traditional culture etc.[12]. GNH regional chapter at Sherubtse College,Bhutan conducted a survey among 1311 villagers and the responses were converted into numeric values. For the analysis of the new method we took the attributes income and health index as shown in figure 1.

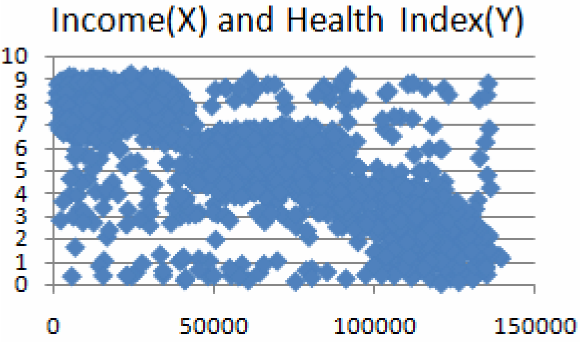


Fig. 1. The data collected for GNH analysis

As we can see from the data set, in Bhutan the low income group maintains better health than high income group since they are self sufficient in many ways. Like any other natural data his data set also contain many outliers which do not belong to any of the groups.

To start the data analysis, we applied the new algorithm and it converged to three cluster centers at (24351,7.83),(67965,4.56) and (123514.29,2.14). From the above figure it can be seen that these points exactly represent the centers of the natural groups.

We also analyzed the behavior of the algorithm for various values of β and the findings are plotted in a graph(figure 2). It is found that the algorithm produces better results for β threshold values within a range of .4 to .5.

For small values of β, most of the initial clusters will be merged to a single cluster. When the value reaches the maximum (one), only very few clusters are merged and it results in large number of clusters.

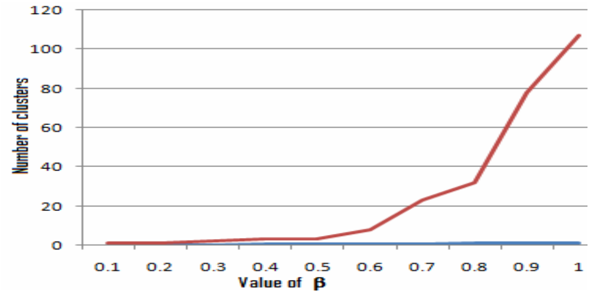


Fig. 2. Behavior of the algorithm with different values of β

5 Conclusion

Most of the supervised clustering algorithms are highly sensitive to the initial values of number of clusters and the centroids. The unsupervised clustering algorithm we proposed does not require any of these initial values and it can locate the clusters naturally based on the principles of fuzzy clustering. The algorithm produces better results with comparatively smaller data sets. The main limitation of the algorithm is that it is computationally expensive but it reduces the overheads of initializing clusters. The presently available tools for cluster count estimation are computationally expensive and they also do not guarantee the correct estimation. So the method we proposed becomes a better choice in exploring natural data sets.

References

1. Pal, K., Mitra, P.: Data Mining in Soft Computing Framework: A Survey. IEEE transactions on neural networks 13(1) (January 2002)
2. Au, W.H., Chan, K.C.C.: Classification with Degree of Membership: A Fuzzy Approach. In: Proceedings IEEE International Conference on Data Mining, ICDM 2001 (2001)
3. Halkidi, M.: Quality assessment and Uncertainty Handling in Data Mining Process, `http://citeseer.ist.psu.edu/halkidi00quality.html`
4. Inmon, W.H.: The data warehouse and data mining. Commun., ACM 39, 49–50 (1996)
5. Fayyad, U., Uthurusamy, R.: Data mining and knowledge discovery in databases. ACM Commun. 39, 24–27 (1996)
6. Thomas, B., Raju, G.: A Modified c-means algorithm for Natural Data Exploration. In: WASET International Conference on Knowledge Management (ICKM), January 2009, vol. 49 (2009) ISSN 2070-3724
7. Thomas, B., Raju, G.: A Fuzzy Threshold Based Unsupervised Clustering Algorithm for Natural Data Exploration. In: Proceedings of International Conference on Database and Data Mining (ICDDM) (June 2010)
8. Keith, C.C., Wai-Ho Au, C., Choi, B.: Mining Fuzzy Rules in A Donor Database for Direct Marketing by A Charitable Organization. In: Proceedings. First IEEE International Conference on Cognitive Informatics, pp. 239–246 (2002)
9. Cox, E.: Fuzzy Modeling and Genetic Algorithms for Data Mining and Exploration. Elsevier, Amsterdam (2005)

10. Klir, G.J., Folger, T.A.: Fuzzy Sets, Uncertainty and Information. Prentice Hall, Englewood Cliffs (1988)
11. Han, J., Kamber, M.: Data Mining Concepts and Techniques. Elsevier, Amsterdam (2003)
12. Donnelly, S.: How Bhutan Can Develop and Measure GNH, `http://www.bhutanstudies.org.bt/seminar/0402-gnh/GNH-papers-1st_18-20.pdf`

Author Index

Abraham, Sajimon 31

Ben Yahia, Sadok 73
Bhat, Veena H. 105
Bondy, Julian 1
Bowerman, Chris 1
Brahmi, Imen 73

Chang, Weiping 22, 28
Chen, Patrick S. 44
Choi, Yoonjung 140
Chou, Shihchieh 113

Erdem, Zeki 1

Hota, Chittaranjan 65
Hsing, Tai-Ping 113
Hsu, Chia-Shih 44

Jawahar, C.V. 154
Jung, Yuchul 140

Kittiphattanabawon, Nichnan 126

Lal, Paulose Sojan 31
Lin, Chien-Lung 15
Li, Pengcheng 98
Liu, Yong-Kuei 44

Mohanty, Vakul 65
Moliya, Dhaval 65
Myaeng, Sung-Hyon 140

Namboodiri, Anoop M. 154
Nantajeewarawat, Ekawit 126

Ozgul, Fatih 1

Patnaik, L.M. 105
Pawar, Pramod S. 52
Poncelet, Pascal 73

R., Venugopal K. 105
Rajarajan, Muttukrishnan 65
Raju, G. 167
Ratan, Ram 88

S., Krishna 105
Shen, Yanping 77
Shenoy, P. Deepa 105
Srinathan, Kannan 154
Srinivasa, Srinath 52
Su, Chingwei 22

Takagi, Tsuyoshi 77, 98
Theeramunkong, Thanaruk 126
Thomas, Binu 167

Upmanyu, Maneesh 154

Wang, Hao 98
Wang, Jau-Hwang 15

Yang, Bo 77, 98

Zhang, Mingwu 77, 98
Zhang, Wenzheng 77