Hsinchun Chen
Christopher C. Yang
Michael Chau
Shu-Hsing Li (Eds.)

LNCS 5477

Intelligence and Security Informatics

Pacific Asia Workshop, PAISI 2009
Bangkok, Thailand, April 2009
Proceedings

Springer

Lecture Notes in Computer Science 5477

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
University of Dortmund, Germany

Madhu Sudan
Massachusetts Institute of Technology, 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 Christopher C. Yang
Michael Chau Shu-Hsing Li (Eds.)

Intelligence and Security Informatics

Pacific Asia Workshop, PAISI 2009
Bangkok, Thailand, April 27, 2009
Proceedings

Springer

Volume Editors

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

Christopher C. Yang
Drexel University, Philadelphia, PA, USA
E-mail: chris.yang@ischool.drexel.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

Library of Congress Control Number: Applied for

CR Subject Classification (1998): H.4, H.3, C.2, H.2, D.4.6, K.4.1, K.5, K.6

LNCS Sublibrary: SL 4 – Security and Cryptology

ISSN 0302-9743
ISBN-10 3-642-01392-9 Springer Berlin Heidelberg New York
ISBN-13 978-3-642-01392-8 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 2009
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12657988 06/3180 5 4 3 2 1 0

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 was started in 2003 and the first four meetings were held in the United States. 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, with over 100 contributors and participants from all over the world. PAISI 2007 (http://isi.se.cuhk.edu.hk/2007/) was then held in Chengdu, China and PAISI 2008 (http://isi.se.cuhk.edu.hk/2008/) was held in Taiwan. These ISI conferences have 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 including 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, and information assurance, among others. We continued the stream of ISI conferences by organizing the 2009 Pacific Asia Workshop on ISI (PAISI 2009) in conjunction with the Pacific Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2009). PAISI 2009 was hosted by the University of Arizona, Drexel University, the University of Hong Kong, and the National Taiwan University. In addition to traditional ISI topics, we also broadened our scope to include research relating to enterprise risk management and information systems security. The one-day program included a keynote speech by Bhavani Thuraisingham and presentations of ten long papers and seven short papers. 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 the workshop Program Committee members, who provided valuable and constructive review comments.

April 2009

Hsinchun Chen
Christopher C. Yang
Michael Chau
Shu-Hsing Li

Organization

Workshop Co-chairs

Hsinchun Chen	The University of Arizona, USA
Christopher C. Yang	Drexel University, USA
Michael Chau	The University of Hong Kong, HKSAR China
Shu-Hsing Li	National Taiwan University, Taiwan ROC

Program Committee Members

Kuo-Tay Chen	National Taiwan University, Taiwan ROC
Tsai-Jyh Chen	National Chengchi University, Taiwan ROC
Reynold Cheng	The University of Hong Kong, HKSAR China
Vladimir Estivill-Castro	Griffith University, Australia
Uwe Glasser	Simon Fraser University, Canada
Raymond Hsieh	California University of Pennsylvania, USA
Hsiang-Cheh Huang	National University of Kaohsiung, Taiwan ROC
Ming-Hui Huang	National Taiwan University, Taiwan ROC
Shi-Ming Huang	National Chung Cheng University, Taiwan ROC
Eul Gyu Im	Hanyang University, Republic of Korea
Paul W.H. Kwan	The University of New England, Australia
Kai Pui Lam	The Chinese University of Hong Kong, HKSAR China
Wai Lam	The Chinese University of Hong Kong, HKSAR China
Sheau-Dong Lang	University of Central Florida, USA
Mark Last	Ben-Gurion University of the Negev, Israel
Ickjai Lee	James Cook University, Australia
You-lu Liao	Central Police University, Taiwan ROC
Ee-peng Lim	Nanyang Technological University, Singapore
Hongyan Liu	Tsinghua University, China
Anirban Majumdar	SAP Research
Byron Marshall	Oregon State University, USA
Dorbin Ng	The Chinese University of Hong Kong, HKSAR China
Jialun Qin	The University of Massachusetts Lowell, USA
Dmitri Roussinov	University of Strathclyde, UK
Raj Sharman	State University of New York, Buffalo, USA
Aixin Sun	Nanyang Technological University, Singapore
Paul Thompson	Dartmouth College, USA
Alan Wang	Virginia Tech University, USA

Jau-Hwang Wang	National Central Police University, Taiwan ROC
Ke Wang	Simon Fraser University, Canada
Philips Wang	City University of Hong Kong, HKSAR China
Shiuh-Jeng Wang	Central Police University, Taiwan ROC
Chih-Ping Wei	National Tsing Hua University, Taiwan ROC
Jennifer Xu	Bentley College, USA
Wei Yan	Trend Micro
Justin Zhan	Carnegie Mellon University, USA
Yilu Zhou	George Washington University, USA
William Zhu	The University of Auckland, New Zealand

Table of Contents

Keynote

Terrorism Informatics and Crime Analysis

Enterprise Risk Management

Emergency Response and Surveillance

Information Access and Security

Data and Text Mining

Building a Geosocial Semantic Web for Military Stabilization and Reconstruction Operations

Bhavani Thuraisingham, Murat Kantarcioglu, and Latifur Khan

The University of Texas at Dallas

The United States and its Allied Forces have had tremendous success in combat operations. This includes combat in Germany, Japan and more recently in Iraq and Afghanistan. However not all of our stabilization and reconstruction operations (SARO) have been as successful. Recently several studies have been carried out on SARO by National Defense University as well as for the Army Science and Technology. One of the major conclusions is that we need to plan for SARO while we are planning for combat. That is, we cannot start planning for SARO after the enemy regime has fallen. In addition, the studies have shown that security, power and jobs are key ingredients for success during SARO. For example, it is essential that security be maintained. Furthermore, it is important to give powerful positions to those from the fallen regime provided they are trustworthy. It is critical that investments are made to stimulate the local economies. The studies have also analyzed the various technologies that are needed for successfully carrying out SARO which includes sensors, robotics and information management. In our research we are focusing on the information management component for SARO. As stated in the work by the Naval Postgraduate School, we need to determine the social, political and economic relationships between the local communities as well as determine who the important people are. This work has also identified the 5Ws (Who, When, What, Where and Why) and the (H).

To address the key technical challenges for SARO, our goal is to utilize the extensive research we have carried out at the University of Texas at Dallas in geospatial information management, social networking and knowledge discovery and develop novel technologies for SARO. In particular, we have defined a Life cycle for SARO and subsequently developing a Temporal Geosocial Service Oriented Architecture System (TGS-SOA) that utilizes Temporal Geosocial Semantic Web (TGS-SW) technologies for managing this lifecycle. We are developing techniques for representing temporal geosocial information and relationships, integrating such information and relationships, querying such information and relationship and finally reasoning about such information and relationships so that the commander can answer questions related to the 5Ws and H.

The presentation will discuss the challenges of SARO and our solutions to SARO that integrates semantic web, social networking, knowledge discovery, geospatial and security and privacy technologies. We will discuss our approach to developing a geosocial semantic web for SARO. Our project has tremendous applications not only in SARO but for many other applications including in emergency response and public health.

H. Chen et al. (Eds.): PAISI 2009, LNCS 5477, p. 1, 2009.
© Springer-Verlag Berlin Heidelberg 2009

Criminal Cross Correlation Mining and Visualization

Peter Phillips and Ickjai Lee

Discipline of IT
James Cook University
Australia
{peter.phillips,ickjai.lee}@jcu.edu.au

Abstract. Criminals are creatures of habit and their crime activities are geospatially, temporally and thematically correlated. Discovering these correlations is a core component of intelligence-led policing and allows for a deeper insight into the complex nature of criminal behavior. A spatial bivariate correlation measure should be used to discover these patterns from heterogeneous data types. We introduce a bivariate spatial correlation approach for crime analysis that can be extended to extract multivariate cross correlations. It is able to extract the top-k and bottom-k associative features from areal aggregated datasets and visualize the resulting patterns. We demonstrate our approach with real crime datasets and provide a comparison with other techniques. Experimental results reveal the applicability and usefulness of the proposed approach.

1 Introduction

Since criminals are creatures of habit, law enforcement agencies can be more effective if they learn from historical data to better understand perpetrators habits and locations they choose to commit crimes. Police and policy makers need an intelligent crime analysis machine that is able to capture the connections that exist between places and events based on past crimes. These patterns can then be used to characterize criminal behavior and discover where, when and why particular crimes are likely to occur. Discovering correlations between crime and spatial features is a core component of intelligence-led policing, and allows for a deeper insight into the complex nature of criminal behavior. Crime activities are geospatially, temporally and thematically correlated therefore a variety of factors can contribute to the formulation of crime. These factors need to be considered when interpreting crime datasets.

Several crime data mining techniques have been developed over recent years [1,2,3,4], however reasoning about crime datasets has received less attention [5,6]. One of the drawbacks of these approaches is that they can typically only reason about positive associative features whereas negative associative features can be just as important. Several works using spatial association rules mining have been proposed in order to mine spatial associations in geospatial databases [7,8,9]. The

H. Chen et al. (Eds.): PAISI 2009, LNCS 5477, pp. 2–13, 2009.
© Springer-Verlag Berlin Heidelberg 2009

main drawback of these approaches is that they capture point-to-point association but ignore spatial association.

The increasing availability of heterogeneous datasets, such as those incorporating socio-economic and socio-demographic factors, geospatial features and crime datasets, has increased the need for intelligent analysis. To discover correlations in these heterogeneous data types, a spatial bivariate correlation measure should be used. Standard bivariate techniques do not account for spatial relations in their computation and are therefore not well suited to identifying spatial structures [10]. Previous efforts in developing a bivariate association measure have been proposed by Wartenberg [11] who based their approach on Moran's I statistic. However, Lee [12] identifies drawbacks with Wartenberg's approach and instead proposes his own bivariate spatial association measure.

In this paper we rank crime and geospatial feature datasets by their spatial co-patterning (similarity) using bivariate correlation measures. This results in the discovery of patterns in the form $crimeX$ is more similar to $featureY$ than $featureX$. We also show that while Lee's L index [12] is an effective spatial correlation measure for crime analysis, it achieves inconsistent results when used with areal aggregated datasets over regions of irregular size and shape. To overcome this we present a more suitable weights matrix where the spatial lag better represents neighboring regions of irregular size. Using this weights matrix, we are then able to extract the top-k and bottom-k associative features. We also present a graph based visualization method that allows the user to easily extract patterns of interest.

Research in the area of crime analysis and reasoning, such as the approach outlined here, can be applied to a wide range of other applications that may lead to economic, social and environmental advantages. This technique could be applied to geospatial datasets in several fields including disaster management, epidemiology, business intelligence, geology, environmental monitoring, marketing and e-commerce. For example, in epidemiology, interesting cross correlations between cancer hotspots and environmental features can be identified.

In Section 2 we review existing bivariate spatial association measures, and perform comparisons between them. Section 3 compares two popular correlation measures and describes the problem statement. In Section 4 we detail the special properties of areal aggregated crime data and show how the choice of the spatial weights matrix affects the bivariate correlation measure. Section 5 provides experimental results with real crime datasets and our visualization technique. We conclude with final remarks and ideas for future work in Section 6.

2 Bivariate Spatial Association

Spatial association is the degree to which a set of univariate observations are similarly arranged over space. It quantifies the distribution of patterns among a dataset, with a strong spatial association occurring when two distributions are similar, weak association describes little similarity and no association occurs when two distributions are dissimilar. Bivariate spatial association measures quantify the spatial relations among many variables in a set of observations.

Many processes, including crime, involve more than one variable, so allowing for their dependence on each other is essential in modeling and in understanding their covariance [13]. Hubert *et al.* [14] make a distinction between the relationship within a pair at each location (point-to-point association) and the relationship between distinct pairs across locations (spatial association). Pearson's r is a common point-to-point association measure, while spatial association is often measured by Moran's I. Pearson's correlation coefficient r for variables X and Y is computed by:

$$r_{X,Y} = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2}\sqrt{\sum (y_i - \overline{y})^2}}, \tag{1}$$

and Moran's I is given by:

$$I_X = \frac{\sum_i \sum_j w_{ij}(x_i - \overline{x})(x_j - \overline{x})}{\sum (x_i - \overline{x})^2}, \tag{2}$$

where i and j are two spatial objects and w_{ij} is a general spatial weights matrix of their locational similarity. The choice of the weights matrix is a critical step in computing the spatial association [10,15].

Several works using spatial association rules mining have been proposed in the data mining community in order to mine spatial associations in massive geospatial databases [7,8,9]. In general, these approaches extend traditional association rule mining [16,17] to the spatial and temporal context. One drawback of these approaches is that they typically capture point-to-point association but ignore spatial association. It is argued that to capture spatial co-patterning, a bivariate spatial association measure should be the combination of both the point-to-point association and spatial association measures [12]. That is, the bivariate measure should be a composite of the univariate spatial associations of two variables and their point-to-point association in a certain form.

There are two main approaches in this regard [11,12]. Wartenberg [11] developed a multivariate extension of Moran's I univariate spatial autocorrelation measure to account for the spatial dependence of data observations and their multivariate covariance simultaneously. Wartenberg's approach can be defined as:

$$I_{X,Y} = \frac{\sum_i (x_i - \overline{x})(\widetilde{y}_i - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2}\sqrt{\sum (y_i - \overline{y})^2}} \cong \sqrt{SSS_Y} \cdot r_{X,\widetilde{Y}}. \tag{3}$$

Lee [12] utilizes the concept of spatial lag (SL), which is defined as the composed weighted averages of neighbors defined by the spatial weights matrix, and is given as:

$$\widetilde{x} = \sum_j w_{ij} x_j \tag{4}$$

The SL is then used to introduce the concept of a spatial smoothing scalar (SSS) that can reveal substantive information about the spatial clustering of a variable. The SSS ranges from 0 to 1, where more clustered variables have higher SSS values. The SSS is given by:

$$SSS_x = \frac{n}{\sum_i (\sum_j w_{ij})^2} \cdot \frac{\sum_i (\sum_j w_{ij}(x_j - \overline{x}))^2}{\sum (x_i - \overline{x})^2}. \tag{5}$$

Lee's L index between two variables can then be defined as the Pearson's correlation coefficient between their SL vectors multiplied by the square root of the product of their SSSs:

$$L_{X,Y} = \sqrt{SSS_X} \cdot \sqrt{SSS_Y} \cdot r_{\widetilde{X},\widetilde{Y}}. \tag{6}$$

3 Correlation Measures and Problem Statement

We evaluate both Lee's L and Wartenberg's I bivariate correlation measures with a synthetic dataset that shows increasing dissimilarity. We use the same base region of 37 regular sized hexagons that Lee [12] uses to illustrate his bivariate spatial correlation measure. Lee compares his L index to Wartenberg's using simple dataset patterns and argues that Cross-MC should not be used as a bivariate spatial association measure.

Figure 1 starts with 4 clusters in $dataset_a$, and for each subsequent dataset we remove 1 cluster. As clusters are removed, $dataset_{b-d}$ become less similar to the original $dataset_a$. The density for non cluster regions are uniformly distributed random values and are the same among datasets. Table 1 and Fig. 1(e) illustrate the expected result; when we rank the similarity for $dataset_a$, the bivariate correlation measure decreases with the number of clusters removed. For this dataset, both Lee's L and Wartenberg's approach show correct results, however only Lee's L shows the expected linear decrease in correlation value.

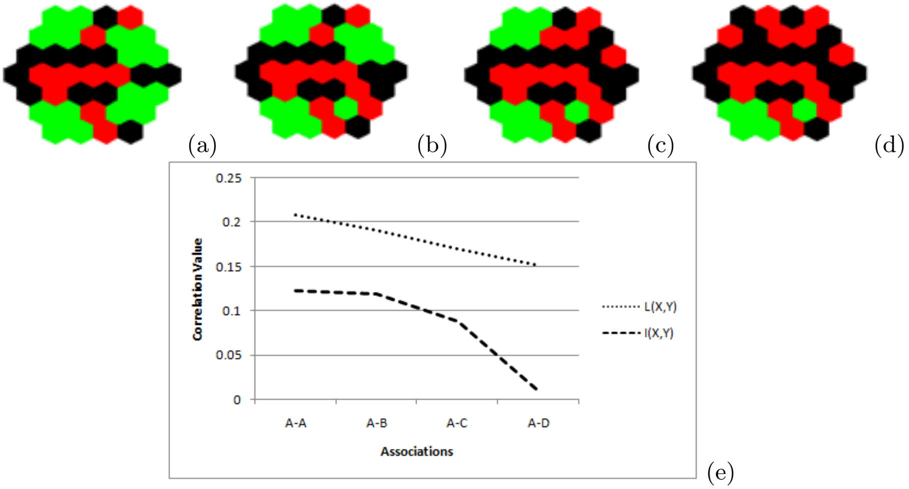


Fig. 1. Synthetic hexagon dataset: (a-d) Datasets $a - d$; (e) Comparison of Lee's L and Wartenberg's I

Table 1. Cross correlations from synthetic hexagon dataset

	SSS_X	SSS_Y	$r_{\tilde{X},\tilde{Y}}$	$r_{X,Y}$	$L_{X,Y}$	$I_{X,Y}$
a-a	0.208	0.208	1.000	1.000	0.208	0.123
a-b	0.208	0.236	0.863	0.896	0.191	0.119
a-c	0.208	0.336	0.642	0.740	0.170	0.089
a-d	0.208	0.387	0.532	0.595	0.151	0.009

4 Shared Border Aware Weight Matrix

Crime, census and geospatial features are often recorded for regions of irregular size and shape. Crime datasets can be aggregated to suburb or police districts and census data using census tracts or collection districts. When the study region contains such irregular regions, the choice of weight matrix is a critical step in the computation of spatial association [15].

Figure 2 shows synthetic areal aggregated datasets over suburbs from Brisbane, Australia. Regions of grey color are low density while green colored regions are high density. This is a subset of the base regions we use in our real world experiments in Section 5. To allow for a deeper insight into the complex question of crime analysis we need to discover crime and spatial features that exhibit strong correlation (similarity). From visual analysis of our synthetic datasets, it can be argued that the ranking of similarity for $dataset_a$ is $dataset_a - dataset_b, dataset_c, dataset_d$. That is, $dataset_a$ is more similar to $dataset_b$ than $dataset_c$ and $dataset_d$. This is because these datasets show a similar density (green high density regions surrounded by grey low density regions) in a similar spatial neighborhood.

Table 2 shows that Lee's L index incorrectly determines that the similarity ranking is $dataset_a - dataset_d, dataset_c, dataset_b$. This is because Lee uses a simple weights matrix that is defined as the row standardized version of the binary connectivity matrix where elements that are neighbors have a value of 1 or otherwise a value of 0. Each neighbor is given the same weighting when calculating the spatial lag vector but with irregular regions this is often not desirable. For example in Fig. 2, $dataset_b$ and $dataset_c$ are both neighbors (share

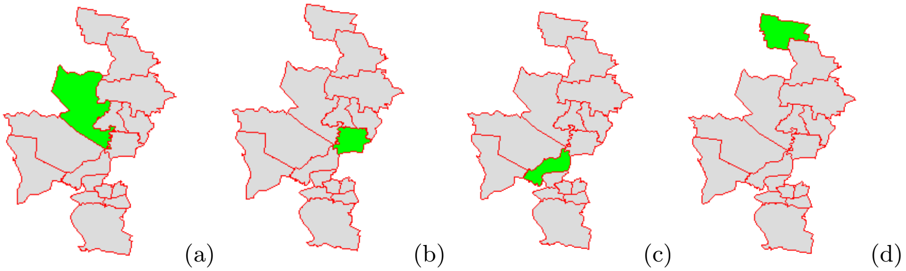


Fig. 2. Synthetic irregular region dataset: (a-d) Datasets $a - d$

Table 2. Cross correlations from irregular dataset

	SSS_X	SSS_Y	$r_{\tilde{X},\tilde{Y}}$	$r_{X,Y}$	$L_{X,Y}$
a-a	1.000	1.000	1.000	1.000	1.000
a-b	1.000	0.865	0.028	0.833	0.027
a-c	1.000	1.000	0.040	0.833	0.041
a-d	1.000	0.741	0.448	0.833	0.388

Table 3. Cross correlations from irregular dataset with modified weights matrix

	SSS_X	SSS_Y	$r_{\tilde{X},\tilde{Y}}$	$r_{X,Y}$	$L_{X,Y}$
a-a	0.994	0.994	1.000	1.000	0.994
a-b	0.994	0.817	0.152	0.833	0.137
a-c	0.994	0.929	0.112	0.833	0.108
a-d	0.994	0.712	-0.107	0.833	-0.0901

a border) with $dataset_a$, however $dataset_c$ shares only a very small border. The spatial weights matrix should be defined so that the common border reflects the weighting and thus the spatial lag. If i and j are neighbors, we define w as:

$$w(i,j) = sharedBorder_{ij}/totalBorder_i, \tag{7}$$

where $sharedBorder_{ij}$ is the shared border between i and j and $totalBorder_i$ is the total border of i.

Given this new spatial weights matrix, Table 3 shows that Lee's L index correctly identifies the similarity ranking of $dataset_a$ as $dataset_a - dataset_b$, $dataset_c$, $dataset_d$. Discovering patterns of similar crime and geospatial/census features is a key component to intelligence-led policing and crime analysis.

5 Experimental Results

This section shows the results collected from real crime datasets of urban suburbs of Brisbane, Australia. A steadily growing trend in recorded violent crime in Australia [18] has been a major concern not only to policing agencies, but also tourism agencies and the public. The Brisbane study region (Fig. 3) is highly dynamic and active and continues to experience significant and sustained population growth and an increase in various criminal activities [19]. The Queensland Police Service (QPS) releases crime data in areal aggregated format due primarily to privacy concerns. We combine these crime datasets with spatial feature datasets and census datasets so that associative patterns can be discovered. The smallest geographic unit at which census data is available from the Australian Bureau of Statistics (ABS) is the Collection District. The ABS produces statistics for other geographic areas by allocating collection districts to each spatial unit in the particular classification [20].

We use a total of 108 datasets in this experiment; 38 crime datasets, 7 geographic features (caravan parks, railway stations, reserves, schools, hospitals,

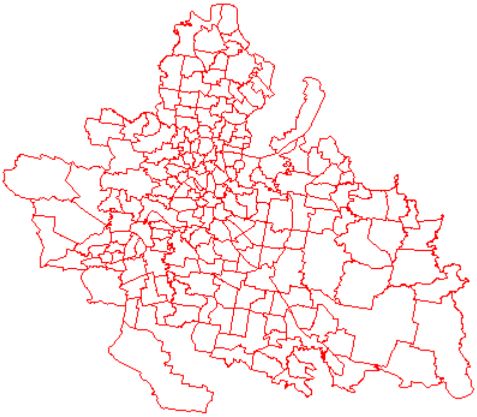


Fig. 3. Brisbane study region

Table 4. Cross correlations from the Brisbane crime dataset

top-k	$L_{X,Y}$
Unlawful Entry With Intent - UNOCCUPIED	0.603
Unlawful Entry With Intent - AGED_25_29	0.580
Unlawful Entry With Intent - UNEMPLOYED	0.564
Other Stealing - OVERSEAS_VISITOR	0.551
Other Theft - OVERSEAS_VISITOR	0.535
bottom-k	
Arson - AGED_85_89	-0.154
Arson - AGED_90_94	-0.152
Liquor (excl. Drunkenness) - SEPARATE_HOUSE	-0.139
Arson - FLAT_UNIT	-0.139
Rape and Attempted Rape - reserve	-0.135

university/colleges and parks) and 63 census classifications. The crime dataset from the QPS has three main categories: personal safety (offences against person), property security (offences against property) and other offences. The census classifications we use include age, dwelling structure, employment status, weekly income, level of education completed and household mobility indicator (change of address). The study region encompasses 186 suburbs of Brisbane that had crime and census data available.

We utilize Lee's L index with the shared border length weights matrix as described in Section 4. We mine the dataset to extract the top-k and bottom-k spatial associative patterns involving salient features or census classifications (for this experiment $k = 5$). We restrict the patterns to those involving at least one type of crime. From the extracted associations shown in Table 4 we can see that there is a strong correlation between the crime *Unlawful Entry With Intent* and *Unoccupied Dwellings* and *Unemployed 25-29 year old persons*. It can be seen that there is also a weak association between *Arson* and *persons aged 85-94*. This information can then be used to further investigate the cause of these specific associative patterns and also allow for targeted crime prevention efforts.

5.1 Visualization

While the associations shown in Table 4 may be common sense, there may be interesting, unknown patterns in the larger subset of top-k and bottom-k associative patterns. The problem becomes one of information overload; how can the user find interesting patterns hidden amongst the other patterns.

Figure 4 shows the visualization that we developed to help the user extract interesting patterns. For this example we use the same datasets as described in Section 5 and extract the top-50 results. Each node (feature) is labeled and its size is a depiction of the correlation strength, where the size of the circle represents the largest L index score of its relationships. The edges between nodes in our visualization depict the top-k associative patterns. The correlation strength is depicted by the edge thickness and color.

Figure 4 clearly shows that the top-50 associative patterns form two clusters, one revolving around *Overseas Visitors, Other Dwelling - Improvised home* and the second cluster around *Unemployed, Age 25-29, Unoccupied Dwelling, Unlawful Entry With Intent, Unlawful Entry Without Violence.* The visualization environment is dynamic so that the user can move and delete any associations that are not of interest. Figure 5 shows the visualization of the bottom-50 associative patterns. In this case the, the visualization easily allows the user to see that *Arson* is not caused by *People aged 80-100*, locations with *Flat or Units* and locations of *Universities and Colleges.*

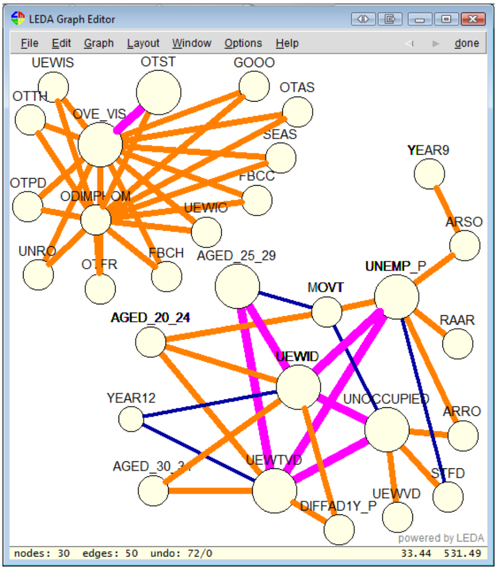


Fig. 4. Visualization of top-50 associative patterns

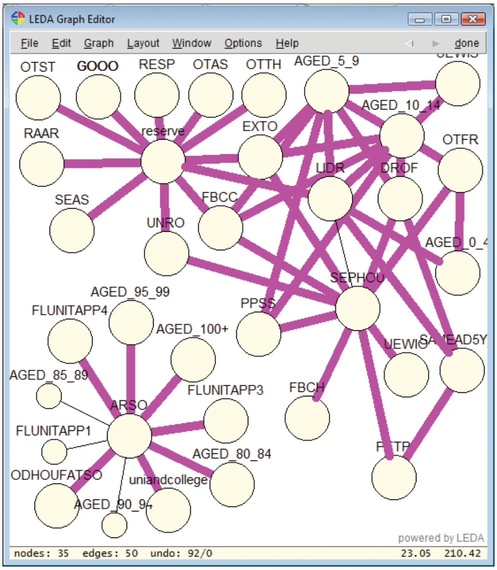


Fig. 5. Visualization of bottom-50 associative patterns

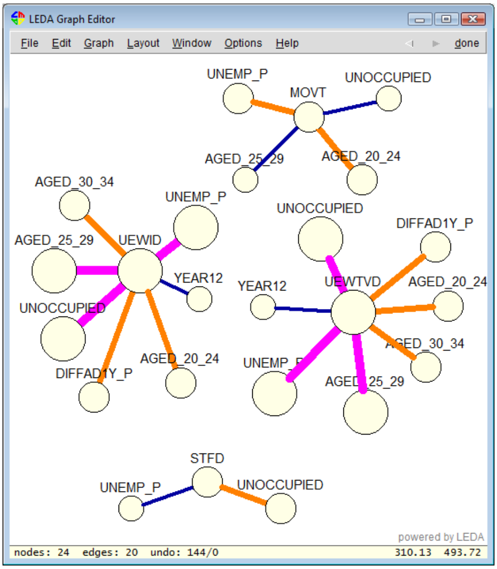


Fig. 6. Visualization of selected top-k crime associative patterns

Users can also select specific crime features to visualize from the top-k/ bottom-k results. Figure 6 shows the associative patterns when the crimes *Unlawful Entry With Intent, Unlawful Entry Without Violence, Motor Vehicle Theft and Stealing from Dwellings* from the top-50 are selected. In this visualization, duplicate geospatial and census features are not removed. As part of a crime decision support system these associative patterns can be used to further investigate and explain the behavior of crime within the study region.

5.2 Comparison with Density Tracing

We compare the spatial associative patterns discovered by the bivariate spatial association approach with our previous Density Tracing approach [21]. Density tracing allows autonomous exploratory analysis and knowledge discovery in areal aggregated crime datasets by transforming the dataset into a density trace that shows density change between regions. The results from the two approaches are not directly comparable, however the associative patterns in the form $crimeX$ is more similar to $featureY$ than $featureX$ are comparable.

Table 5. Cross correlations using Density Tracing

Patterns	*dissimilarity*
Unlawful Entry With Intent - UNEMPLOYED	0.425599
Unlawful Entry With Intent - AGED_90_94	0.45303
Unlawful Entry With Intent - AGED_20_24	0.459116
Other Stealing - OVERSEAS_VISITOR	0.115436
Other Theft - OVERSEAS_VISITOR	0.116533

We examine the top-5 patterns from Table 4 against Density Tracing. We use each crime in the top-5 as a reference feature f in the Density Tracing algorithm and select the most similar non-crime feature. For *Unlawful Entry With Intent* we report the three most similar features. From the results in Table 5 it can be seen that for *Other Stealing* and *Other Theft* both approaches report the same most similar associative patterns. The results for *Unlawful Entry With Intent* are due to the differences in the way the two approaches calculate similarity.

6 Final Remarks

Since crime activities are geospatial phenomena, they are geospatially, thematically and temporally correlated. Thus, crime datasets must be interpreted and analysed in conjunction with various factors that can contribute to the formulation of crime. We propose a bivariate spatial association approach for crime analysis using an enhanced Lee's L index [12] to extract the top-k and bottom-k associative features. We have shown why the choice of spatial weight matrix is an important consideration for spatial association analysis of crime datasets of irregular size and shape. In addition, we introduced a visualization approach that helps users find interesting patterns.

This research is part of an ongoing project with the aim to build a crime knowledge discovery machine, as a crime decision support system for large areal-aggregated crime datasets, that explains the behavior of crime with the first order effect of crime (concentrations and deviations) and the second order effect of crime (links and associations). Future work includes incorporating temporal data into the bivariate approach. We wish to compare the results from this study to census and crime data for other years. We are also investigating graph based implementations for computational and memory efficiency. A comparison between this technique and an Association Rules Mining approach such as [8] is also planned.

References

1. Chen, H., Atabakhsh, H., Zeng, D., Schroeder, J., Petersen, T., Casey, D., Chen, M., Xiang, Y., Daspit, D., Nandiraju, S., Fu, S.: Coplink: visualization and collaboration for law enforcement. In: Proceedings of the 2002 annual national conference on Digital government research, pp. 1–7 (2002)
2. Craglia, M., Haining, R., Wiles, P.: A Comparative Evaluation of Approaches to Urban Crime Pattern Analysis. Urban Studies 37(4), 711–729 (2000)
3. Hirschfield, A., Brown, P., Todd, P.: GIS and the Analysis of Spatially-Referenced Crime Data: Experiences in Merseyside. U. K. Journal of Geographical Information Systems 9(2), 191–210 (1995)
4. Ratcliffe, J.: The Hotspot Matrix: A Framework for the Spatio-temporal Targeting of Crime Reduction. Police Practice and Research 5, 5–23 (2004)
5. Chen, H., Chung, W., Xu, J.J., Wang, G., Qin, Y., Chau, M.: Crime Data Mining: A General Framework and Some Examples. Computer 37(4), 50–56 (2004)
6. Oatley, G., Ewart, B., Zeleznikow, J.: Decision Support Systems for Police: Lessons from the Application of Data Mining Techniques to Soft Forensic Evidence. Artificial Intelligence and Law 14(1), 35–100 (2006)
7. Koperski, K., Han, J.: Discovery of Spatial Association Rules in Geographic Information Databases. In: Proceedings of the 4th International Symposium on Large Spatial Databases, Portland, Maine. LNCS, pp. 47–66. Springer, Heidelberg (1995)
8. Lee, I., Phillips, P.: Urban crime analysis through areal categorized multivariate associations mining. Applied Artificial Intelligence 22(5), 483–499 (2008)
9. Shekhar, S., Huang, Y.: Discovering Spatial Co-location Patterns: A Summary of Results. In: Jensen, C.S., Schneider, M., Seeger, B., Tsotras, V.J. (eds.) SSTD 2001. LNCS, vol. 2121, pp. 236–256. Springer, Heidelberg (2001)
10. Dray, S., Saïd, S., Débias, F.: Spatial ordination of vegetation data using a generalization of Wartenberg's multivariate spatial correlation. Journal of Vegetation Science 19, 45–56 (2008)
11. Wartenberg, D.: Multivariate spatial correlation: A method for exploratory geographical analysis. Geographical Analysis 17, 263–283 (1985)
12. Lee, S.: Developing a bivariate spatial association measure: An integration of Pearson's r and Moran's I. Journal of Geographical Systems 3(4), 369–385 (2001)
13. Morrison, D.F.: Multivariate Statistical Methods, 2nd edn. McGraw-Hill, New York (1976)

14. Hubert, L.J., Golledge, R.G., Costanzo, C.M., Gale, N.: Measuring association between spatially defined variables: an alternative procedure. Geographical Analysis 17, 36–46 (1985)
15. Tiefelsdorf, M., Griffith, D.A., Boots, B.: A variance-stabilizing coding scheme for spatial link matrices. Environment and Planning A 31(1), 165–180 (1999)
16. Agrawal, R., Imielinski, T., Swami, A.N.: Mining Association Rules between Sets of Items in Large Databases. In: Buneman, P., Jajodia, S. (eds.) Proceedings of the ACM SIGMOD 1993 International Conference on Management of Data, pp. 207–216. ACM Press, Washington (1993)
17. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, San Francisco (2000)
18. Australian Institute of Criminology: Comparing International Trends in Recorded Violent Crime. In: Crime Facts Info No. 115 (2006), http://www.aic.gov.au/publications/cfi/cfi115.html
19. Murray, A.T., McGuffog, I., Western, J.S., Mullins, P.: Exploratory Spatial Data Analysis Techniques for Examining Urban Crime. British Journal of Criminology 41, 309–329 (2001)
20. Australian Bureau of Statistics: Australian Standard Geographical Classification (ASGC) (2005)
21. Phillips, P., Lee, I.: Areal Aggregated Crime Reasoning through Density Tracing. In: International Workshop on Spatial and Spatio-temporal Data Mining in conjunction with IEEE International Conference on Data Mining, Omaha, NE, USA (October 2007)

A Cybercrime Forensic Method for Chinese Web Information Authorship Analysis

Jianbin Ma[1], Guifa Teng[1], Yuxin Zhang[1], Yueli Li[1], and Ying Li[2]

[1] College of Information Science and Technology, Agricultural University of Hebei,
Baoding 071001, China
[2] College of Sciences, Agricultural University of Hebei, Baoding 071001, China
{majianbin}@hebau.edu.cn

Abstract. With the increasing popularization of the Internet, Internet services used as illegal purposes have become a serious problem. How to prevent these phenomena from happening has become a major concern for society. In this paper, a cybercrime forensic method for Chinese illegal web information authorship analysis was described. Various writing-style features including linguistic features and structural features were extracted. To classify the author of one web document, the SVM(support vector machine) algorithm was adopted to learn the author's features. Experiments on Chinese blog, BBS and e-mail dataset gained satisfactory results. The accuracy of blog dataset for seven authors was 89.49%. The satisfactory results showed that it was feasible to put the method to cybercrime forensic application.

Keywords: cybercrime forensic, Chinese web information, authorship analysis, Support Vector Machine, feature extraction.

1 Introduction

With the rapid growth in computer and information technology, especially the increasing popularization of Internet, Internet services such as blog, BBS, e-mail has become convenient means for information communication. As the Internet provides convenience to people, the Internet is also used as illegal purpose which has become a serious problem. Pornographic information, retroactive information, racketeering information, terroristic information are common on the Internet. Misuse of Internet does harm to people's daily life, even affecting social stabilization and national security. How to prevent these phenomena from happening has become a major concern for society.

Passive filtering methods are not effective method to prevent these phenomena from happening. If the illegal cybercrime action can be cracked down on by means of law and the criminal can be justly punished, the illegality of internet can be prevented. Now a lot of states have formulated laws and regulations concerning network criminal acts. But many cases cannot be put on trial because the court lacks necessary and effective evidence. The anonymous nature of the internet causes difficulty to identify the authorship. Therefore analyzing the web content by a technical method is important in order to collect effective evidence. Authorship analysis techniques have been

H. Chen et al. (Eds.): PAISI 2009, LNCS 5477, pp. 14–24, 2009.
© Springer-Verlag Berlin Heidelberg 2009

used to analyze the authorship of literary works by capturing the unconscious features of an author's style and making use of statistical methods to measure these features. However, the web information differs from literary works, which determines the methods for authorship analysis of web information are different from that of literary works. Section 2 show many researchers have begun to investigate Arabic, English, Japanese documents' authorship analysis. However there are not related authorship analysis researches on Chinese language. The characteristic of Chinese language determines that the feature extracted methods are different from that of other language. In this paper, the method for Chinese web information authorship analysis is described. Far-ranging features representing author's writing style are extracted. Support vector machine algorithms are used to learn the features.

The remainder of the paper is organized as follows. Section 2 presents a general review of stylometry and previous contributions. Section 3 describes the framework of web information authorship analysis. Section 4 is our feature extraction methods. Section 5 provides our experimental methodology and analysis the experimental results. Section 6 is the conclusion of the paper.

2 Related Work

2.1 Stylometry and Authorship Analysis

Stylometry is a linguistic discipline that applies statistical analysis to literary style [3], which is often used to attribute authorship to anonymous or disputed documents. It is based on the assumption that every author's writing style has certain features inaccessible to conscious manipulation. Stylometry is the basis of authorship analysis. There are three major approaches to authorship analysis: authorship attribution, authorship characterization and plagiarism detection [4].

Authorship attribution determines the author of a piece of unidentified writing on the basis of writing style between the author's known works and the disputed one. The Federalist Papers are a good example of authorship attribution problem. The authorship attribution of the twelve federalist papers had been a disputed problem all along since United States was founded. The papers might be written by Madison or Hamilton. Mosteller and Wallace came to the conclusion that the twelve disputed papers were written by Madison by comparing against the word usage rate between Madison and Hamilton [15]. Another well-known study is the attribution of disputed Shakespeare works. Elliot and Valenza compared the writing style to the Earl of Oxford. The writing style include unusual diction, frequency of certain words, choice of rhymes, and habits of hyphenation have been used as testing for authorship attribution [11].

Authorship characterization attempts to formulate author profile by making inferences about gender, education, and cultural backgrounds on the basis of writing style. De Vel investigated the language and gender cohort attribution from e-mail text documents [9].

Plagiarism detection is used to calculate the degree of similarity between two or more pieces of text, without necessarily determining the authors, for the purposes of determining if a piece of text has been plagiarized. Authorship attribution and authorship characterization are quite distinct problems from plagiarism detection. The authorship analysis of web information belongs to the problem of authorship attribution. In fact, the authorship attribution is the classification problems.

The most extensive and comprehensive application of authorship analysis is in literature. Program code analysis is another authorship analysis application. Program code authorship has been researched for software plagiarism detection, cyber attacking and virus intrusion forensics. Some features such as typographical characteristics, stylistic metric, and programming structure metrics have been researched [12][14][16]. In the context of cybercrime, authorship analysis for the purpose of computer forensic appeared. The forensic analysis attempts to match text to authors for criminal investigation[5]. Currently forensic analysis has become increasingly popular in identification of online messages due to augmented misused of the Internet [1][18].

2.2 Authorship Features

The main focus of authorship analysis is what the features can represent the writing style of authors. Various features have been selected in previous authorship analysis research. Although the primary problems in the field are that there is no consensus of fixed features set. Abbasi and Chen presented a comprehensive analysis on the stylistics features, namely, lexical, syntactical, structural, content-specific, and idiosyncratic features [1].

Lexical features are the words-based features. Several typical lexical features are vocabulary richness, word usage, word length distribution, etc. Syntactic features include the distribution of function words [15] (such as “while”, “enough”, “upon”), punctuation, etc. Structural features are used to measure the overall layout and organization of text. For instance, average sentence length, average paragraph length, presence of greetings within the e-mail documents are common structural features [8]. Content-specific are keywords or phrases on certain topics. Idiosyncratic features include misspellings, grammatical mistakes, etc.

2.3 Web Information Authorship Analysis

Authorship analysis has been widely used in resolving authorship attribution of literary and conventional writing. With the increasing cybercrime arising on the Internet, web information authorship analysis began to draw researchers’ attention.

E-mail is one special type of web information. With rapid growth of e-mail misuse phenomena, e-mail authorship analysis has been researched in forensic investigations. De Vel(2000) applied support vector machine classification model over a set of linguistic and structural features for e-mail authorship attribution for the forensic purpose[6][7][8]. Corney (2003) investigated extensive stylometric features from e-mail texts and support vector machine was used as learning algorithm in his master’s thesis[4]. Tsuboi (2002) studied authorship attribution of e-mail messages and World Wide Web documents written in Japanese [18]. He used the sequential word patterns or word n-grams with n=2 and 3 from each sentence in the documents as features set. Farkhund Iqbal(2008) mined the frequent pattern for authorship attribution in e-mail forensic[13].

Zheng (2006,2003) analyzed the authorship of web-forum, using a comprehensive set of lexical, syntactical , structural features, and content-specific features[18][19]. Ahmed Abbasi(2005,2006,2008) analyzed the authorship identification and similarity detection of web information[1][2][3].

The above researches are English, Japanese, and Arabic documents authorship analysis. But, techniques of authorship analysis used for feature extraction are almost language dependent, and in fact differ dramatically from language to language. For example, Chinese do not have word boundaries explicitly in texts. In fact, word segmentation itself is a difficult problem in Chinese languages. So feature extraction methods for Chinese documents are different to other language such as English and other Indo-European languages. In this paper, the feature extraction methods of web information authorship analysis for Chinese language were analyzed in detail.

Machine learning techniques are the most common analytical approach in recent years. These techniques include decision trees, neural network, and support vector machine. The distinctive advantage of the SVM is its ability to process many high-dimensional applications such as text classification and authorship attribution. Therefore reduction of dimensionality is not necessary and importance weights can be used to quantity how important a specific given feature is in the documents of a text collection[10]. The former studies[3][18] have drawn the conclusion that SVM significantly outperforms neural network and decision trees in authorship analysis. So in this paper, SVM was chosen as the learning algorithm.

3 The Framework of Web Information Authorship Analysis

Figure 1 presents the framework of web information authorship analysis. There are four major components, namely, web information collection, feature extraction, authorship learning, and authorship classification.

The precondition of authorship analysis of web information was to collect as much of the suspects' web information as possible. The computer can not solve the web information directly due to disorderly information such as photos, sound, and advertising information. So it is necessary to pre-process the web information and retain the texts of web information to be analyzed.

In the feature extraction component, the extensive stylometric features that could represent the author's writing style were extracted. The vector space model (VSM) was adopted to represent the writing features.

In the authorship learning component, machine learning techniques was utilized. The support vector machine algorithm was used to learn the writing features and transform into classifier.

The unknown authorship e-mail documents could be categorized into some authors list automatically by the classifier that was trained in learning component.

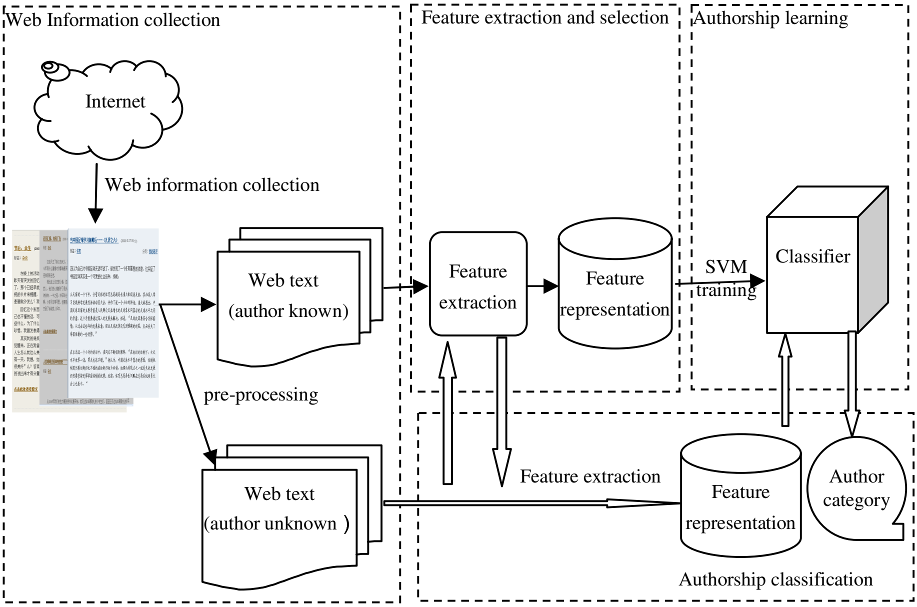


Fig. 1. The framework of web information authorship analysis

4 Feature Extraction Methods

In comparison with literary texts, web texts have unique characteristics. Firstly, the content of web texts should be inputted by keyboard. To save time authors always ignore some information, such as essential punctuation. Wrongly written characters can be seen frequently. Secondly, the authors' writing style is free. The authors don't obey the rules of writing. They type space key or enter key at will. Each author has personal intrinsic habits. So the space and blank lines can offer significant clues to authorship analysis. Thirdly, the length of web text is short. Only several words within web texts are common. The above characteristics of web texts determine that the feature extraction method of web information is different from literary texts. So in this paper linguistic features and structural features were extracted.

Vector Space Model(VSM) was adopted to represent the authors' writing features. In VSM, text-based documents are represented as vectors in a high-dimensional vector space where the value of dimensions is based on the feature in that document. Each document is represented as a vector of term and weight pairs. Namely document d will be represented by a vector $V_d=((t_1,w_1),(t_2,w_2),\cdots,(t_n,w_n))$. The following is the weight calculating methods of linguistic and structural features.

4.1 Linguistic Features

The linguistic features are lexical based. The purpose of linguistic features is to obtain the preference for usage of some specific words unconsciously. The function words have been used as features in some studies[15]. What kind of function words extracted are key problems to be considered. Fixed function words don't always represent the writing style of authors. So in this study, lexical features that could be extracted automatically by *tf-idf* techniques from web texts were proposed.

Chinese texts don't have natural delimiter between words. To extract words from Chinese texts Chinese word segmentation software is needed. In our study, the word segmentation software named ICTCLAS developed by Chinese academy of sciences was used for word segmentation and part of speech tagging.

Formula 1 is the weight calculating methods of linguistic features.

$$W(t,\vec{d}) = f(t,\vec{d}) \times \log(N/n_t + 0.01) \tag{1}$$

Where $W(t,\vec{d})$ is the weight of term t in document *d*, $tf(t,\vec{d})$is the frequency of term *t* in document *d*, *N* is the total number of documents, n_t is the number of documents that contain term *t*.

4.2 Structural Features

In web texts authors always ignore some punctuation or use incorrect punctuation. The authors can write freely on the premise of expressing the author's meaning. So the structure of web texts is loose. Furthermore, the authors have a preference for part of speech usage which can reflect the authors' degree of education. So we extracted three aspects of structural features, namely, punctuations features, structural characteristics, and part of speech features.

The word of web text should be inputted by keyboard. To save labor, the authors always ignore switching the Input Method Editor. Chinese punctuations and corresponding English punctuations were extracted. Table 1 was the punctuations features.

Table 1. Web documents' punctuations features

Chinese Punctuations				English Punctuations	
——	…	。	，	.	,
、	；	：	？	:	?
！	“	”	〔	(	)
〕	《	》	·	“	!
．	‘	’	－	;	‘

The weight of the punctuations features was calculated by the formula 2.

$$W_p(t,\vec{d}) = \frac{\text{Number of punctuation t in document d}}{\text{Total number of punctuations in document d}} \tag{2}$$

Structural characteristics deal with the text's organization and layout. We extracted 10 structural characteristics which were listed in table 2.

Table 2. Web documents' structural characteristics

Feature type
Mean sentence length
Mean paragraph length
Number of digital characters/total number of words
Number of lowercase letters/total number of words
Number of uppercase letters/total number of words
Number of space/total number of words
Number of blank lines/total number of lines
Number of indents/total number of words
Number of distinct punctuations/total number of punctuations
Number of distinct words/total number of words

The rate of part of speech can reflect the preference for word class usage. For example some authors always use exclamation, whereas some authors never do. The usage of part of speech can reflect the authors' degree of education. Chinese has 12 categories part of speech in common use which were listed in table 3.

Table 3. Web documents' part of speech features

Number	Part of speech features	Number	Part of speech features
1	noun	7	adverb
2	verb	8	preposition
3	adjective	9	conjunction
4	numeral	10	auxiliary
5	quantity	11	exclamation
6	pronoun	12	onomatopoeia

The weight of the part of speech features was calculated by the formula 3.

$$W_s(t,\vec{d}) = \frac{\text{Number of part of speech } t \text{ in document } d}{\text{Total number of part of speech in document } d} \tag{3}$$

5 Experiments

5.1 Experimental Methods

To test whether the method was effective, experiments were made on blog, BBS and e-mail dataset. Blog dataset was collected from http://blog.sina.com.cn/. BBS dataset was gained from one university forum. Due to public e-mail dataset was difficult to gain, E-mail dataset came from personal e-mails in our laboratory. Table 4 was the summary statistics of dataset in the experiment.

Table 4. Summary statistics of blog, BBS and e-mail dataset in the experiment

Dataset	Number of authors	Average number of documents	Document size(number of words)		
			Minimum	Maximum	Average
blog	7	197	9	1309	485
BBS	5	82	2	191	17
E-mail	5	19	5	573	76

Since there was only a small amount of data to produce a model of authorship, the performance of each feature was measured by 5-fold cross-validation to provide a more meaningful result. A SVM implementation, Libsvm-2.88 which was developed by Chih-Jen Lin was used for this experiment with linear kernel function.

5.2 Experimental Results and Discussions

Table 5 summarizes authorship identification accuracy results for the comparison of the different feature types.

Table 5. The experimental results of different features types combination (accuracy %)

Dataset	Linguistic	Structural	Linguistic + Structural
Blog	82.39	75.07	89.49
BBS	65.69	68.37	73.97
E-mail	76.53	75.51	80.61

From Table 5, we can get the following Figure 2. In the chart, AL denotes the Linguistic features, AS denotes the Structural features, and ALS denotes the Structural features and Linguistic features combination.

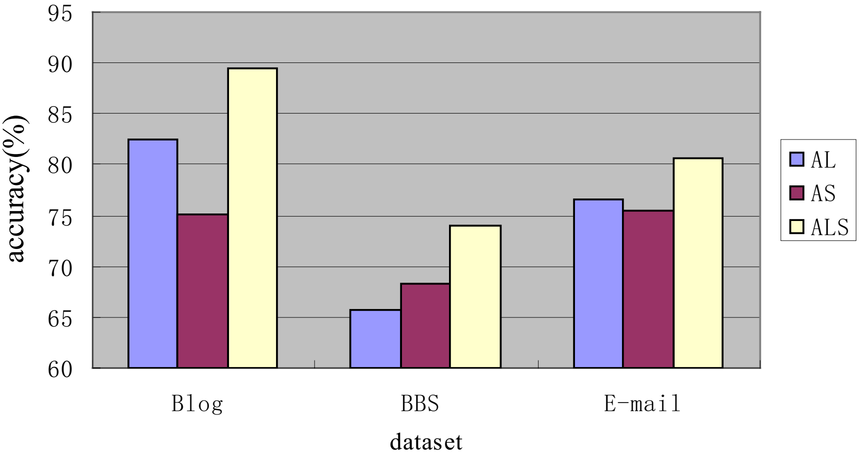


Fig. 2. The experimental results of different features types combination

From the experimental results of Table 5 and Figure 2 we draw three conclusions.

Firstly, the more the number of words in documents the better the experimental results. The number of words in blog dataset was more than that of bbs or e-mail dataset. We could see that experimental results on blog dataset were best. The results on bbs dataset were worst. More words in documents could embody the author's writing style distinctly.

Secondly, the experimental results of linguistic and structural features combination were better than the linguistic or structural features separately. The linguistic and structural features combination were more comprehensive and could express the author's writing style better.

Thirdly, linguistic features and structural features were all effective. The results of linguistic features and structural features were 82.39% and 75.07%, which showed that the linguistic and structural had discrimination function in authorship analysis.

To test the effect of the number of authors on experimental results, we did experiments on blog dataset for different number of authors. The experimental results were showed in Table 6 and Figure 3.

Table 6. The experimental results of different number of authors (accuracy %)

Feature type	7	6	5	4	3	2
Linguistic	82.39	82.45	83.01	84.75	88.17	90.75
Structural	75.07	84.75	85.76	88.88	88.17	90
Linguistic + Structural	89.49	88.98	90.74	92	92.5	94.25

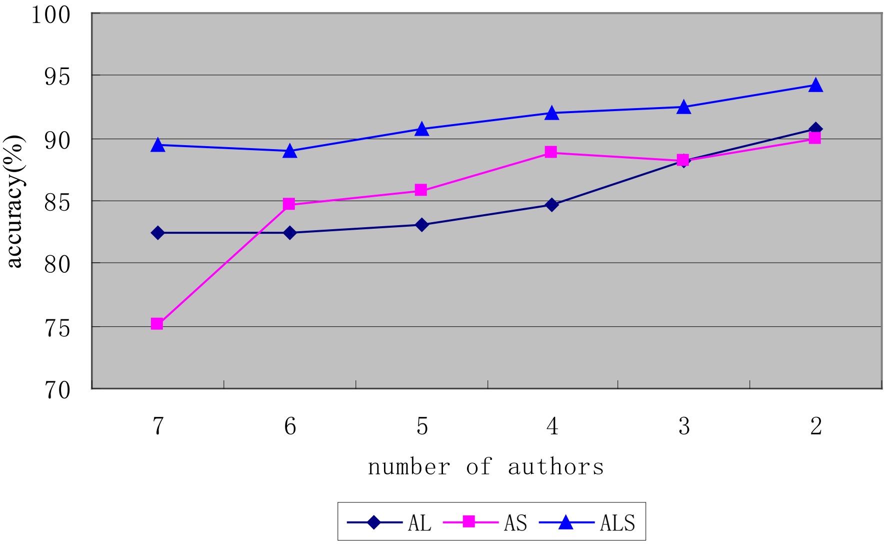


Fig. 3. The experimental results of different number of authors

From Table 6 and Figure 3, we could see that the accuracy of authorship analysis decreased with the authors increasing in number. The experimental results were the same as that of Abbasi(2008) and Zheng(2006)[1][18]. Possibly that was because the larger number of authors could disruptors effectively differentiating authorship.

6 Conclusion

A Cybercrime forensic method for Chinese web information authorship analysis was proposed in this paper. Based on authors' writing style, linguistic features, and structural features were extracted. The support vector machine algorithm was adopted as learning algorithm. To test the effect of the method, blog, BBS, and e-mail dataset was tested. The experimental results showed that the accuracy of linguistic and structural features combination was better than that of features separately. The accuracy of linguistic and structural features combination on blog dataset for seven authors was 89.49%. Furthermore, the conclusion that the accuracy of authorship analysis decreased with authors increasing in number was made. The experimental results were satisfying, which showed that the method was feasible to apply to cybercrime forensic.

References

1. Abbasi, A., Chen, H.: Writeprints: A Stylemetric Approach to Identity-Level Identification and Similarity Detection in Cyberspace. ACM Transactions on Information Systems 26(2) (2008)
2. Abbasi, A., Chen, H.: Visualizing authorship for identification. In: Proceeding of IEEE International Conference on Intelligence and Security Informatics, San Diego, pp. 60–71 (2006)
3. Abbasi, A., Chen, H.: Applying Authorship Analysis to Extremist- Group Web Forum Messages. IEEE Intelligence System 20(5), 67–75 (2005)
4. Corney, M.: Analysing E-mail Text Authorship for Forensic Purpose. Australia, University of Software Engineering and Data Communications (2003)
5. Crain, C.: The Bard's fingerprints, Lingua Franca, pp. 29–39 (1998)
6. De, Vel, C.: Mining E-mail Authorship. In: KDD 2000 Workshop on Text Mining, ACM International conference on knowledge Discovery and Data Mining, Boston, MA, USA (2000)
7. De, Vel, C., Anderson, A., Corney, M., Mohay, G.: Multi-Topic E-mail Authorship Attribution Forensics. In: ACM Conference on Computer Security - Workshop on Data Mining for Security Applications, Philadelphia, PA (2001)
8. De Vel, C., Anderson, A., Corney, M., Mohay, G.: Mining E-mail Content for Author Identification Forensic. SIGMOD Record 30(4), 55–64 (2001)
9. De, Vel, C., Corney, M., Anderson, A., Mohay, G.: Language and gender author cohort analysis of e-mail for computer forensics. In: Proceeding of digital forensic research workshop, New York, USA (2002)
10. Diederich, J., Kindermann, J., Leopold, E., Paass, G.: Authorship Attribution with Support Vector Machines. Applied Intelligence (19), 109–123 (2003)
11. Elliot, W., Valenza, R.: Was the Earl of Oxford the true Shakespeare? Notes and Queries (38), 501–506 (1991)

12. Frantzeskou, G., Gritzalis, S., MacDonell, S.: Source Code Authorship Analysis for supporting the cybercrime investigation process. In: Proc. 1st International Conference on e-business and Telecommunications Networks (ICETE 2004), vol. 2, pp. 85–92 (2004)
13. Iqbal, F., Hadjidj, R., Fung, B.C.M., Debbabi, M.: A novel approach of mining write-prints for authorship attribution in e-mail forensics. Digital Investigation 5(1), 42–51 (2008)
14. Krsul, I., Spafford, E.: Authorship analysis: Identifying the author of a program. Computers and Security (16), 248–259 (1997)
15. Mosteller, F., Wallace, D.L.: Inference and Disputed Authorship. In: The Federalist, Addison-Wesley Publishing Company, Inc., Reading (1964)
16. Sallis, P., MacDonell, S., MacLennan, G., Gray, A., Kilgour, R.: Identified: Software Authorship Analysis with Case-Based Reasoning. In: Proc. Addendum Session Int. Conf. Neural Info. Processing and Intelligent Info. Systems, pp. 53–56 (1997)
17. Tsuboi, Y.: Authorship Identification for Heterogeneous Documents. Nara Institute of Science and Technology, University of Information Science (2002) (Japanese)
18. Zheng, R., Li, J., Huang, Z., Chen, H.: A framework for authorship analysis of online messages: Writing-style features and techniques. Journal of the American Society for Information Science and Technology 57(3), 378–393 (2006)
19. Zheng, R., Qin, Y., Huang, Z., Chen, H.: Authorship analysis in cybercrime investigation. In: Proceedings of the first international symposium on intelligence and security informatics, Tucson AZ USA, pp. 59–73 (2003)

Prediction of Unsolved Terrorist Attacks Using Group Detection Algorithms

Fatih Ozgul[1], Zeki Erdem[2], and Chris Bowerman[1]

[1] Department of Computing &Technology, University of Sunderland, SR6 0DD
Sunderland, United Kingdom
{fatih.ozgul,chris.bowerman}@sunderland.ac.uk
[2] TUBITAK- Marmara Research Centre, Information Technologies Institute,
41470 Gebze, Kocaeli, Turkey
zeki.erdem@bte.mam.gov.tr

Abstract. Detection of terrorist groups using crime data has few examples currently; this is because of lack of detailed crime data which contain terrorist groups' attacks and activities. In this study, a novel prediction model; CPM is applied to a crime dataset which includes solved and unsolved terrorist events in Istanbul, Turkey between 2003 and 2005, aiming to predict perpetuators of terrorist events which are still remained unsolved. CPM initially learns similarities of crime incident attributes from all terrorist attacks and then puts them in appropriate clusters. After solved and unsolved attacks are gathered in the same "umbrella" clusters, CPM classifies unsolved cases to previously known terrorist groups. Recall and precision results and findings of CPM regarded as successful then a baseline system; TMODS.

Keywords: Terrorist groups, crime data mining, matching and predicting crimes, clustering, classification, offender networks, group detection.

1 Introduction

Terrorist attacks regularly occur in some cities; some of them are solved and identified for which terrorist organizations planned and realized them, whereas some of the terrorist attacks remain unsolved. Current terrorism informatics, which aims to help security officials using data mining techniques, is mainly focused on using social network analysis (SNA) for structural and positional analysis of terrorist networks [2] where required information is provided from non-crime data. Using crime data has few examples currently; this is because of lack of requiring terrorist organization's consistent crime data with their useful attributes about the attacks. In this study, we apply Crime Prediction Model (CPM) which relies on our previous models: OGDM and GDM [10]. Our datasets include terrorist events happened in a province of Turkey where unsolved and solved crime incidents are together. CPM aims to predict perpetuators of some terrorist events which are remain unsolved. Prediction results are classified according to known terrorist groups in this data set and findings of CPM algorithm are accepted as successful.

H. Chen et al. (Eds.): PAISI 2009, LNCS 5477, pp. 25–30, 2009.
© Springer-Verlag Berlin Heidelberg 2009

2 Predicting Past Terrorist Attacks

Open source intelligence, terrorist web sites, intelligence records, email and telephone signal information are accepted as basic sources to detect terrorist networks. Detecting terrorism from raw data can be posed as an unsupervised clustering problem. If raw data contains unsolved crimes and activities and known terrorist groups then this task can be seen as supervised prediction of unsolved events and a classification problem. Using crime incident data for predicting past unsolved terrorist attacks is not common. Crime data may contain descriptions of the activities and features of single events, but relationships among each record feature act in ways that they are correlated. We can consider them as logical links between crimes, or "behavioral links". For example, datasets may contain information about locations regularly used for terrorist attacks over time, using particular method of conducting events (like bombing, throwing Molotov cocktails), particular memorial days or anniversaries of terrorist groups and their choice of location. Our aim here firstly must be obtaining cohesive clusters of generalized events, and then trying to match known terrorist groups to obtained clusters, and finally classifying and predicting perpetuator of an unsolved crime to a known terrorist group. It is difficult to predict perpetuator of series of attacks which belong to unheard single terrorist group, if attacks are not identical.

Although there has been some remarkable research on predicting unsolved crimes using crime classification, clustering, and artificial intelligence, only prediction work on terrorist networks is done with Terrorist Modus Operandi Detection System (TMODS), which is developed by 21st Century Technologies [1,2,3,5,6,7] where it uses various graph-matching, ontological matching and graph isomorphism methods for terrorist threat matching scenario against the graph database of recorded previous crimes, intelligence records, email and telephone signal information. In TMODS, user defined threatening activity or a graph pattern can be produced with a possible terrorist network ontology and this can be matched against graph database. At the end, human analyst views matches that are highlighted against this input graph pattern.

3 Crime Prediction Model (CPM)

CPM is developed to from the same understanding over Group Detection Model (GDM) [8] and Offender Group Detection Model (OGDM) [8, 9].

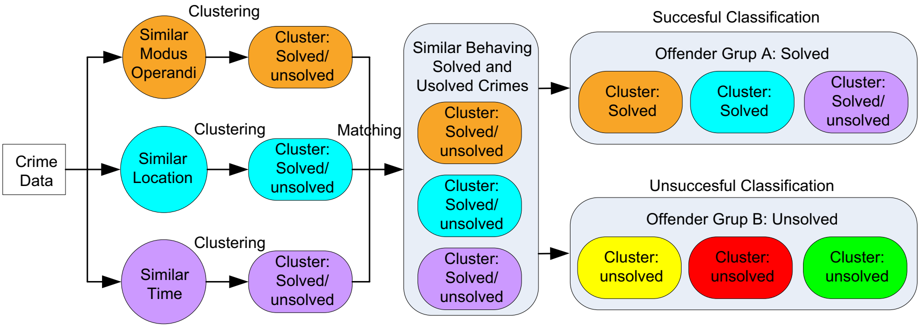

Crime Data
Clustering
Similar Modus Operandi
Cluster: Solved/ unsolved
Clustering
Similar Location
Cluster: Solved/ unsolved
Clustering
Similar Time
Cluster: Solved/ unsolved
Matching
Similar Behaving Solved and Usolved Crimes
Cluster: Solved/ unsolved
Cluster: Solved/ unsolved
Cluster: Solved/ unsolved
Succesful Classification
Offender Grup A: Solved
Cluster: Solved
Cluster: Solved
Cluster: Solved/ unsolved
Unsuccesful Classification
Offender Grup B: Unsolved
Cluster: unsolved
Cluster: unsolved
Cluster: unsolved

Fig. 1. Crime Prediction Model

Contrary to TMODS, To predict unsolved crimes, CPM uses only crime data by looking towards similarity in time, location, modus operandi and for crimes. CPM takes following steps as exhibited in figure 1. CPM uses solved and unsolved crime information, basically learning from attributes of each crime. Similar modus operandi cluster for crimes, similar located crimes cluster and (the – same day – same month – not the same year –) similar dated crime clusters using inner join query on crime data table. These clusters contain solved and unsolved crimes in specific modus operandi, date and location type clusters like “umbrellas”. For example one cluster contain Crime A and Crime B under the same umbrella have one or more than one similar features (i.e. location, date, modus operandi) which can be measured with J(A, B), a Jaccard coefficient [4] similarity score as;

$$\mathcal{J}(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (1)$$

For each crime A in a cluster, it has at least one of three similarity score with crime B; namely similarity score for modus operandi, similarity score for location, and similarity score for date. When we represent those similarity scores as $\mathcal{J}\ modus\ operandi(A,B)$,$\mathcal{J}\ location(A,B)$, $\mathcal{J}\ date\ (A,B)$ then total similarity $\triangle\ \mathcal{J}\ (A,B)$ could be measured by Euclidian distance of these scores;

$$\triangle\ \mathcal{J}\ (A,B) = \sqrt{\mathcal{J}\ modus\ operandi(A,B)^2 + \mathcal{J}\ location(A,B)^2 + \mathcal{J}\ date(A,B)^2} \quad (2)$$

Getting Euclidian distance for $\triangle\ \mathcal{J}\ (A,B)$ provides full matching of clusters, and within these clusters putting similarly behaving solved and unsolved crimes in more accurate clusters.

Threshold value for $\triangle\ \mathcal{J}$ such as 0.25 or 0.5 should be decided upon the size of clusters containing unsolved crimes to be predicted. If threshold value is decided lower, all predicted crimes will have similar counterparts with scores but it will also have false positives. If threshold value gets higher there will be more accurate results but some of crimes will get similarity scores under the threshold, so will remain unpredicted. Finally, CPM looks for possible perpetuators of crimes, assuming that majority of crimes in a single cluster were committed by the same single offender group. If majority of crimes are not addressed to more than single offender group, then similarity scores with high $\triangle\ \mathcal{J}$ values are addressed to prevailing offender group. This can also be done with following equations (equation 3, 4) [4]. One possible technique for this is comparison of means and variances of $\triangle\ \mathcal{J}$. For example CPM offers two candidate groups; Group A and B. To decide which is between these two terrorist groups we look for firstly comparison of variances as in equation 3.

$$Selection(A,\ B) = \sqrt{\frac{var(\triangle\mathcal{J}(Group\ A))}{number\ of\ Group\ A\ cases} + \frac{var(\triangle\mathcal{J}Group\ B)}{number\ of\ Group\ B\ cases}} \quad (3)$$

Secondly, means of Jaccard similarity value is divided by this selection score as in equation 4, then we get group selection score for two candidate groups. These steps can be taken for three, four or more candidates to predict one of them is the perpetuator.

$$Group\ Selection\ (A) = \frac{mean(\Delta\mathcal{J}(Group\ A))}{Selection\ (A,B)} \tag{4}$$

4 Evaluation and Results

We evaluated the performance of CPM model using precision and recall that are widely used in information retrieval. Precision, in this scenario, is defined as the percentage of correctly detected group in cluster that committed unsolved crimes divided by the total number of within-cluster groups. Recall is the percentage of matching group within-cluster divided by correctly detected overall crimes of matching group.

$$precision = \frac{TP}{TP+FP} \qquad recall = \frac{TP}{TP+FN} \tag{5}$$

Dataset contains 1244 terror crimes which took place between 2003 and 2005. Out of 1244 crimes, 1037 of them are solved, 239 of them unidentified 207 of them are unsolved. Main terrorist groups responsible for the crimes are PKK (400 crimes), DHKP/C (198 crimes), MLKP (154 crimes), TKP/ML-TIKKO (44 crimes) and Hizbut-tahrir (32 crimes).

Types of unsolved crimes to be predicted are exhibited according to their types. Most of the unsolved crime types are attacking with Molotov cocktail (firebomb), leaving bomb somewhere with intention to explode, illegal riots and demonstrations,

Table 1. Performance matrix for CPM.

	Perpetuator group of unsolved crime	Perpetuator group of other crimes
Terrorist groups considered refer to the unsolved crime	True Positive (TP)	False Positive (FP)
Terrorist groups considered refer to other crimes (not to the unsolved crime)	False Negative (FN)	True Negative (TN)

Table 2. Results for predicted crimes. In total 237 previously unsolved crimes predicted, 39 previously unsolved crimes are non-matched and not predicted.

Terrorist Group	Before Prediction (Solved Crimes)	After Prediction	Predicted
PKK	400	511	111
DHKP/C	198	228	29
MLKP	154	184	30
TKP/ML-TIKKO	44	53	9
Hizbut-Tahrir	32	37	5
TOTAL	1037	1274	237

arson and leaving illegal billboards to propagandize terrorist groups. We applied CPM algorithm that we developed to cluster the solved and unsolved crimes in the dataset. As we mentioned in equation 1 and 2, having a high total Jaccard value ($\Delta \mathcal{J}$) against other unidentified groups gives the possible perpetuator of selected unsolved crime. If there are more than one possible terrorist group, we apply feature selection method as mentioned in equation 3 and 4. Results for predictions are exhibited in Table 2.

Precision were high especially for big terrorist groups (many members) but going downwards when the size of terrorist groups get smaller (few members). Average precision was 0,64. Recall values were fluctuating but generally didn't be smaller than precision where average recall was 0,71. Figure 2 represents precision and recall values for terrorist group prediction. It is told that in a benchmark exercise, in a specific problem domain, TMODS can achieve 55% recall and 71.6% precision in search results at the stated processing rate [6]. In general bigger groups get high precision values, whereas small groups get high recall values. This is probably because of matching a few similar crimes pointing at small-sized terrorist groups can cluster many unsolved crimes with few solved crimes, since unsolved crimes are evaluated non-desired false negatives. False negatives decrease recall value as exhibited in table 1.

There are also some very high recall values (1,00) for some terrorist groups, that is because of non existence of false negatives. Very specific attacks, which holds characteristics to very few groups, are matched for these terrorist groups, none of the rest of the crimes are similar to these attacks so they produced high recall values with low precision.

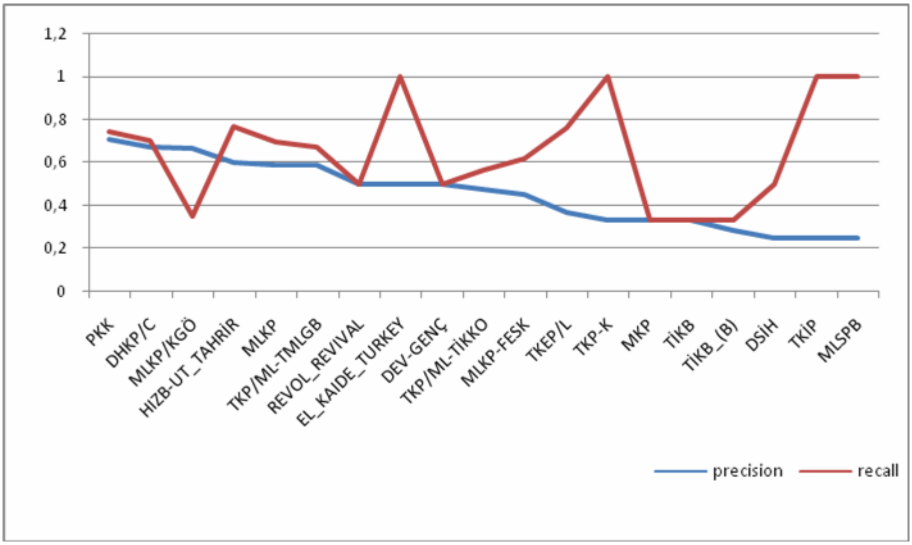


Fig. 2. Precision and recall values according to predicted terrorist groups

5 Conclusion and Further Recommendations

We proved in this study that crime prediction can also be made with the same understanding with group detection algorithms using crime incident data. Rather than analyzing an existing terrorist group which was detected previously from non-crime data, more research should be done for using terrorist attacks crime data. CPM performed well on behavior attributes of crime to predict terrorist activities.

CPM is developed for terrorist crime incident data, but as the further step demographics of terrorist groups should also be used for unsolved crime prediction. Another suggestion can be developing CPM for producing better and consistent recall values. If prediction of unsolved crimes are made with phone calls or e-mail data, attributes of these data sources should be sought for similarities as CPM does for crime data.

References

1. Coffman, T., Greenblatt, S., Marcus, S.: Graph-based technologies for intelligence analysis. Communication of ACM 47(3), 45–47 (2004)
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. 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)
4. Kantardzic, M.: Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons, New York (2003)
5. Marcus, S., Coffman, T.: Terrorist Modus Operandi Discovery System 1.0: Functionality, Examples, and Value. In: 21st Century Technologies, Austin TX (2002)
6. 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)
7. Moy, M.: Using TMODS to run best friends group detection algorithm. In: 21st Century Technologies, Austin, TX (2005)
8. 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)
9. 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)

Exploring Fraudulent Financial Reporting with GHSOM

Rua-Huan Tsaih[1], Wan-Ying Lin[2], and Shin-Ying Huang[1]

[1] Department of Management Information Systems, National Chengchi University ,
Taipei 11605, Taiwan
[2] Department of Accounting, National Chengchi University, Taipei 11605, Taiwan

Abstract. The issue of fraudulent financial reporting has drawn much public as well as academic attention. However, most relevant researches focus on predicting financial distress or bankruptcy. Little emphasis has been placed on exploring the financial reporting fraud itself. This study addresses the challenge of obtaining an enhanced understanding of the financial reporting fraud through the approach with the following four phases: (1) to identify a set of financial and corporate governance indicators that are significantly correlated with fraudulent financial reporting; (2) to use the Growing Hierarchical Self-Organizing Map (GHSOM) to cluster data from listed companies into fraud and non-fraud subsets; (3) to extract knowledge from the fraudulent financial reporting through observing the hierarchical relationship displayed in the trained GHSOM; and (4) to provide justification to the extracted knowledge.

Keywords: Financial Reporting Fraud, Growing Hierarchical Self-Organizing Map, Knowledge Extraction.

1 Fraudulent Financial Reporting and GHSOM

Fraudulent financial reporting can lead to not only significant investment risks for stockholders, but also financial crises for the capital market. Fraudulent financial reporting were often detected with a very low frequency but with severe impacts [1]. Given the infrequency of synthetic reporting, most auditors cannot develop sufficient experiences and knowledge on fraudulent detection [7]. Furthermore, top management may be involved in providing less fairly represented financial statements. Beasley found that 83% of top managements of the U.S. listed firms, chief executive officer, chief financial officer, sometimes even both, are related to financial statement fraud [3]. Tipgos notes that internal control is designed in a “giver-receiver” model [14]. It means that management implements the internal control and employees are expected to follow it. The internal control mechanism aims to prevent employee frauds, not management frauds. In other words, since managers could bypass the internal control, it created a significant condition of financial statement fraud to lead to bamboozle auditors deliberately [10]. The standard audit procedures are insufficient to detect malfeasance for managers who understand the limit of audit [7].

H. Chen et al. (Eds.): PAISI 2009, LNCS 5477, pp. 31–41, 2009.
© Springer-Verlag Berlin Heidelberg 2009

Table 1. Literature summary of fraud detection techniques

Author	Technique	Variable	Sample	Findings
Persons (1995)	Stepwise logistic model	➢ 9 financial ratios ➢ Z-score	Matched-pairs design	The study found four significant indicators: financial leverage, capital turnover, asset composition and firm size
Fanning and Cogger (1998)	Self-organizing artificial neural network	62 variables ➢ Financial ratios ➢ Other indicators: corporate governance, capital structure etc.	Matched-pairs design: 102 fraud and 102 non-fraud	➢ Neural network is more effective ➢ Financial ratios are over half of 8 significant indicators such as debt to equity, ratios of accounts receivable to sales, trend variables etc.
Bell and Carcello (2000)	Logistic regression	46 fraud risk factors	77 fraud samples and 305 non-fraud samples	Logistic regression model was significantly more effective than auditors for fraud samples, but for non-fraud samples both made no difference.
Kirkos et al. (2007)	● Decision tree ● Back-propagation neural network ● Bayesian belief network	➢ 27 financial ratios ➢ Z-score	Matched-pairs design: 38 fraud and 38 non-fraud	➢ Training dataset: neural network is the most accurate ➢ Validation dataset: Bayesian belief network is the most accurate
Hoogs et al. (2007)	Genetic algorithm	➢ 38 financial ratios ➢ 9 qualitative indicators	1fraud vs. 8 non-fraud design	Integrated pattern had a wider coverage for suspected fraud companies while it remained lower false classification rate for non-fraud ones

There are numerous studies dealing with prediction of financial statement fraud using either logistic regression or neural network [4] [7] [9] [11][15]. Table 1 summarizes the fraud detection techniques.

Although several studies have shown the benefits of fraud prediction using regression model or neural network, they are often criticized that they are difficult to deal with high-dimensional data and limited samples. The specific criticism of neural network was a black box of classification process so that auditors were unable to understand the adopted factors and to verify validity of the model.

This study presents the application of Growing Hierarchical Self-Organizing Map (GHSOM) [12] to extract rules and knowledge about financial reporting fraud. Self-Organizing Map (SOM) is designed with the concept of unsupervised learning network to handle high-dimensional data and visualize results. It can also help to discover hidden hierarchies and to extract rules and knowledge from data. Unfortunately, SOM requires predefined number of neural processing units, static architecture of this model and has limited capabilities for the representation of hierarchical relations of the data. The GHSOM will automatically grow additional layers to display refined maps of individual nodes exhibiting high degrees of heterogeneity, thus providing higher levels of classification as needed. In addition, the size of each map was relatively smaller so that users can easily analyze and quickly obtain an overview. In the practical aspects, several studies [5][6][13] applied GHSOM into information extraction field. Hence, the study applies GHSOM solution to dynamic network structure and hierarchical relationship to help auditors effectively extract rules or features of financial reporting fraud.

2 Experimental Design and Results

The research process can be described as Figure 1. In sampling stage, we first use the following sources to identify the occurrence of financial statement fraud. The first source is Summary of Indictments and Sentences for Major Securities Crimes issued by the Securities and Futures Bureau of the Financial Supervisory Commission in Taiwan. The second source is Summary of Group Litigation Cases issued by Securities and Futures Investors Protection Center in Taiwan. The third source is the law and regulations retrieving system of The Judicial Yuan which provides information to verify whether the accused companies committed the financial statement fraud.

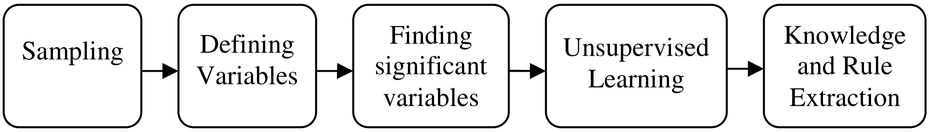


Fig. 1. Research Framework

Table 2. Definition of Fraud code and Color code

Fraud code		Color code		
0	Non-fraud Year	0	□ White	For fraud firms, the years which might be not investigated after detection year were assigned 0. All years of non-fraud firms would be assigned the same color code.
1	Fraud Year	1	■ Light blue	For fraud firms, the years under investigation and shown non-fraud after the last fraud year and before detection year were assigned 1.
		2	■ Deep blue	For fraud firms, the year preceding the first fraud year could be investigated and shown non-fraud.
		3	■ Gray	For fraud firms, other years before the first fraud year with low possibility of investigation were given 3.
		4	■ Red	For fraud firms, the years that appeared fraud in the indictments or judgments were marked 4.

If a company was prosecuted or judged according to the following enactments of Taiwan, it is a fraud firm:

(1) Securities and Exchange Law：Paragraph 2, Article 20;
(2) Securities and Exchange Law：Sub-paragraph 5, Paragraph 1, Article 174;
(3) Business Accounting Law：Article 71;
(4) Criminal Law：Article 342.

For the fraud firm, indictments and judgments do state the detection year that was investigated by prosecutors' offices and the fiscal year of financial statements that are fraudulent. For each fraud company, there is a five-year sampling period of financial statements, which covers two years before and after the first fault year. Based on the fraud years and detection year, we further distinguish the difference among non-fraud year financial statements of fraud firms. The detail of fraud code and color code can be explained as Table 2.

We use the matched-pairs concept to create a sample pool of 58 fraud firms and 58 non-fraud firms, all of them are Taiwan publicly traded companies. For each fraud firm, we pick up a non-fraud counterpart that is in the same industry and the total assets are near to those of the fraud firm in the year before the first fraud year. There are totally 117 fraud financial-statement samples and 463 non-fraud financial-statement samples. The imbalanced samplesconsistent with facts which fraud cases were infrequent relatively.

Every sample data has a code composed of industry abbreviation, four-digit stock code, two-digit year abbreviation, one-digit fraud code, and one-digit color code. Take a fraud firm in electron industry for example as Table 3:

Table 3. An example of sample encoding.

Industry abbrev	Stock code	Year abbrev.	Fraud code	Color code	**Sample code**
E	8295	00	0	0	E82950000
		99	0	1	E82959901
		98	1	4	E82959814
		97	0	2	E82959702
		96	0	3	E82959603

We employed one nominal dependent variable-FRAUD, which is dichotomous and expressed as 1 and 0 according to whether the year financial statement is fraud or non-fraud. Initially, we use 25 independent variables from financial and corporate governance dimensions to be indicators of financial reporting fraud. The financial ratios were employed to measure profitability, liquidity, operating ability, financial structure and cash flow ability of a firm. Moreover, corporate governance variables and Z-score were utilized to examine probability of financial distress.

The measurement for profitability indicators we adopt are: Gross profit margin (GPM), Operating profit ratio (OPR), Return on assets (ROA), [8]), Growth rate of net sales (GRONS), Growth rate of net income (GRONI).

The liquidity ability indicators we adopt are: Current ratio (CR), Quick ratio (QR).

The operating efficiency indicators we adopt are: Accounts receivable turnover (ART), Total asset turnover (TAT), Growth rate of accounts receivable (GROAR), Growth rate of inventory (GROI), Growth rate of Accounts receivable to gross sales (GRARTGS), Growth rate of Inventory to gross sales (GRITGS), Accounts receivable to total assets (ARTTA), Inventory to total assets (ITTA).

The financial structure indicators we adopt are: Debt ratio (DR), Long-term funds to fixed assets (LFTFA).

The cash flow ability indicators we adopt are: Cash flow ratio (CFR), Cash flow adequacy ratio (CFAR), Cash flow reinvestment ratio (CFRR).

The corporate governance indicators we adopt are: Stock Pledge ratio (SPR), Deviation between control rights and cash flow rights (DBCRCFR), Deviation between ratio of controlled board seats and cash flow rights (DBCBSCFR), Z-score.

Then we adopt the CANDISC to do tolerance test for multi-collinearity and significance test to derive significant variables. The result of multi-collinearity test suggested that the GRITGS variable should be excluded since its tolerance was extremely lower than other independent variables. Then, we use the F-value to determine the significance of each of remaining 24 independent variables. The result indicated that the following eight variables had statistically significant effects: ROA, CR, QR, DR, CFR, CFAR, SPR and Z-Score. The result of structure coefficient (i.e. discriminant loadings) shown that ROA had the greatest effect on the function, followed by Z-Score and CFAR has the smallest effect.

These eight significant variables examine a company from different dimensions:

(1) Profitability: ROA can be used to assess a firm's ability to generate profits by the use of its own assets. [11] indicated that lower profit may give management an incentive to overstate revenues or understate expenses.

(2) Liquidity: CR and QR can be used to measure a firm's liquidity which means its short-term ability to pay a debt. QR excludes inventory and prepaid expenses whose ability to realize is lower than cash or accounts receivable.

(3) Financial structure: DR can be employed to inspect a firm's financial structure. [11] found that fraud firms have higher financial leverage than non-fraud firms.

(4) Cash flow ability: CFR and CFAR can be used to test a company's ability to paying debts and other disbursement such as capital expenditures, inventory additions and cash dividends using cash flows from operating activities.

(5) Stock pledge ratio: SPR can be utilized to measure the financial pressure on leverage degree of directors and supervisors by pledging their stocks to obtain funds.

(6) Financial condition: Z-score can be used to measure a company's financial situation to determine the relationship between financial distress and fraud.

The procedure of GHSOM experiment can be mentioned as Figure 2. We use SOM toolbox and GHSOM toolbox in the platform of Matlab R2007a to conduct the GHSOM experiment. The trial and error would be performed in different breadth, depth and normalization method to get a suitable GHSOM model for analysis.

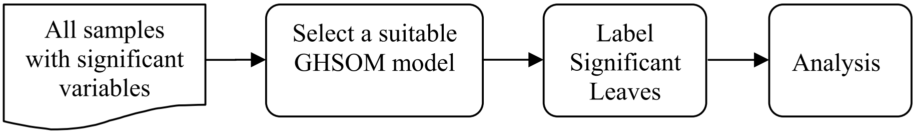

All samples with significant variables
Select a suitable GHSOM model
Label Significant Leaves
Analysis

Fig. 2. Procedure of GHSOM experiment

At the beginning, we used GHSOM and significant variables for total samples (i.e. fraud samples and non-fraud samples) to get several hierarchical structures under different parameters including breadth, depth and normalization method. Next, we compared these GHSOM models and selected a suitable model base on the depth and map size. Finally, we chose some significant leaves with better explanatory power to label and analyze.

The growing process of GHSOM is primarily dominated by two parameters, τ_1 and $\tau_{2.}$ While τ_1 controls the size of individual SOM, τ_2 determines the minimum data granularity and the global termination criterion. Therefore, a resultant hierarchy is not necessarily balanced and individual SOM could have different number of units and configurations.

(Where is on one hand?)On the other hand, each map will be named according to its upper layer, current layer which it was located on and its order in the same upper map. For example, a map named "L1m2-L2m1 " indicates that it was from the second map of layer 1 and its current location is in the first map of layer 2.

The study attempts to present data hierarchy and to extract knowledge from clusters. An applicable model was also capable of comparing among clusters and analyzed within a cluster. Hence, the criteria of a GHSOM model can be defined as:

(1) The depth of a model should be greater than two layers.
(2) The breadth of individual map should consist of two firms at least. It meant that a map would be expected to have over ten samples.
(3) New maps shouldn't extremely cluster in a minority of the parent.

The study performed canonical discriminant analysis to do tolerance test for multi-collinearity and significance test for selecting significant variables. In addition, the analysis also presented prediction rate of discriminant function.

The result of multi-collinearity test suggested that one variable, namely GRITGS, were excluded since their tolerance were extremely lower than other independent variables. The detail was described as Table 4. As a result, we acquired 24 independent variables as input to the Canonical Discriminant Analysis after deleting the variable.

Table 4. Tolerance Test of Variable

Deleted Variables	Variance within groups	Tolerance
GRITGS	89043100.972	.001

The study verified whether or not the discriminant function could show the significant difference by means of Wilks' Λ statistic. The corresponding P-value of Wilks' Λ value which was less than level of significance(α=0.05) proved that a significant effect of the discriminant function. The result can be shown as Table 5.

Table 5. Wilks' Λ Statistic

Function	Wilks' Lambda value	χ^2	D.F.	Significance
1	.766	151.095	24	.000

We made use of F-value to determine the significance of each independent variable. The result of corresponding p-value indicates that eight variables have statistically significant effects, including ROA, CR, QR, DR, CFR, CFAR, SPR and Z-Score.

Next, we utilized structure coefficient (i.e. discriminant loadings) to compare the discriminant power of individual variable. In brief, it is used to estimate the relative importance of each variable to the discriminant function based on the absolute value of structure coefficient. The result shows that ROA had a greatest effect on the function, the secondary is variable Z-Score and variable CFAR is the smallest. The study would analyze the variation of significant variables across a number of years.

The result shown in Table 6 lists the consistency between significance and relative importance from the rank of influence. The significant variables, such as ROA and Z-score, also provided stronger discrimination for the function. With regard to direction of influence, the result pointed out variable DR and SPR among all significant variables appeared negative correlation. It also indicates that a company whose most of significant indicators became bigger may tend to health.

Table 6. Significance and Relative Weights of Independent Variable

Variable	Structure Coefficient	F value	Significance	Rank of Influence	Direction of Influence
GPM	0.14	3.51	0.061		
OPR	-0.03	0.16	0.688		
ROA	0.77	105.82	**0.000*****	1	+
GRONS	0.06	0.63	0.427		
GRONI	-0.02	0.05	0.822		
CR	0.34	20.59	**0.000*****	5	+
QR	0.28	13.42	**0.000*****	7	+
ART	0.09	1.58	0.210		
TAT	0.19	6.38	0.012		
GROAR	0.03	0.12	0.731		
GROI	0.07	0.90	0.344		
GRARTGS	0.00	0.00	0.997		
ARTTA	0.11	2.25	0.134		
ITTA	0.12	2.37	0.125		
DR	-0.42	30.46	**0.000*****	4	-
LFTFA	0.02	0.09	0.764		
CFR	0.33	19.21	**0.000*****	6	+
CFAR	0.24	9.89	0.002***	8	+
CFRR	0.19	6.41	0.012		
SPR	-0.47	38.85	**0.000*****	3	-
SMLSR	-0.19	6.18	0.013		
DBCRCFR	0.02	0.04	0.835		
DBCBSCFR	-0.05	0.41	0.524		
Z-score	0.64	72.74	**0.000*****	2	+

Note. *** p < 0.01.

Table 7. Prediction rate of discriminant function

Class			Predict	
			0	1
Original	No.	0	390	77
		1	44	69
	%	0	83.5	16.5
		1	38.9	61.1

As a whole, the canonical discriminant function incorrectly classify 21.9% of samples, that is to say, prediction power of the function achieved 79.1%. The result was listed in Table 7. In the study dealing with financial reporting fraud prediction, it is less costly to classify a non-fraud firm as a potential candidate for fraud than a potentially fraudulent firm as non-fraud. Based on the principle, type I error should be lower to reduce misclassification costs. Type I error is 16.5% and type II error is 38.9%.

H_0: The firm did not commit financial reporting fraud

H_1: The firm committed financial reporting fraud

The selected GHSOM model develops a tree with three layers and 41 leaves as Figure 3. The tree creates four maps including L1m1, L1m2, L1m3 and L1m4 in the first layer. By means of comparing ratio of fraud to non-fraud sample among them, L1m2 having the highest fraud ratio indicated over half of fraud sample could be classified into its clusters in the following layers. By contrast, L1m3 with the lowest fraud ratio was expected to probably produce some pure non-fraud clusters. The detail can be listed in Table 8.

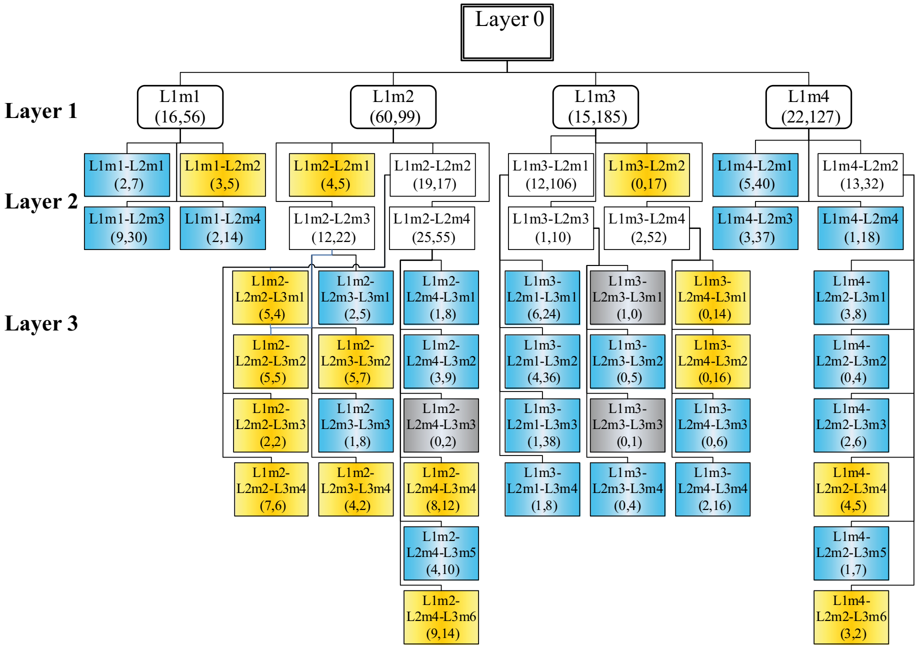


Fig. 3. GHSOM Tree with three layers

Table 8. Ratio of fraud to non-fraud in Layer1

Layer1	Number		Fraud ratio (%)
	Fraud	Non-fraud	
L1m1	16	56	**28.57**
L1m2	60	99	**60.61**
L1m3	15	185	**8.11**
L1m4	22	127	**17.32**

This GHSOM model generates 41 leaves, in which L1m2-L2m4-L3m3, L1m3-L2m3-L3m1 and L1m3-L2m3-L3m3 leaves have less than three sample data. These three leaves are excluded from the further discussion. The overall ratio of number of fraud sample data to number of non-fraud sample data is 113:467; this ratio information is adopted as the norm for picking up the leaves whose ratio is quite deviated from this value. Figure 3 shows 15 pick-up leaves in yellow, deleted leaves in gray, and other leaves in blue. Amongst 15 pick-up leaves, there are three leaves whose number of non-fraud sample data is much larger than of fraud ones.

The study would classify 15 significant leaves into high fraud risk group, mixed group and healthy group according to their characteristics are shown in Figure 4. In GHSOM model, all of high fraud risk groups were generated from L1m2 with the highest fraud ratio while all of healthy groups were produced from L1m3 with the lowest fraud ratio. That is to say, L1m2 and L1m3 had superior discrimination power.

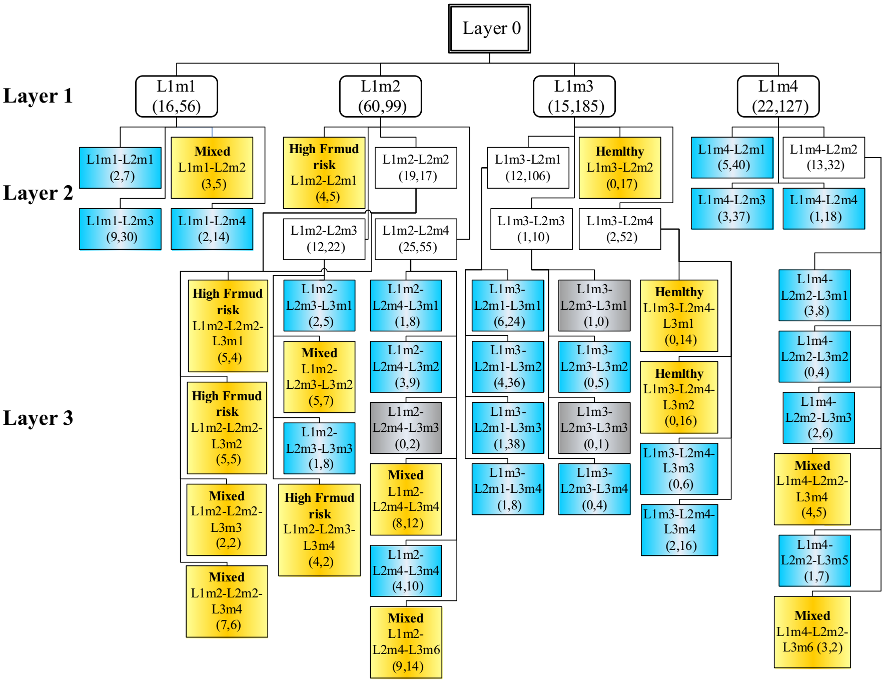


Fig. 4. GHSOM Tree with Label

1. High fraud risk group

The group had totally four leaves whose non-fraud samples all came from fraud firms. The significant features were the worst profitability, liquidity, financial structure and Z-score. The results showed techniques which fraud samples cooked the books probably were not skillful and thorough so financial adversity could be detected from the preceding dimensions. On the other hand, fraud samples have manipulated the related accounts so their financial situation may be more terrible.

2. Mixed group

The group consisted of fraud samples and non-fraud ones which belonged to fraud firms or non-fraud firms. In terms of comparison among leaves, some leaves had few but unique features which were primarily in the worst cash flow ability and higher stock pledge level.

3. Healthy group

The group with three leaves had excellent characteristics in financial structure, stock pledge level and Z-score. Because few fraud firms had good financial state and lower stock pledge level, they probably represented financial statements correctly.

3 Conclusions

The research attempted to present a hierarchical structure from data and to extract knowledge related to financial reporting fraud through Growing Hierarchical Self-Organizing Map (GHSOM) as well as a set of financial and corporate governance indicators. Our financial sample excludes financial industry and includes 113 frauds and 467 non-frauds firm-year observations. First of all, discriminant analysis and GHSOM was applied to obtain eight significant variables and to develop a suitable hierarchical model. Next, 15 leaves with explanation power was selected to perform analysis among and within clusters.

The research result appeared distinctions among high fraud risk group, mixed group and healthy group were described as below: In terms of consistency, all of leaves had smaller coefficient of variation in liquidity, financial structure and stock pledge level while the discrimination among them was in profitability, cash flow ability and Z-score.

To sum up, many fraud samples in 15 leaves belonged to Iron& Steel and Building Material& Construction industry and committed shenanigans in 1998 and 1999 which East Asian Financial Crisis occurred seriously. Because of bear market, the operation of fraud firms deteriorated sharply and could not create cash flow. Under the pressure of capital, they borrowed short-term loans to meet operating demand so financial structure became worse. The vicious circle motivated evil top management to misappropriate corporate money for keeping stock price rising. In the meantime, fraudulent financial reporting could conceal the embezzlement to make investors and banks trust them. Moreover, frequent schemes included overstating revenues through fictitious sales, obtaining money in nominal accounts such as temporary payment or prepayment for purchases, recording loans from related party into accounts receivable, falsified some accounts like accounts receivable etc.

References

1. Association of Certified Fraud Examiners. Report to the nation on occupational fraud & abuse [Electronic Version], `http://www.acfe.com/documents/2006-rttn.pdf`
2. Altman, E.I.: Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The Journal of Finance 23(4), 589–609 (1968)
3. Beasley, M.S., Carcello, J.V., Hermanson, D.R.: Fraudulent financial reporting: 1987-1997 an analysis of U.S. public companies (1999)
4. Bell, T.B., Carcello, J.V.: A Decision Aid for Assessing the Likelihood of Fraudulent Financial Reporting. Auditing 19(1), 169–184 (2000)
5. Dittenbach, M., Merkl, D., Rauber, A.: The Growing Hierarchical Self-Organizing Map. In: The Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks- IJCNN 2000 (2000)
6. Dittenbach, M., Rauber, A., Merkl, D.: Uncovering hierarchical structure in data using the growing hierarchical self-organizing map. Neurocomputing 48(1-4), 199–216 (2002)
7. Fanning, K.M., Cogger, K.O.: Neural network detection of management fraud using published financial data. International Journal of Intelligent Systems in Accounting, Finance & Management 7(1), 21–41 (1998)
8. Hoogs, B., Kiehl, T., Lacomb, C., Senturk, D.: A genetic algorithm approach to detecting temporal patterns indicative of financial statement fraud. Intelligent Systems in Accounting Finance and Management 15(1/2), 41–56 (2007)
9. Kirkos, E., Spathis, C., Manolopoulos, Y.: Data Mining techniques for the detection of fraudulent financial statements. Expert Systems with Applications 32(4), 995–1003 (2007)
10. Loebbecke, J.K., Eining, M.M., Willingham, J.J.: Auditors' experience with material irregularities: frequency, nature, and detectability. Auditing 9(1), 1–28 (1989)
11. Persons, O.S.: Using financial statement data to identify factors associated with fraudulent financial reporting. Journal of Applied Business Research 11(3), 38–46 (1995)
12. Rauber, A., Merkl, D., Dittenbach, M.: The Growing Hierarchical Self-Organizing Map: Exploratory Analysis of High-Dimensional Data. IEEE Transactions on Neural Networks 13(6), 1331–1341 (2002)
13. Shih, J.-Y., Chang, Y.-J., Chen, W.-H.: Using GHSOM to construct legal maps for Taiwan's securities and futures markets. Expert Systems With Applications 34(2), 850–858 (2008)
14. Tipgos, M.A.: Why management fraud is unstoppable. CPA Journal 72(12), 34–41 (2002)
15. Virdhagriswaran, S., Dakin, G.: Camouflaged fraud detection in domains with complex relationships. In: The Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (2006)

Identifying Firm-Specific Risk Statements in News Articles

Hsin-Min Lu[1], Nina WanHsin Huang[1], Zhu Zhang[1], and Tsai-Jyh Chen[2]

[1] Management Information Systems Department, The University of Arizona
1130 E. Helen Street, Tucson, Arizona 85721
hmlu@email.arizona.edu, wanhsin.huang@gmail.com,
zhuzhang@u.arizona.edu

[2] Department of Risk Management and Insurance, National Chengchi University
No. 64, Sec.2, ZhiNan Rd., Wenshan District, Taipei City 11605, Taiwan
tjchen@nccu.edu.tw

Abstract. Textual data are an important information source for risk management for business organizations. To effectively identify, extract, and analyze risk-related statements in textual data, these processes need to be automated. We developed an annotation framework for firm-specific risk statements guided by previous economic, managerial, linguistic, and natural language processing research. A manual annotation study using news articles from the Wall Street Journal was conducted to verify the framework. We designed and constructed an automated risk identification system based on the annotation framework. The evaluation using manually annotated risk statements in news articles showed promising results for automated risk identification.

Keywords: Risk management, epistemic modality, evidentiality, machine learning.

1 Introduction

Risk management has long been a topic of interest for researchers and an important issue for business professionals. From the corporate governance prospective, risk management can help managers identify important adverse future events a company may be facing and help establish procedures to measure, report, mitigate, and manage risk. For investors who hold the stocks or bonds of a company, risk management can help them assess potential losses and adjust their portfolios accordingly.

To be able to achieve the expected benefits of risk management, one needs to be able to collect and analyze relevant information from a broad range of data sources. Many of these data sources can be found within a company. For example, records for IT failure, system downtime, or errors caused by production systems [1] are valuable indicators for risks related to production technology. Public data sources such as newspapers and newswires, on the other hand, also play an important role in providing critical information for risk management. For example, various mergers, acquisitions, and business development events, which are important indicators for strategic risk [1, 2], can be found in public data sources.

H. Chen et al. (Eds.): PAISI 2009, LNCS 5477, pp. 42–53, 2009.
© Springer-Verlag Berlin Heidelberg 2009

One important characteristic of public data sources is that the majority are textual data. News articles typically report a company's past, current, and possible future events considered important for its stakeholders. The reporters of these events, based on their professional judgment, may hint at the potential impacts of these events. Investors need to digest these textual data before meaningful decisions can be made.

The advance of information technology has made the task of information retrieval much easier. For example, by using popular search engines, one can easily locate a set of documents relevant to a company. Only some of these documents, however, are relevant for a user interested in assessing a company's risk factors. To the best of our knowledge, there are few information systems that can help their users further refine and digest risk-related information. Current technology limitations force users to conduct manual analysis on a set of documents that is beyond their capacity.

This information digesting bottleneck may be addressed by systematically studying the characteristics of risk and how the messages are conveyed through textual expressions. Eventually, an information system can be built to help investors better analyze the information and make educated decisions.

Few previous studies have addressed the problem of identifying risk-related statements in textual data or attempted to construct information systems that can assist users in performing relevant tasks. In this study, we proposed a framework for identifying risk-related statements based on previous economic, managerial, linguistic, and natural language processing research. We conducted a preliminary study to verify the framework by manually annotating news articles from the *Wall Street Journal.* A prototype information system that can identify risk-related statements was designed and evaluated.

The rest of the paper is organized as follows. Section 2 summarizes previous studies that are relevant to our goal. Section 3 presents research gaps and questions. We present the annotation framework for risk-related statements in Section 4. Section 5 summarizes the annotation results. In Section 6, we describe the information system that can automatically identify risk-related statements in textual data. The system is evaluated using the human annotated data created in Section 5. We conclude the discussion in Section 7.

2 Background

Rich literature exists for risk management. Most studies have analyzed the problems from an economic or managerial perspective. The other branch of literature comes from current linguistic and natural language processing research. We summarize related studies in this section.

2.1 Definition of Risk

Risk is the possibility of loss or injury [3]. In the context of risk management for business organizations, it can be interpreted as the events and trends that can devastate a company's growth trajectory and shareholder value [1, 4]. Risk is also interpreted simply as "uncertainty" in some microeconomic discussions [5]. These definitions reveal three basic dimensions of risk: timing, uncertainty, and company value.

The first dimension is timing: risk can only exist in events that have not yet happened. The second dimension, uncertainty, is one of the most important characteristics of risk. Uncertainty can be interpreted as the possibility that more than one outcome may occur, which is a typical setting for discussing the decision making process under uncertainty (see, for example, [5]). Finally, risk must have certain impact on company value. The company value can be defined as the company's expected market value.

Note that some studies only consider losses as a contributing factor of risk. Other authors adapt a broader definition and consider the deviation from expectation as risk. For purposes of our study, we adopted the narrower definition and consider an event risky if and only if it may occur in the future and have a negative impact on company value.

2.2 Natural Language Processing Perspective

Despite the lack of direct treatments of risk-related statements in textual data, some studies in natural language processing (NLP) did provide relevant insights. We summarize here the research of subjectivity identification and certainty identification that are related to risk assessment.

Subjectivity Identification. The studies of subjectivity in NLP focus on recognizing expressions that are used to express opinions, emotions, evaluations, and speculations [6, 7]. Articles from newspapers or newswires often contain both subjective and objective expressions. Distinguishing these two kinds of expressions is valuable in applications such as information retrieval or document summarization.

Early studies of subjectivity identification focused on document and sentence level annotation. Annotators were asked to make the judgment of "whether or not the primary intention of a sentence is to objectively present material that is factual to the reporter" [6]. Later studies moved into expression level (words and phrases). Detailed annotation schemes for subjective sentences including the source and target of private state, together with the intensity and polarity of private state used [7]. Note that uncertainty is considered one kind of subjective information. However, this dimension is not emphasized in this line of research.

Certainty Identification. Certainty is the quality or state of being free from doubt, especially on the basis of evidence about the past, present, or future factual or abstract information, expressed by the writer or reported by the writer about others, directly or indirectly involved in the events in the narrative. Certainty is a pragmatic position instead of a grammatical feature.

To capture certainty from textual expressions, Rubin and Liddy [8] proposed a categorization model that characterizes certainty by four dimensions: degree of certainty, perspective of narration, focus, and timing. Four degrees of certainty (absolute, high, moderate, and low) were proposed. They also annotated whether the message was reported from the writer's point of view or from a third party's perspective. Focus was divided into abstract information (opinions, judgments) or factual information (concrete facts). Finally, past timing referred to completed or recent states or events; present timing referred to current, immediate, and incomplete states or events; and future timing referred to predictions, plans, warnings, etc.

Previous studies manually annotated news articles at both the expression and sentence levels [8-10]. Although expression level annotation can provide the richest information, the inter-rater agreement, measured by Cohen's kappa, was low [10]. It indicates that certainty information may be the results of a complicated interaction of various expressions. Although sentence level inter-rater agreement has not been studied yet, it is reasonable to expect better results.

2.3 Linguistic Perspective

Epistemic modality and evidentiality are two semantic categories that are closely related to our study. Epistemic modality is speaker's evaluation toward an event while evidentiality is concerned with the source of information. We discuss them below.

Epistemic Modality. Epistemic modality is "concerned with the speaker's assumptions, or assessment of possibilities, and, in most cases, it indicates the speaker's confidence or lack of confidence in the truth of the proposition expected" [11]. Linguistic devices for epistemic modality include: modal auxiliaries (e.g. could, may, should), epistemic lexical verbs (e.g. appear, seem, suggest), epistemic adverb (e.g. perhaps, possibly, supposedly), epistemic adjectives (e.g. likely, doubtful, apparent), and epistemic nouns (e.g. proposal, suggestion, probability).

Evidentiality. Evidentiality can be narrowly defined as the source of information. A wider definition also include the speaker's attitude (e.g. attested, reported, inferring) toward the source of information. Some authors believe that evidentiality is part of epistemic modality while others believe that these two concepts can be separated . The distinction can be made by noting that "evidentials assert the nature of the evidence for the information in the sentence, while epistemic modals evaluate the speaker's commitment for the statement." [12].

3 Research Gaps and Questions

It is clear from the above discussion that most risk management studies incorporated only numerical accounting and capital market data. Little research has systematically investigated suitable dimensions for risk-related text quantification. We have seen few studies that attempted to develop automatic information systems to help identify and quantify risk-related information.

We notice that, from the previous NLP studies, expression level annotation is the most time consuming and most difficult to conduct. Despite the rich information one may obtain, expression level annotation results are much noisier than the sentence level and document level annotations. Since sentences in a document may or may not contain risk-related information, document level annotation is not a reasonable choice. Sentence level annotation, on the other hand, can provide detailed information with less noise. The results can be aggregated to the document level or, with the help of various statistical methods, drilled down to the expression level.

News articles are one popular data source for obtaining risk-related information. As a result, we chose to focus on firm-specific news articles from the *Wall Street Journal*, one of the most widely read newspapers today.

We set forth a preliminary study that aimed at: 1) developing a suitable framework for quantifying sentence-level risk-related information in news documents, and 2) developing an information system for automatic risk identification and quantification.

4 Risk Quantification Framework

We developed the risk quantification framework based on the characteristics of risk defined in economic and managerial literature as well as various semantic categories mentioned in previous NLP and linguistic studies.

A news sentence can be analyzed using the following five dimensions: timing, epistemic modality, evidentiality, abstract or factual information, and polarity. For the purpose of extracting risk-related information, we focused on the potential impact on three key aspects: future timing, uncertainty, and company value. A risk-related statement can be defined as a sentence that includes future timing, indicates uncertainty, and implies negative impact on company value. We will discuss the framework in detail using sample news articles statements below.

Kodak, based in Rochester, N.Y., said it ***expects*** *net earnings of $1.15 to $1.30 cents a share for the full year.* (1)

Mercedes ***will*** *invite potential buyers to hot restaurants and special concerts where it will let them test-drive a C240 or C320.* (2)

Analysts had forecast the company would report earnings of 90 cents a share. (3)

Future Timing. Future timing refers to the expressions that indicate (possible) upcoming events or states. For instance, "expects" in (1) and "will" in (2) indicate future timing. Note that (3) does not have future timing because the forecast is for past company earnings. Binary classification (yes/no) is used for this dimension in our annotation study.

Although Univision has emerged as the ***likely*** *buyer, Disney and Tribune also have expressed interest.* (4)

National Semiconductor Corp. cut its ***revenue forecast*** *for the fiscal second quarter, citing inventory and backlog adjustments from customers and distributors.* (5)

Uncertainty. Uncertainty can come from epistemic modality, evidentiality, or both. The expression "likely" in (4) is an example of uncertainty that comes from epistemic modality. In (5), "revenue forecast" indicates the potential uncertainty that is associated with the source of the information.

Various levels of uncertainty can be inferred from the expression. Previous studies had tried to divide the certainty-uncertainty continuum into four or five categories [8-10]. As a preliminary study, we decided to perform a binary classification for this dimension first. A sentence is classified as uncertain or not uncertain.

Disappointing *ICD sales were offset by a 7% **increase** to $248 million in sales of pacemakers.* (6)

Brown ***boosted*** *its* ***rating*** *on the Denver telecommunications services provider.* (7)

Boeing ***fell*** *2.55 to 58.02.* (8)

Company Value. This dimension captures how the information impacts the market's expectation of the company value. Both abstract and factual information can contribute to this dimension. The polarity in a sentence may also hint at the direction of the impact. It should be pointed out that this dimension cannot be treated as a three-way classification (good, bad, no effect), as a result of our focus on sentence-level information. One sentence may have both good and bad implications. For example, the first half of (6) has a negative implication while the second half has a positive implication. One possibility is to annotate for the net effect on company value. However, the meaning of the original message may be distorted by assuming that only the net effect matters.

We therefore considered company value along two separate dimensions: good news and bad news. A sentence belongs to the good news category if the underlying message has a positive implication for company value and vice versa. A sentence can belong to both categories simultaneously.

Another complication arose from the complex nature of assessing company value. Assessments may differ because of different levels of world knowledge of annotators. It is unrealistic to assume that an annotator (and for that matter, a computer algorithm) can have perfect knowledge about the world. A practical solution is to assume that the annotator has basic business knowledge but does not possess detailed knowledge of a particular event. Under this assumption, (7) is good news because the company rating is a good indicator for the overall financial status of a company; (8) is bad news because stock price reflects current market value of a company.

5 A Manual Annotation Study Using the Wall Street Journal

We annotated news sentences from the Wall Street Journal (WSJ) based on the framework proposed in the previous section. The research test bed is described first, followed by a discussion of the annotation results.

5.1 Research Testbed

We used the following procedure to create our research test bed. First, we drew a random sample of 200 news articles from the WSJ published between 8/4/1999 and 3/2/2007. One annotator manually filtered out non firm-specific news, leaving 103 firm-specific articles. For each firm-specific article, the full text was split into sentences. Each sentence was attached to the original document ID. An Excel file that contained the sentences and document IDs was created. The original order of the sentence in an article was preserved. Only the first 988 sentences from 46 articles were made available to the annotator.

The distribution of article length (measured by # of sentences) is bimodal. A large number of articles contained less than 10 sentences. Another peak was at the group of 30-40 sentences per article. Separating the two groups by a threshold of 20 sentences per article, we can see that 28 out of 46 articles had length less than or equal to 20. However, this group only contributed to 235 of the total 988 sentences.

The annotator read the sentences in the Excel file in sequence. The experience was much like reading the original article except that the sentences had been split and listed row by row. Each sentence was annotated with four 0s or 1s indicating whether it belonged to the four dimensions proposed: future timing, uncertainty, good news, and bad news.

5.2 Annotation Results

On average, 14.3% of sentences were marked with future timing, 17% uncertainty, 12.9% good news, and 24.6% bad news. The proportion of risk-related sentences (those with future timing, uncertainty, and bad news simultaneously) was only 4.7%. The number of bad news sentences was about twice as much as those with good news.

Separating the collection into short and long articles (at the cutoff point of 20 sentences per article), we can see clearly that these two types of news articles had different distributions in terms of the four dimensions under consideration. Short articles contained more future timing and uncertainty than the longer ones. On the other hand, bad news sentences were more prevalent in long articles. Interestingly, the proportion of risk-related sentences was about the same in these two types of articles.

The differences may reflect the nature of these two types of articles. Long articles, in many cases, contained comments and analysis of current and past events. These articles might be, in part, stimulated by recent bad news about a company. Short articles, in many cases, were quick notices of recent and future events.

6 An Automatic Risk Identification System

Based on the risk quantification framework and the manual annotation results, we designed a risk identification system to extract risk-related information at the sentence level using the four dimensions proposed. Each dimension was separately modeled.

For each input news articles, the system first splits the full text into individual sentences. A feature extraction routine then converts each sentence into the bag-of-words representation. Four binary classifiers are used to identify information related to future timing, uncertainty, good news, and bad news. The scores from the 4 classifiers then can be used to compute the risk scores of a sentence.

It is clear from the design that the performance of the system depends heavily on the four classification modules. As a preliminary study, we decided to evaluate the system performance based on the four classifiers individually.

6.1 Features

We adopted the baseline bag-of-words representation, together with the part of speech (POS) tags. The Porter stemmer was used to find the lemma of individual words [13]. Four combinations of features are considered:

1. Bag-of-words (no stemming)
2. Bag-of-words (stemmed)

3. Bag-of-words (no stemming) with POS tags
4. Bag-of-words (stemmed) with POS tags

When POS tags were considered, the same word with different POS tags was treated as a different feature.

6.2 Machine Learning Approaches

We considered two popular machine learning approaches in this study. The first approach is the maximum entropy (Maxent) model [14]. This model is mathematically equivalent to the discrete choice model. To handle the large number of features, Gaussian priors with mean 0 and variance 1 is imposed on the parameters. The second model considered is the support vector machine (SVM) classifier with a linear kernel [15].

6.3 Baseline Models

We considered two baseline models in this study. The first baseline model is the majority classifier. The majority classifier classifies each instance to the majority class. For example, 14.3% of sentences in the test bed have future timing. The majority classifier assigns all sentences to the class of no future timing since the majority of sentences do not have this tag.

The second baseline classifier is the agnostic classifier. The agnostic classifier assigns a random score between 0 and 1 to each instance. Given a threshold, all instances above the threshold are positively classified while instances below the threshold are negatively classified. The ROC curve of an agnostic classifier is a straight line connecting the origin to (1, 1) [16]. Note that the majority classifier is a special case of the agnostic classifier. Depending on whether the negative or positive tagging is the majority class, an agnostic classifier with cutoff equals 0 or 1 is equivalent to the majority classifier.

6.4 Performance Measures

We considered accuracy, recall, precision, and F-measure in this study. Let TP, FP, TN, and FN denote true positive, false positive, true negative, false negative. These measures can be computed as follows:

Accuracy = (TP+TN) / (TP+FP+TN+FN)
Recall = TP / (TP+FN)
Precision = TP / (TP+FP)
F-measure = 2 * Precision * Recall / (Precision + Recall)

Accuracy is the probability that a case is correctly assigned. Recall is the probability that a positive case is correctly assigned. Precision is the probability that an assigned positive case is correct. F-measure is the harmonic mean of precision and recall. Most of these measures had been used in prior classification studies [17, 18].

6.5 Experimental Results

We used 10-fold cross validation to test the performance of the 4 classifiers. Different performance can be achieved by choosing different threshold values. To make the subsequent discussion easier, we present the figures that maximized the F-measure for each classification task. Tables 1 and 2 summarize the performance of the 4 classification tasks. The bag-of-words features with no stemming were denoted as "B" in the parentheses of the first column; "BS" denotes bag-of-words with stemming. We reported figures that maximized the F-measure. Adding POS tags had only minor impacts on the performance. The outcomes using POS tags, as a result, were omitted to save space.

The majority classifiers have accuracy rates of 85.7%, 83%, 75.4%, and 87.1% for future timing, uncertainty, bad news, and good news, respectively. However, since the majority classes are all negative (the absent of a characteristic), the recalls were all zero and the precisions could not be calculated (since there were no positively assigned cases). Compared to the performance of the Maxent and SVM models, the accuracy was similar but the recalls were much lower.

The agnostic classifier, which can be considered as a generalization of the majority classifier, achieved much lower performance compared to the Maxent and SVM models. The accuracy across four classification tasks was at the range of 21.6% to 33.7%. The F-measures were all less than 39.5%. It is interesting to observe that the recall rates were much higher than those for precision. One possible reason is that the agnostic classifier cannot distinguish cases with different tags. As a result, the only way to boost the F-measure is to have positive assignments to most instances.

The SVM model outperformed the Maxent model in all feature-task combinations. On average, the F-measures of the SVM model were about 10% higher than those of the Maxent model. In most cases, the performance gaps came from higher recall of the SVM model.

In most classification tasks (future timing, uncertainty, and good news), stemming improved the performance. However, the best performance of bad news classification came from bag-of-words without stemming. Error analysis shows that stemming mapped words with different semantics to the same lemma (e.g. "willfully" to "will") caused false positives in subsequent classification.

Table 1. Performance Summary for Future Timing and Uncertainty

	Future Timing				Uncertainty			
	Acc.	Recall	Prec.	F	Acc.	Recall	Prec.	F
Majority	85.7	0	NA	NA	83.0	0	NA	NA
Agnostic	22.4	92.9	14.8	25.5	21.6	97.0	17.5	29.6
Maxent (B)	91.7	57.4	78.6	66.4	82.0	42.3	46.7	44.4
SVM (B)	93.1	75.2	76.3	75.7	83.4	72.0	50.8	59.6
Maxent (BS)	91.8	55.3	81.3	65.8	82.3	42.9	47.7	45.1
SVM (BS)	**94.0**	**75.9**	**81.1**	**78.4**	**87.9**	**56.0**	**67.1**	**61.0**

Table 2. Performance Summary for Bad News and Good News

	Bad News				Good News			
	Acc.	Recall	Prec.	F	Acc.	Recall	Prec.	F
Majority	75.4	0	NA	NA	87.1	0	NA	NA
Agnostic	24.6	100.0	24.6	39.5	33.7	80.3	13.9	23.7
Maxent (B)	73.3	44.4	45.6	45.0	80.0	47.2	31.4	37.7
SVM (B)	**73.4**	**62.6**	**46.9**	**53.6**	88.1	40.9	54.7	46.8
Maxent (BS)	62.3	68.3	36.0	47.2	83.9	40.9	38.2	39.5
SVM (BS)	72.1	64.2	45.2	53.1	**88.4**	**45.7**	**55.8**	**50.2**

To gain a better understanding of the classification task, we computed the conditional probability associated with each lemma from the Maxent model. The model is trained using all training data with words stemmed. The conditional probability is the probability that the Maxent model will have positive assignment given a particular lemma. For example, given the lemma "outlook," we computed the conditional probability that a sentence belonged to the future timing category given that the sentence contained only the word "outlook." A similar procedure was repeated for the remaining three dimensions. The conditional probability is a good proxy for the importance of each lemma.

We sorted the conditional probability of each lemma in descending order. Top 10 lemmas from each classifier were analyzed. We observe from the results that lemmas such as "will," "expect," and "estim" were good indicators for future timing. Lemmas such as "expect," "if," and "may" hinted at uncertainty. Bad news sentences may contain lemma such as "fall," "problem," or "risk" while good news sentences may contain lemma such as "strong," "unit," or "rose."

Note that some lemmas were not semantically related to the underlying classes. For example "cable-tv" in future timing, uncertainty, and bad news did not have a clear connection to these three dimensions. One possible reason is that our training dataset was small and the Maxent model over-generalized from this particular dataset.

7 Concluding Remarks

We developed an annotation framework for risk identification based on the previous literature. The framework models risks along four dimensions: future timing, uncertainty, good news, and bad news. We applied the framework on firm-specific news articles from the *Wall Street Journal*. The annotation results showed that bad news is the most commonly annotated dimension across the four dimensions considered. We designed an automatic risk identification system based on the annotation framework and trained the underlying classifiers using the manual annotation results.

Using the bag-of-words representation, we achieved F-measures between 50.2% and 78.4% for the four classification tasks under consideration. Important features of these four classifiers showed consistent semantics as indicated by the definitions of these four dimensions. The results are promising for the development of a full-fledged system.

We are currently recruiting and training more human annotators to conduct manual annotation based on the proposed framework. The validity of the framework can be further confirmed by analyzing the annotation results from multiple sources. We also plan to continue the research by developing an information system that can automatically identify risk-related statements from various business-related news sources.

Acknowledgments. This work was supported in part by Taiwan National Science Council through Grant #NSC97-2410-H002-002 and U.S. National Science Foundation through Grant #IIS-0428241 ("A National Center of Excellence for Infectious Disease Informatics"). The authors wish to thank Dr. Hsinchun Chen and Dr. Shu-Hsing Li for their constructive comments.

References

1. Slywotzky, A.J., Drzik, J.: Countering the Biggest Risk of All. Harvard Business Review, 1–11 (2005)
2. Chapelle, A., Crama, Y., Hubner, G., Peters, J.-P.: Practical methods for measuring and managing operational risk in the financial sector: A clinical study. Journal of Banking & Finance 32, 1049–1061 (2008)
3. Merriam-Webster: Risk: Definition from the Merrian-Webster Online Dictionary (2008)
4. COSO: Enterprise Risk Management - Intergrated Framework. COSO(Committee of Sponsoring Organizations of the Treadway Commission) (2004)
5. Mas-Colell, A., Whinston, M., Green, J.R.: Microeconomic Theory, Oxford (1995)
6. Bruce, R.F., Wiebe, J.M.: Recognizing subjectivity: a case study of manual tagging. Natural Language Engineering 1, 1–16 (1999)
7. Wiebe, J., Wilson, T., Cardie, C.: Annotating Expressions of Opinions and Emotions in Language. Language Resources and Evaluation 39, 165–210 (2005)
8. Rubin, V.L., Liddy, E.D.: Certainty Identification in Texts: Categorization Model and Manual Tagging Results. In: Shanahan, J.G., Qu, Y., Wiebe, J. (eds.) Computing Attitude and Affect in Text: Theory and Applications. Springer, Heidelberg (2005)
9. Rubin, V.L.: Certainty Categorization Model. In: AAAI Spring Symposium: Exploring Attitude and Affect in Text: Theories and Applications, Stanford, CA (2004)
10. Rubin, V.L.: Starting with Certainty or Starting with Doubt: Intercoder Reliability Results for Manual Annotation of Epistemically Modalized Statements. In: Proceedings of NAACL HLT 2007, pp. 141–144 (2007)
11. Coates, J.: Epistemic Modality and Spoken Discourse. Transactions of the Philological Society 85, 110–131 (1987)
12. de Haan, F.: Evidentiality and Epistemic Modality: Setting Boundaries. Southwest Journal of Linguistic 18, 83–101 (1999)
13. Porter, M.F.: An Algorithm for Suffix Stripping. Program 14, 130–137 (1980)
14. Berger, A.L., Pietra, V.J.D., Pietra, S.A.D.: A maximum entropy approach to natural language processing. Comput. Linguist. 22, 39–71 (1996)
15. Joachims, T.: Making large-Scale SVM Learning Practical. In: Schölkopf, B., Burges, C., Smola, A. (eds.) Advances in Kernel Methods - Support Vector Learning. MIT-Press, Cambridge (1999)

16. Cortes, C., Mohri, M.: Confidence Intervals for the Area under the ROC Curve. In: Advances in Neural Information Processing Systems (NIPS 2004), vol. 17. MIT Press, Vancouver (2005)
17. Abbasi, A., Chen, H.: Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. ACM Trans. Inf. Syst. 26, 1–29 (2008)
18. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. Elsevier, Amsterdam (2005)

Predicting Future Earnings Change Using Numeric and Textual Information in Financial Reports

Kuo-Tay Chen[1,*,**], Tsai-Jyh Chen[2], and Ju-Chun Yen[1]

[1] Department of Accounting, College of Management,
National Taiwan University
ktchen@management.ntu.edu.tw
[2] Department of Risk Management and Insurance,
National Chengchi University

Abstract. The main propose of this study is to build a more powerful earning prediction model by incorporating risk information disclosed in the textual portion of financial reports. We adopt the single-index model developed by Weiss, Naik and Tsai as a foundation. However, other than the traditionally used numeric financial information, our model adds textual information about risk sentiment contained in financial reports. We believe such a model can reduce specification errors resulting from pre-assuming linear relationship, thus can predict future earnings more accurately. The empirical results show that the modified model does significantly improve the accuracy of earning prediction.

Keywords: Single-index model, earnings prediction, risk sentiment, textual information.

1 Introduction

The neoclassical security valuation model determines a firm's value as the present value of expected future dividends. The ability of a firm to distribute dividends in the future can be assessed by its expected future earnings. As a result, future earnings prediction has become an important research issue. A number of studies have employed various factors in their models to predict earnings. These factors include bottom-line number of income statement (e.g. time-series pattern of earnings)[2], components of earnings [8], and accounting ratios based on income statement and balance sheet [13]. These studies utilize only numeric information in financial reports and do not incorporate textual information in their models. Since textual information such as footnotes and management discussion and analysis (MD&A) contain lots of information related to future earning, these models might have reduced their prediction power by excluding textual information.

Previous studies show that managers may have incentives to disclose risk in order to reduce the influence of future earnings shock and to avoid litigation responsibilities

[*] Corresponding author.

[**] Corresponding address: 1 Roosevelt Road, Sec. 4, Taipei, Taiwan, 106.

H. Chen et al. (Eds.): PAISI 2009, LNCS 5477, pp. 54–63, 2009.
© Springer-Verlag Berlin Heidelberg 2009

[14]. Recently Li (2006) [12] uses risk-related words in financial reports as a measurement of risk sentiment and finds that risk sentiment is associated with future earnings and can be used to predict the change of earnings. This implies that risk sentiment in the textual part of financial reports may have information content about future earnings. As a result, our study builds an earning prediction model by incorporating the risk sentiment information contained in financial reports. Moreover, we do not assume that the risk sentiment has linear relation with future earnings. Because the disclosure of risk is manager's decision, future earnings may not definitely decrease when managers disclose low risk. On the other hand, large decrease of future earnings may be expected when high risk is disclosed. In other words, the relationship between risk sentiment and future earnings cannot be clearly specified as a linear relationship. Therefore, we might create a specification error if we pre-assume linear regression model without knowing the true relationship.

In this paper, we construct the research model based on Weiss, Naik and Tsai (2008) [16]. They employed a new method, single-index model, to estimate the market expectations of future earnings change (called Market-Adapted Earnings, MAE in [16]) and used MAE to predict future earnings. Compared with previous studies, single-index model allows a nonlinear relation between dependent and independent variables. Hence it can reduce the specification error committed by the pre-assumed linear model. Weiss et al. [16], however, consider merely numeric accounting ratios while assessing MAE. As we mentioned before, the textual part of financial reports may also contain useful information for market participants to predict future earnings. For this reason, we adjust Weiss et al. [16] model by incorporating the textual portion of financial reports to re-examine MAE, and hope to build an earnings predicting model with more predicting power.

In summary, we design our study in two parts. First, we construct a single-index model that includes a risk sentiment variable which represents textual information in financial reports. From this model we assess MAE to predict future earnings change. Second, we calculate predicted market expectation of future earnings change (predicted MAE) and compare it with those generated by Weiss et al. predicting model to determine the relative strength of our model.

2 Literature Review and Hypothesis Development

In this section we first review previous studies on earnings prediction and textual information. We then develop our research hypotheses.

2.1 Earnings Prediction

2.1.1 Factors to Predict Earnings

In 1970s, several studies try to find out the time-series pattern of earnings (e.g. Ball and Watts 1972 [2]). Those studies suggest that earnings process is close to random walk with a drift. Beaver, Lambert and Morse (1980) [3] assert that earnings are a compound process of the earnings series which will reflect events affecting price and those which will not affecting price. Those studies use the bottom-line earnings number to predict future earnings. Brown et al. (1987) [5] compare the quarterly earnings predicting results of three different time-series earnings model and analyst earnings forecast from Value Line. Among one-quarter ahead through three-quarter ahead earnings predicting results, however, all of the three time-series earnings model do not outperform than analyst forecasts by examining forecast error.

After that, some researchers try to construct more specific and accurate earnings predicting models. Rather than using bottom-line earnings, Fairfeild et al. (1996) [8] suggest that different components of earnings may contain different content to predict earnings. They disaggregate earnings into several components in different way to see which approach of classification will provide the best ability of predicting one-year ahead return on equity. The result shows that non-recurring items of earnings do not have better predicting ability than recurring items. In addition, their model performs better than the model with aggregated bottom-line item.

Besides, since the time-series of earnings model only extract the information from income statement, several studies turn to exam the content of the balance sheet items and accounting ratios to predict future earnings. Ou (1990) [13] uses non-earnings annual report numbers to predict earnings change. In this study, logit model is used to predict the sign of earnings change in next period. Lev and Thiagarajan (1993) [11] choose 12 accounting ratios as candidates and calculate an aggregate fundamental quality score. First they assign 1 to positive signal and 2 to negative signal for each of the 12 accounting ratios, and then compute an average score for each firm and year. They find that this fundamental quality score has positive relation with subsequent earnings changes. Abarbanell and Bushee (1997) [1] employ an earnings predicting linear model with accounting ratios tested by Lev and Thiagarajan [11].

Other than accounting information, Beaver et al. (1980) [3] suggest that price reflect the information not presented in current earnings. As a result, price may also contain the information content and lead earnings. Beaver et al. (1987) [4] use percentage of price change as an independent variable to predict percentage of earnings changes.

2.1.2 Statistics Methods to Predict Earnings

The previous subsection focuses on the factors or elements which can be used to approximate future earnings. The other studies line is to adopt different research methods to improve the ability of predicting earnings. Most of the previous studies use linear OLS regression model (e.g. [3], [4], [8]), logit model (e.g. [13]). However, it is unknown that what the true pattern of relation between the predicting factors and future earnings is. For example, some have found S-shape relationship in returns-earnings relationship studies [6]. Therefore, the use of parametric estimation model may result in the problem of specification error.

The main propose of Weiss et al. (2008) [16] is to develop a new index to extract forward-looking information from security price. This study originates from the price lead earnings researches of Beaver et al. who suggest that price may have information content about future earnings. Thus, Weiss et al. take a semi-parametric statistic method, single-index model, to connect the relation between earnings change and returns via market expected earnings change. It extracts information from both security prices and accounting ratios to calculate a new index representing market's expectations of future earnings change, which they called MAE. The use of single-index model allows for a nonlinear returns-earnings relationship. They use earnings change and four accounting ratios such as change in inventory, change in gross margin, change in sales and administrative expenses and change in accounts receivable as independent variables and annual returns as dependent variable. This paper has two contributions: (1) the single-index model performs more explanatory power than the linear model with fundamental accounting signals by Lev and Thiagarajan (1993)[11], and (2) the predicted MAE index, which is used to predict future earnings change, has better forecasting ability than previous random-walk model and accounting-based forecasting model by Abarbanell and Bushee (1997) [1]. But still, MAE index does not outperform analyst earnings forecast from I/B/E/S.

2.2 Textual Information

Li (2006) [12] examines the relation between risk sentiment in annual reports and future earnings. Li measures risk sentiment by counting frequency of words about risk or uncertainty in textual part of annual reports. The counting rules in his paper are (1) count the frequencies of the "risk" words (including "risk", "risks" and "risky") and the "uncertainty" words (including "uncertain", "uncertainty" and "uncertainties"). (2) "risk-" format words are excluded because it may relate to "risk-free". Li finds a negative relation between risk sentiment and next period earnings.

2.3 Hypothesis Development

After MAE has been estimated, we can use predicted MAE to forecast future earnings. Because the risk sentiment in financial reports may affect future earnings rather than current earnings, it may have information content to future earnings. That means, the change of risk sentiment in financial reports may add incremental earnings prediction ability. Accordingly, the hypothesis is developed as follows:

Hypothesis: The earnings forecast errors of the modified model with risk sentiment variable are lower than previous models.

3 Research Design

3.1 Introduction to Single-Index Model

As we mentioned in previous sections, using parametric method when unknowing the true pattern may cause specification errors. Although nonparametric method can be used to solve this problem, but the cost of reducing specification errors can be very high because (1) estimation precision will decrease when the dimension of independent variables increase, (2) it's hard to interpret the results in multidimensional independent variables, and (3) it cannot be used to predict [9]. Since that, semi-parametric method can both solve the problem from nonparametric methods and reduce the specification errors from parametric methods.

Single-index model is one of the semi-parametric models. It aggregates the multi-dimensional X into single-dimensional index first, and then estimates the function connecting the dependent variable and the index by parametric estimation methods. The basic form of single-index model is as follows:

$$Y_i = G(X\beta) + \varepsilon_i \tag{1}$$

where â is the vector of coefficients and *X*â is the index. Note that we do not have to assume the type of G(.) in priori. G(.) can be an identity function (then the function (1) becomes linear model), cumulative distribution function (then the function (1) becomes probit model) or nonlinear function. In turn, G(.) is determined endogenously.

In this paper, we employ the same estimating method with Weiss et al. [16] as follows:

1. Estimating â: There are different approaches can be adopted to estimate â(e.g. nonlinear least square), even without knowing G(.). One simple method without solving optimal problem is sliced inverse regression [7]. First the data should be sorted by the increasing value of dependent variable, and then divided into several groups, or slices. After slicing, calculate a new covariance matrix by slice means and then estimateâ. This method is without solving optimal problem and link-free. That is, â can be estimated when G(.) is unknown.
2. Estimating G(.): Afterâis estimated, the index *X*â can be calculated. As a result, we can use *Y* and *X*â to estimate G(.) by nonparametric method since the multidimensional *X* has been aggregated into single dimensional index.

3.2 Model Construction

3.2.1 Single-Index Model

Following Weiss et al. [16], we also release the relation pattern between returns and earnings by setting G(.) allowing linear or nonlinear relation. However, we want to examine whether the risk sentiment in the textual part of financial reports has information content to future earnings. Therefore, we incorporate additional variable to capture risk sentiment in 10-K reports in Weiss et al. model and construct as follows:

$$\begin{aligned} & R_{it} = G(MAE_{it}) + \varepsilon_{it} \\ & \text{Where } MAE_{it} = \Delta E_{it} + \beta_{1t}\Delta INV_{it} + \beta_{2t}\Delta GM_{it} + \beta_{3t}\Delta SGA_{it} + \beta_{4t}\Delta REC_{it} + \beta_{5t}\Delta RS_{it} \end{aligned} \tag{2}$$

- R_{it}: Annual abnormal returns, which is measured by accumulated 12 months of monthly raw stock returns starting from the fourth month of the beginning of fiscal year to the fourth month of the ending of fiscal year, and then less the equally weighted monthly returns for the same periods in CRSP [16].
- ΔE_{it}: Change of earnings per share before extraordinary items and discontinued operations, deflated by the stock price at the beginning of fiscal year. [16] [10] Note that we set the coefficient of ΔE_{it} equals one for scale normalization. [9]

- ΔINV_{it} : Change in inventory measured by $\Delta Inventory - \Delta Sales$ [1], which is a signal of logistic operations[2]. (Compustat #78 or 3, #12)
- ΔGM_{it} : Change in gross margin measured by $\Delta Sales - \Delta GrossMargin$, which is a signal of profitability of sales. (#12,#12-41)
- ΔSGA_{it} : Change in sales and administrative expense measured by $\Delta SalesAndAdministrativeExpense - \Delta Sales$, which is a signal of marketing and administrations. (#189, #12)
- ΔREC_{it} : Change in accounts receivable measured by $\Delta AccountsReceivable - \Delta Sales$, which is a signal of management of clientele.
- ΔRS_{it} : Change in risk sentiment in the MD&A and footnote parts of annual report. This variable is used to capture the annual reports' information which will affect future earnings but not recognized in financial statements yet. In other words, this variable presents textual information rather than numerate information in annual reports. Following Li [12], ΔRS_{it} is calculated as follows:

$$\Delta RS_{it} = \ln(1 + NR_{it}) - \ln(1 + NR_{it-1}) \tag{3}$$

Where NR_{it} are the numbers of occurrences of risk related words in the annual report of year *t*. The risk related words are the words including "risk" (e.g. Risk, risks, risky) and "uncertainty" (e.g. Uncertain, uncertainty, uncertainties) and excluding "risk-".

Note that based on the definitions of the above variables and the analysts' interpretation, it was defined as "good news" when the value of the variable is negative, and vise versa. As a result, we predict the signs of all the coefficients are negative, including ΔRS_{it}.

3.2.2 MAE Prediction

Since we want to compare the predicting ability of our model with Weiss et al. model, we adopt the same procedure with Weiss et al. After theβare estimated, estimated *MAE* can be calculated: [16]

$$\begin{aligned} M\hat{A}E_{it} &= \Delta E_{it} + \textstyle\sum_j S_{jit}\hat{\beta}_{jt} \\ &= \Delta E_{it} + \hat{\beta}_{1t}\Delta INV_{it} + \hat{\beta}_{2t}\Delta GM_{it} + \hat{\beta}_{3t}\Delta SGA_{it} + \hat{\beta}_{4t}\Delta REC_{it} + \hat{\beta}_{5t}\Delta RS_{it} \end{aligned} \tag{4}$$

[1] Ä means "percentage change of the variable between its actual amount and expect amount, where the expected amount is the average amount in the previous two year. E.g. $E(Sales_t) = (Sales_{t-1} + Sales_{t-2})/2$, $\Delta Sales = [Sales_t - E(Sales_t)]/E(Sales_t)$. See Lev and Thiagarajan (1993) [11].

[2] Originally we should use change in cost of goods sold as a benchmark rather than change in sales. But analysts usually use change in sales and previous study showed the identical results [11]. In order to compare with previous studies, we also choose change in sales as benchmark.

Next, in order to estimate future earnings, the estimated MAE is transformed as follows[3]: [16]

$$M\hat{A}E_{it}^{*} = M\hat{A}E_{it} \times \frac{\overline{\Delta E_{it}}}{\overline{M\hat{A}E_{it}}} = \frac{M\hat{A}E_{it}}{\overline{M\hat{A}E_{it}}} \times \overline{\Delta E_{it}} \qquad (5)$$

Where $\overline{\Delta E_{it}}$ and $\overline{M\hat{A}E_{it}}$ are the means of actual earnings change and estimated MAE. After transforming, the mean of $M\hat{A}E_{it}$ will equal the mean of ΔE_{it}.

4 Empirical Results

4.1 Data

To measure the fundamental accounting ratios, we use data from Compustat fundamental annual and calculate the percentage change of inventory, gross margin, sales and administrative expense, and receivables. In addition to the accounting ratios, we use earnings per share before extraordinary items in Compustat to calculate the change of reported earnings, which is deflated by the share price at the beginning of the fiscal year.

For abnormal stock returns, we use data of raw stock monthly returns from Compustat and equally weighted monthly return from CRSP. Abnormal annual stock returns are calculated by cumulating raw stock monthly returns minus equally weighted monthly return for 12 months (from the fourth month after the beginning of fiscal year to the third month after the end of the fiscal year) [16].

Following Li, we extract number of risk related words in 10-K reports to calculate change of risk sentiment. The counting method of risk related words is similar to Li[4].

The data period is from 1998 to 2006. The sample firm with missing data will be dropped for that sample year.

4.2 Earnings Forecast Error

In Weiss et al. 2008 paper, the predicting results of this model has been proved outperforming than random-walk model and fundamental accounting ratios model proposed by Abarbanell and Bushee (1997) [1]. Accordingly, we can only compare our model, which contains textual information in financial reports, with Weiss et al. model by median absolute forecast errors. Our purpose is to see whether risk sentiment in 10-K reports can improve earnings predicting ability; therefore, we compare earnings

[3] In Weiss et al., they state that transforming can let the sum of noise in earnings change be zero across firms. Moreover, transformed MAE is equivalent to random-walk model. [16].

[4] Different from Li, we do not delete the title items in 10-K reports before counting risk words. However, we expect similar results since we use changes of risk words but no absolute number of risk words for the year.

predicting errors of SIM model with change of risk sentiment variable with those of the original SIM model in Weiss et al.

To compare the predicting abilities among the models, absolute earnings forecast errors are calculated: [16]
Absolute earnings forecast error =

$$\left|\frac{Actual - Forecastd}{Actual}\right| = \left|\frac{\Delta E_{t+1} - Forecasted\Delta E_{t+1}}{E_{t+1}}\right| = \left|\frac{\Delta E_{t+1} - M\hat{A}E^{*}_{t+1}}{E_{t+1}}\right| \tag{6}$$

After absolute earnings forecast errors are calculated for the two models, median of the forecast errors is reported rather than mean forecast errors to prevent the influence of huge value owing to deflation.

In traditional linear earnings forecast model [1], earnings in period t and fundamental ratios in period t-1 are used to construct model, and then we can input fundamental ratios in the model to predict earnings in period t+1. That is, we need two year data to predict next year earnings. However, SIM model needs only one year data, rather than two years, to construct MAE*. Therefore, we estimate next year's earnings change using both the model construct by last year's data (out of sample prediction) and the model construct by this year's data (in sample prediction). The results are showed in table 1 and table 2.

When using last year's model and this year's data (out-of-sample prediction), incorporating additional risk sentiment reduce median forecast error by nearly 60% in full sample (from 0.146244 to 0.058998). In addition, the nine year average of median forecast error also declines by 70% (from 0.36856 to 0.10973).

Table 1. Median forecast errors of MAE* with and without use of risk sentiment. (out-of-sample)

Year	Median forecast error	
	MAE* with no risk sentiment	MAE* with risk sentiment
1998	--	--
1999	0.068852	0.082984
2000	0.065603	0.065896
2001	0.159866	0.056722
2002	1.477695	0.243611
2003	0.423893	0.364171
2004	0.722144	0.033904
2005	0.011548	0.011429
2006	0.018875	0.01912
Average	0.36856	0.10973
Full sample	0.146244	0.058998

Table 2. Median forecast errors of MAE* with and without use of risk sentiment. (in sample)

Year	Median forecast error	
	MAE* with no risk sentiment	MAE* with risk sentiment
1998	0.040108	0.015707
1999	0.073343	0.074516
2000	0.064788	0.071767
2001	0.027909	0.027053
2002	0.122653	0.119972
2003	0.687346	0.18015
2004	0.047517	0.035202
2005	0.011333	0.011427
2006	0.034801	0.077692
Average	0.123311	0.068165
Full sample	0.061508	0.055131

When using this year's model and data to prediction next's year's earnings change (in sample prediction), considering additional risk sentiment reduce median forecast error by 10% (from 0.061508 to 0.055131) in full sample. The nine year average of medina forecast error is reduced by 45% (from 0.123311 to 0.068165). Comparing to the results of out-of-sample prediction, the percentages of reduction are less than out-of-sample prediction, but the absolute values of forecast error in in-sample prediction are smaller.

As a result, the use of risk sentiment in textual part of annual reports, additional to numeric information in financial statement, can improve the earnings predicting ability.

5 Conclusion

Many studies have attempted to make better predictions of earnings. Nevertheless, numeric information may only provide partial information for future earnings. Thus, in this study we incorporate textual information in financial reports to examine whether it has incremental earnings prediction ability. The results show that incorporating risk sentiment in 10-K reports can significantly improve the one-year ahead earnings prediction ability by using a single-index model.

References

1. Abarbanell, J.S., Bushee, B.J.: Fundamental Analysis, Future Earnigns, and Stock Returns. Journal of Accounting Research 35, 1–24 (1997)
2. Ball, R., Watts, R.: Some Time Series Properties of Accounting Income. Journal of Finance, 663–682 (June 1972)

3. Beaver, W., Lambert, R., Morse, D.: The Information Content of Security Prices. Journal of Accounting and Economics 2, 3–28 (1980)
4. Beaver, W., Lambert, R.A., Ryan, S.G.: The Information Content of Security Prices: a Second Look. Journal of Accounting and Economics 9 (1987)
5. Brown, L.D., Hagerman, R.L., Griffin, P.A., Zmijewski, M.E.: Security Analyst Superiority Relative to Univariate Time Series Models in Forecasting Quarterly Earnings. Journal of Accounting and Economics 9, 61–87 (1987)
6. Das, S., Lev, B.: Nonlinearities in the Returns-Earnings Relation: Tests of Alternative Specifications and Explanations. Contemporary Accounting Research 11, 353–379 (1994)
7. Duan, N., Li, K.C.: Slicing Regression: A Link-Free Regression Method. Annals of Statistics 19, 503–505 (1991)
8. Fairfield, P.M., Sweeney, R.J., Yohn, T.L.: Accounting Classification and the Predictive Content of Earnings. The Accounting Review 71(3), 337–355 (1996)
9. Horowitz, J.L.: Semiparametric Methods in Econometrics. Springer, New York (1998)
10. Kothari, S.P.: Price-Earnings Regressions in the Presence of Prices Leading Earnings. Journal of Accounting and Economics 15, 143–171 (1992)
11. Lev, B., Thiagarajan, R.: Fundamental Information Analysis. Journal of Accounting Research 31, 190–215 (1993)
12. Li, F.: Do Stock Market Investors Understand the Risk Sentiment of Corporate Annual Reports? Working paper (2006)
13. Ou, J.A.: The information Content of Nonearnings Accounting Numbers as Earnings Predictors. Journal of Accounting Research 28, 144–163 (1990)
14. Skinner, D.J.: Why Firms Voluntarily Disclose Bad News. Journal of Accounting Research 32, 38–60 (1994)
15. Wang, T.W., Rees, J.: Reading the Disclosures with New Eyes: Bridging the Gap between Information Security Disclosures and Incidents. Working paper (2007)
16. Weiss, D., Naik, P.A., Tsai, C.L.: Extracting Forward-Looking Information from Security Prices: A New Approach. The Accounting Review 83, 1101–1124 (2008)

When Generalized Voronoi Diagrams Meet GeoWeb for Emergency Management

Christopher Torpelund-Bruin and Ickjai Lee

School of Business,
James Cook University, Cairns Campus,
McGregor Rd., Smithfield, QLD4870, Australia

Abstract. This article is to investigate a Voronoi-based computational model for Geographic Knowledge Discovery (GKD) from Geo-reference Web 2.0 datasets to provide detailed emergency management analysis of various geospatial settings including various distance metrics; weights modeling different speeds, impacts, sizes, capacities of disasters; point, line and areas of influence representing disasters, obstacles blocking interactions such as political boundaries, rivers, and so on; higher order neighbors in case the first k-nearest neighbors are currently busy or not functioning; any combination of these settings. The framework is analyzed for efficiency and accuracy and tested in a variety of emergency life-cycle phases consisting of real datasets extracted from GeoWeb 2.0 technologies.

1 Introduction

An emergency is a natural or man-made situation which poses an immediate risk to health, life, property or environment. Most emergencies require urgent intervention from a range of agencies in order to prevent worsening of the situation. The four phases of emergency management are traditionally known as: *planning, response, recovery and mitigation*. Recently, Emergency Management Information Systems (EMIS) have been used by many initiatives and researchers for hazard and emergency decision support [1,2,3,4,5]. Unfortunately, many of these information systems are proprietary and developed specifically for marketing purposes to satisfy the stakeholders in the organizations. This is because currently the proliferation of information concerning all aspects of emergencies is extremely expensive. For an organization to have this kind of detailed information requires major investments into the ways of gathering often massive amounts of distributed data from multiple sources. On top of this, many of the information systems are limited to producing cartographic mappings and visualization for response and recovery rather than emergency analysis and predictive modeling [6]. Effective response and recovery can only be possible with rigorous mitigation and preparedness planning and prediction modeling. One of the best ways that emergencies can be planned for is with the use of a Knowledge Discovery (KD) process which aims to find previously unknown information by comparing every aspect of the known information. The latest research has

H. Chen et al. (Eds.): PAISI 2009, LNCS 5477, pp. 64–75, 2009.
© Springer-Verlag Berlin Heidelberg 2009

produced a Voronoi based model that supports emergency analysis by utilizing higher order Voronoi diagrams for what-if emergency management [7].

Currently, no EMIS exists that can provide detailed analysis and decision support on a global scale. This is due to lack of a proper infrastructure whereby contributed global information can be aggregated and fully utilized. Traditional EMIS are domain-specific and limited to response and recovery phases focused on a particular region and a specific target. They fail to provide decision support for the four phases simultaneously. They typically suffer from lack of effective interfaces and data visualizations for discovered knowledge analysis.

This paper investigates Geographic Knowledge Discovery (GKD) from GeoWeb model combined with spatial data mining with Generalized Voronoi Diagrams (GVD) for use in an EMIS that overcomes the drawbacks of traditional EMIS, and allows for effective dynamic decision support and optimization. The advantages of using GKD from GeoWeb Model include: 1) the extraction of implicit, previously unknown, and potentially useful information from data; 2) efficient and relatively cheap access to enormous amounts of participation driven information; 3) geographic referencing allows geospatial analysis; 4) constantly dynamically updating information as well as temporal-referencing allow the time dimensions to be added to the GKD process for further analysis. The advantages of using GVD include: 1) succinctly models topology and geometry together; 2) can provide geographic analysis from easily obtainable raster mappings; 3) is a mathematically robust data modeling technique; 4) supports various generalizations for effective modeling of many various real world situations.

The aims of the project included the development of: 1) the clear definition of stages to allow knowledge discovery from information extracted from the Internet; 2) a list of well defined Web 2.0 protocols to use during the data selection and pre-processing stages; 3) data mining techniques that include spatial and temporal dimensions; 4) spatial analysis techniques with the use of GVD; 5) methods for interpretation and visualization of discovered information that can lead to knowledge discovery. Based on the functionality provided by our model, the following three types of essential queries for diverse emergency decision making are supported. Its capability encompasses: 1) TYPE I: neighboring (adjacency) support; 2) TYPE II: districting (segmenting) support; 3) TYPE III: location optimization support.

2 Preliminaries

2.1 Geographic Knowledge Discovery

GKD is data mining applied to geospatial datasets. GKD emphasizes the importance and uniqueness of geographical space. GKD must consider the peculiar characteristics of geoinformation (dependence and heterogeneity) that makes space special in order to detect geographically interesting patterns that are either descriptive or predictive [8]. GDM techniques can be classified according to tasks and methods as follows:

- Geospatial clustering: identifying a set of spatial aggregations with similar characteristics that summarizes and describes the data (descriptive pattern).
- Geospatial characterization: generalizing the data to find compact descriptions (descriptive pattern).
- Geospatial deviation detection: finding outliers that deviate from spatial aggregations (descriptive pattern).
- Geospatial classification: assigning data to one of predefined classes (predictive pattern).
- Geospatial trend detection: detecting changes and trends along a spatial dimension (predictive pattern).
- Geospatial association: finding interesting dependencies among attributes (predictive pattern).

The first three are descriptive pattern mining tools that summarize and characterize the data while the last three are predictive pattern mining tools that draw inferences from the data for predictions.

GKD as a quantitative study of geo-referenced phenomena, explicitly focuses on the spatial extent including location, distance, interaction and spatial arrangement. Thus, spatial-dominant generalization, where identifying spatial clusters is a fundamental and core operation, is more important than aspatial-dominant generalization. In spatial-dominant generalization, clustering works as a summarization or generalization operation on the spatial component of the data. Later, an agent (computer or human) can explore associations between layers (themes or coverages). In fact, other layers are explored to attempt to find associations that may explain the concentrations found by the clustering algorithm.

2.2 Generalized Voronoi Diagrams

Let $P = \{p_1, p_2, ..., p_k\}$ be a set of *generator* points of interest in the plane in R^m space. For any point p in the plane, $dist(p, p_i)$ denotes the distance from point p to a generator point p_i. The distance metric can be of Manhattan, Chessboard, Euclidean or another metric and the resulting dominance region of p_i over p_j can be defined as:

$$dom(p_i, p_j) = \{p \mid dist(p, p_i) \leq dist(p, p_j)\}\,. \tag{1}$$

For the generator point p_i, the Voronoi region of p_i can be defined by:

$$V(p_i) = \cap_{j \neq i} dom(p_i, p_j). \tag{2}$$

The partitioning into subsequent Voronoi regions $V(p_1)$, $V(p_2)$, $\cdots$, $V(p_k)$ is called the *generalized Voronoi diagram*. The bounding regions of $V(p_i)$ are known Voronoi boundaries and depending on the primitive used as the generator, such as points, lines or polygons and the metric space used, will result in a series of polygons and arcs made up of lines and Bezier curves. These geospatial dominance regions provide natural neighbor relations that are crucial for many topological queries in geospatial modeling and analysis.

The way generalized Voronoi diagrams can be represented is virtually limitless. Voronoi diagrams are based on the distance metrics used, the types of primitives used and how generators interact in the plane. Common metrics used include Euclidean, Manhattan and Chessboard distance. While Euclidean distance offers high precision, Manhattan distance metrics are known to better approximate real world situations [9]. While our system generates many generalizations of Voronoi diagrams for various geospatial analysis, three fundamental models are the Multiplicatively Weighted Voronoi Diagram (MWVD), Higher Order Voronoi Diagrams (HOVD) and the Voronoi Diagram with Obstacles (VDO).

In many situations, each point has different influential factors such as different sizes, powers, speeds and weights. Using a variety of functional forms for calculating the weighted distance, many weighted Voronoi diagrams are possible. The dominance regions generated by MWVDs are defined by the following function:

$$MWVD : d_{mw}(p, p_i) = \frac{1}{w_i} |x - x_i| \qquad (w_i > 0). \tag{3}$$

The real world is obviously non-trivial and it is the goal of GIS to represent real world objects and obstacles that affect them as accurately as possible. Spatial analysis with the presence of obstacles is a natural candidate for representing dangerous areas and non-passable terrain (mountains, rivers, oceans etc). Let $O = \{o_1, ..., o_n\}$ represent a set of line segments and polygons that are obstacles in the plane. The shortest path distance from point p to element e_i that does not intersect any obstacles is commonly referred to as the *geodesic distance* in computational geometry literature [10]. The shortest path distance from p to p_i with obstacle set O is found with the aid of a *visibility graph*, denoted by $VG(Q, L_{vis})$ where $Q = q_1, ..., q_n, p, e_i$ represent the vertices that define the obstacle polygon, and p and e_i represent the point p and element e_i where the distance value is being calculated. L_{vis} denotes the set of line segments $\overleftrightarrow{q_i q_j}$ where q_j is visible from q_i. The visibility graph contains all visible links from p to e_i along L_{vis}. The shortest path is found by computing the shortest combination of links in $VG(Q, L_{vis})$ from p to e_i. Using the shortest path from p to e_i the *shortest-path Voronoi region associated with O* can be constructed, defined as:

$$\mathcal{V}_{sp}(p_i) = \{p | d_{sp}(p, e_i) \leq d_{sp}(p, e_j)\} . \tag{4}$$

Emergency response is often concerned with trying to find the nearest k units available or when the first nearest unit is in use. This modelling can be generated with what is known as *higher-rrder Voronoi diagrams*. This algorithm can be defined as:

$$V(P_i^k) = p | max_{P_h} \left\{ d(p, p_h) | p_h \in P_i^k \right\} \leq min_{P_i} \left\{ d(p, p_j) | p_j \in P / P_i^k \right\} . \tag{5}$$

The *Order-k Voronoi Diagram* ($\mathcal{V}^{(k)}$) is a set of all order-k Voronoi regions $\mathcal{V}^{(k)} = \left\{ V(P_1^{(k)}), ..., V(P_n^{(k)}) \right\}$, where the order-k Voronoi region $V(P_i^{(k)})$ for a random subset $P_i^{(k)}$ consists of k points out of P.

3 The Geographic Knowledge Discovery from GeoWeb Model

The GKD from GeoWeb process models the steps of KD process except with the extension that data has included spatial and temporal dimensions. GKD from GeoWeb reflects a growing awareness of the need to interpret geospatial information in the broadest possible terms to provide a platform where information sharing and collaboration is utilized to increased productivity, efficiency and vastly improved decision making. Information is extracted from a set of well defined Web 2.0 protocols and undergoes cleaning, normalization and reduction in order to create a refined information space for the data mining layer.The discovered patterns from the data mining layer are interpreted and visualized using previous known information retrieved from the Internet in order to create

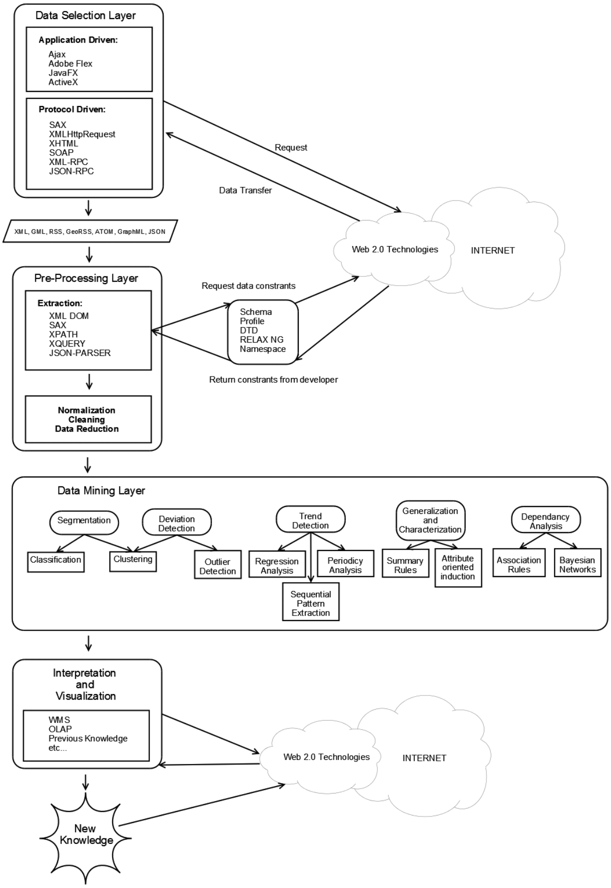


Fig. 1. The GKD from GeoWeb process

new hypotheses which can eventually lead to new knowledge. The GKD from GeoWeb process is illustrated in Fig. 1.

4 GeoWeb Based Emergency Management

4.1 Framework

This section details the functionality the GKD from GeoWeb framework utilizing the generalized Voronoi diagram model for an EMIS giving detailed spatial analysis information that can be used for accurate decision support. The case studies follow the framework described in Fig. 2. In the following cases, Yahoo Local [1] and the Global Disaster Alert and Coordination System [2] (GDACS) are used for retrieving the geo-referenced information. k-medoids clustering is combined with the GVD framework for TYPE I: neighbouring, TYPE II: districting and TYPE III: location optimization in the data mining layer to identify spatial patterns. The extracted new knowledge is represented on Web Map Servers (WMS) which include Open Street Maps[3], Yahoo Maps[4], Google Maps[5] and NASA Blue Marble[6] for detailed spatial analysis and new knowledge elicitation. The information displayed for the case studies is real-world information and results represent real spatial analysis.

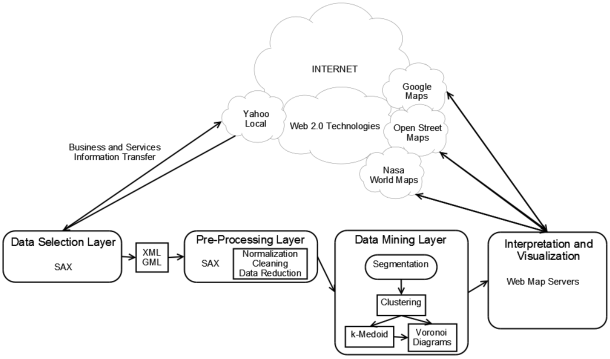


Fig. 2. The process used for the EMIS utilizing Web 2.0 technologies and spatial analysis

[1] http://developer.yahoo.com/local/

[2] http://www.gdacs.org/

[3] http://www.openstreetmap.org/

[4] http://maps.yahoo.com/

[5] http://maps.google.com/

[6] http://earthobservatory.nasa.gov/Features/BlueMarble/

4.2 Raster Generalized Voronoi Diagrams

The development of our efficient and effective model required the incremental implementation of algorithms that allow the generation of all the generalized Voronoi diagrams. Voronoi diagram generalizations tend to exhibit a hierarchy of connectivity with the root being the Ordinary Voronoi Diagram (OVD). The subsequent subclasses of the OVD modify its general properties to produce new VD. We developed the generalized algorithms based on this hierarchy. Before moving onto the next generalization required the previous to be implemented and working correctly. The incremental development of our project is shown in Fig. 3.

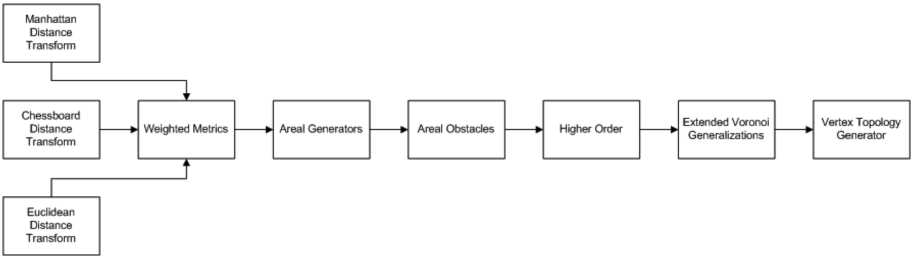

Manhattan Distance Transform
Chessboard Distance Transform
Euclidean Distance Transform
Weighted Metrics
Areal Generators
Areal Obstacles
Higher Order
Extended Voronoi Generalizations
Vertex Topology Generator

Fig. 3. The incremental stages of development

The sequential scan methods for the Manhattan, Chessboard and Euclidean OVDs serve as a platform for the development of the other generalizations. We use a modified version of Shin and Wu [11]'s sequential scan exact Euclidean distance transformation. Unlike Shin and Wu's algorithm which focuses on finding the relative distance from p, our algorithm considers the direction of connection needed for the dominance regions generated by complex primitives. Because all generalizations support the complex primitives (point, line and polygon). Details of our algorithm are beyond of the scope of this paper and interested readers may refer to [12].

5 Case Studies

5.1 San Francisco Police Optimization

The main San Francisco police department has been assigned the task of reducing criminal activities which has been on the rise in the southern districts of the city. The local council has agreed to the funding of the development of new police departments. However, due to budget constraints only a small number of new departments will be granted. It is clear that optimal locations for police departments must be determined that will provide maximum effective coverage. To do this, the San Francisco police employs the TYPE III: location optimization query (*Mitigation and Preparedness phases*).

The first stage is to locate the trouble areas in the south district of the city. A recent study has shown that the large number of prisons in the area might be contributing to the rise of criminal activities. To test this theory, the GKD from GeoWeb framework is used to cluster prison areas within San Francisco. The results are shown in Fig. 4(a), and indicate that indeed there exists a high number of prisons in the southern area compared to the rest of the city. The prisons area is focused on and combined with school locations to determine the levels of residency within the area, shown in Fig. 4(b). The close proximity of prisons to neighborhoods suggests that people could be in danger if police coverage is not adequate. To determine the police coverage TYPE I query can be used. Figure 4(c) shows the nearest police department for areas in south San Francisco in the Manhattan metric. Higher-order Voronoi diagrams can be used to determine the coverage when multiple departments may be needed. Figure 4(d) shows the 3rd-order Voronoi diagram with the Manhattan metric. The results show that police coverage is quite inadequate for the south-eastern districts, especially with the association of high crime in areas around prisons. This area is chosen to be the target of the project.

The next stage is to determine the optimal placement of police departments in order to be in equi-distance to each of the prisons. However, not all of the prisons have the same significance. Clearly the group of prisons should be given higher priority and have police departments in a closer proximity for more effective coverage. To determine this optimal placement, the Voronoi diagram computa-

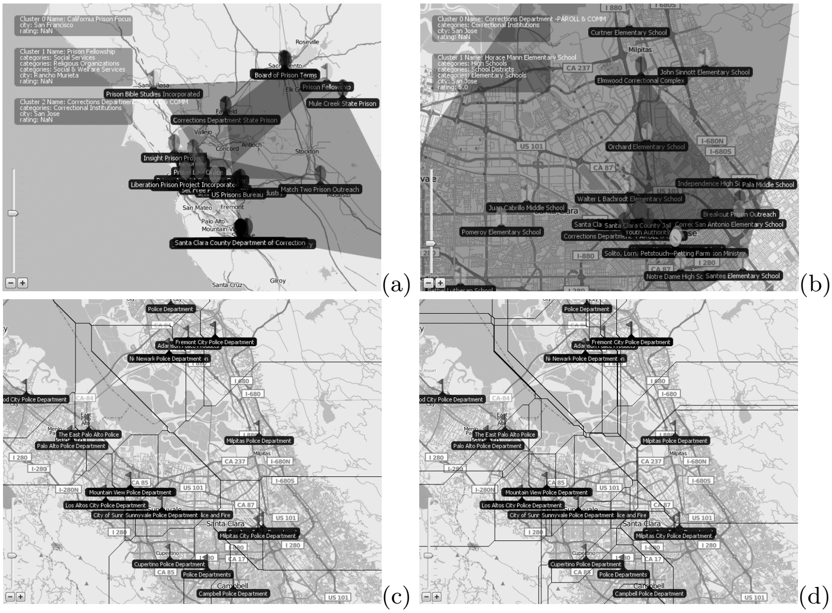

(a)
Cluster 0 Name: California Prison Focus
city: San Francisco
rating: NaN
Cluster 1 Name: Prison Fellowship
categories: Social Services
categories: Religious Organizations
categories: Social & Welfare Services
city: Rancho Murieta
rating: NaN
Cluster 2 Name: Corrections Department -PAROLL & COMM
categories: Correctional Institutions
city: San Jose
rating: NaN
Prison Bible Studies Incorporated
Board of Prison Terms
Prison Fellowship
Mule Creek State Prison
Corrections Department State Prison
Insight Prison Project
Liberation Prison Project Incorporated
US Prisons Bureau
Match Two Prison Outreach
Santa Clara County Department of Correction
Roseville
Sacramento
Fairfield
Vallejo
Concord
Antioch
Stockton
San Mateo
Fremont
Palo Alto
Mountain View
Santa Cruz
Gilroy
(b)
Cluster 0 Name: Corrections Department -PAROLL & COMM
categories: Correctional Institutions
city: San Jose
rating: NaN
Cluster 1 Name: Horace Mann Elementary School
categories: High Schools
categories: School Districts
categories: Elementary Schools
city: San Jose
rating: 5.0
Curtner Elementary School
Milpitas
John Sinnott Elementary School
Elmwood Correctional Complex
Orchard Elementary School
Independence High School
Pala Middle School
Juan Cabrillo Middle School
Walter L Bachrodt Elementary School
Breakout Prison Outreach
San Antonio Elementary School
Pomeroy Elementary School
Santa Clara County Jail
Youth Authority
Corrections Department
Petting Farm
Notre Dame High School
Santee Elementary School
US 101
I-880
I-680N
I-680S
CA 237
CA 87
I 280
(c)
(d)
Police Department
Fremont City Police Department
Newark Police Department
Milpitas Police Department
The East Palo Alto Police
Palo Alto Police Department
Mountain View Police Department
Los Altos City Police Department
City of Sunnyvale Police Department
Milpitas City Police Department
Cupertino Police Department
Police Departments
Campbell Police Department
Santa Clara
Cupertino
Campbell
I 280
CA 85
US 101
I-680
CA 237
CA 87
CA 17
I-880

Fig. 4. San Francisco prison datasets and police coverage

tion model can be used to generate a high-priority low-weighted support vector topology. It is determined that the group of prisons in the south should have 1st priority, with the remaining two having 2nd priority. There should be four support units available in order to deal with the 1st priority area. The 2nd priority prison will only require two support units. Because the chances of criminal activity in the areas occurring simultaneously are low, it is determined that the support vectors should be shared across each of the prison areas. The optimal location determined is shown in Fig. 5(a).

This information is presented to the council, who agree that this is the correct course of action. However, the council also assumes the problem of criminal activity in the area is due to the over-crowding within the prisons. In order to facilitate effective rehabilitation of criminals, it is determined a new facility within the area will be used to share the load of criminals. The new prison will be used to house very naughty kids from the high-security prisons from the south. The naughty kids prison is considered as less of a risk than the other prisons and will therefore be given 3rd priority. The addition of a new prison in the area requires more intensive coverage from the police departments. The 1st priority prisons will continue to have a minimum of four support units. However, the 2nd priority prisons will now be assigned three support units. The naughty kids prison will only require one support unit. The optimal arrangement of police departments is shown in Fig. 5(b).

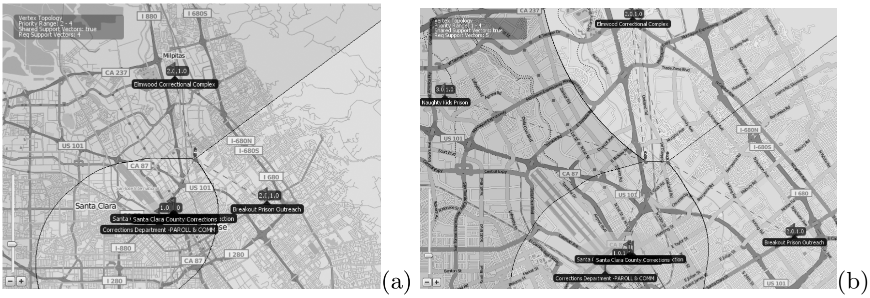


Fig. 5. San Francisco optimal placement of police departments: (a) The initial optimized police department location; (b) The set of optimized locations for police departments taking into account the new prison

5.2 Manhattan Fire Emergency

In this scenario, Manhattan island has just been hit by a massive tornado which has caused dangerous fires within the three major areas of Manhattan. Emergency services have been assigned the task of effectively coordinating the situation (response & recovery phase). The GKD from GeoWeb framework and Voronoi diagram computation model can be used to provide real-time decision support by utilizing all three of the query types. The first step is to locate the

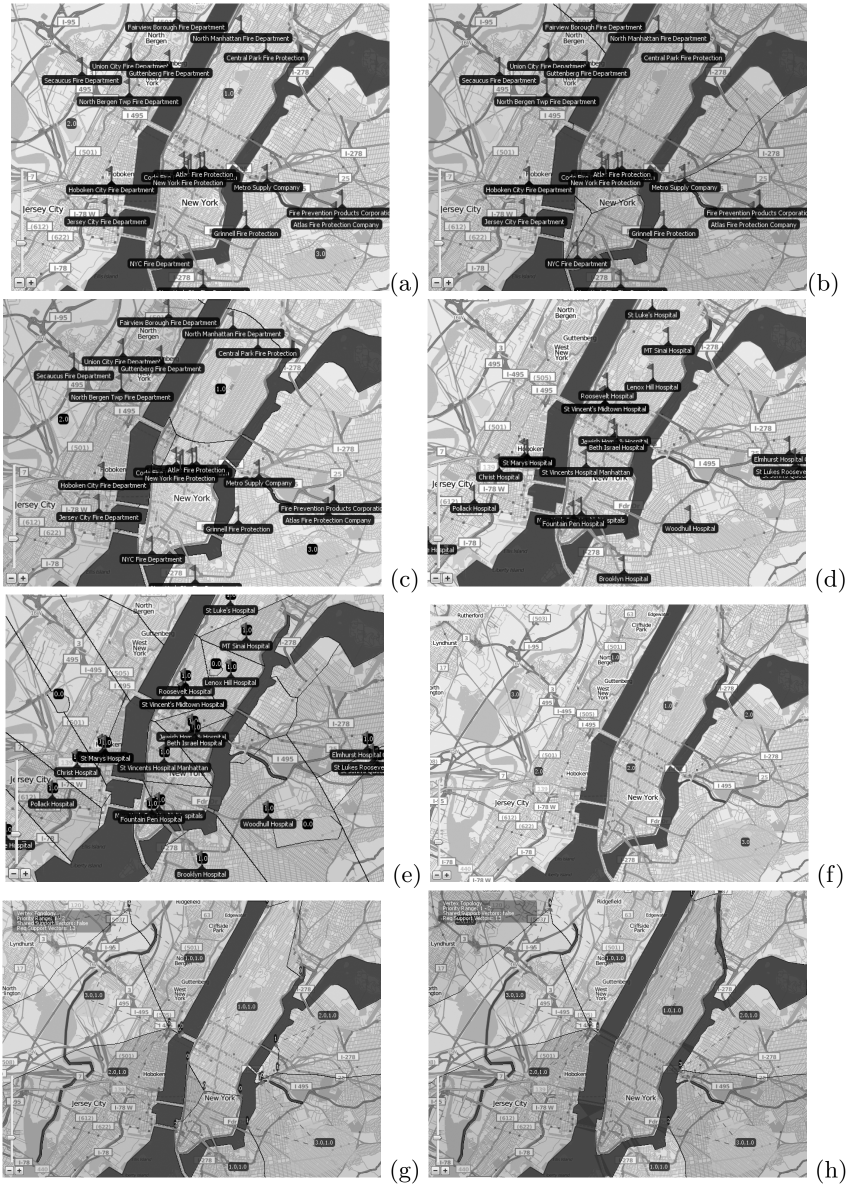


Fig. 6. Manhattan fire emergency: (a) The location of fires and fire departments in Manhattan; (b) The initial districting query with fires and fire departments; (c) The districting query with fires taking into account significance values; (d) The location of hospitals within Manhattan; (e) Hospitals represented with the neighborhood query; (f) The new additional set of fires; (g) The initial optimal set of evacuation points; (h) The optimal set of evacuation points avoiding Manhattan island

fires and local fire departments within the area, which is shown in Fig. 6(a). The districting query with the fires as targets, shown in Fig. 6(b), locates the nearest sets of fire departments that can provide support. However, some of the locations of the fires suggest that some fires require higher priority/concentrated support than others. The fire in central park is given the lowest priority, while the industrial area on the left gets middle priority. The highest priority goes to the fire within the residential area located in the south-east corner. Using these new weighted priorities, the new districting query gives the results shown in Fig. 6(c). The new results indicate that most of the fire departments located in Manhattan should be used on the residential area. While the fire department coordination effort is in progress, the hospitals also needed to be coordinated as to how best to provide support for the victims of the fires. The hospitals located within Manhattan are described in Fig. 6(d). Using the neighborhood query, shown in Fig. 6(e), the hospitals neighboring the fires are easily located and dispatched to provide assistance.

Meanwhile, the fire department coordination effort has taken a turn for the worse with extreme damage in the area causing even more widespread fires. The new fire location information is shown in Fig. 6(f). The current events are now defining a disaster and the authorities decide it is in the best interests of the civilian population that they be evacuated to areas deemed safe away from the fires while the emergency services continue to battle the fires. The safest evacuation areas can be determined by using a high-priority high-weighted support vector topology. It is determined that for each of the first and second priority fires (weight 2,3), two dedicated hospital units should be available to aid victims of the fires. For the third priority fires (weight 1), one hospital unit should be dedicated. The initial plan for the optimal set of evacuation sites is shown in Fig. 6(g). However, it is decided that Manhattan island should be avoided where possible to prevent evacuating citizens blocking bridges being used by the emergency services. Taking this into account, the new set of optimal evacuation sites is shown in Fig. 6(h).

6 Final Remarks

In this article, we proposed a Voronoi diagram based computational model for data mining in the GKD from GeoWeb model. The spatial analysis provided by generalized Voronoi diagrams can be effectively used in an EMIS for decision support and optimizations. In reality there could be an infinite number of disciplines that this framework could be applied to because of its versatility. The GKD from GeoWeb framework allows an enormous amount of data to be extracted and analyzed based on all forms of geography (human geography, environmental geography, business geography etc). Web 2.0 technologies are constantly updating, meaning discovered patterns based on the dynamic data represents real-time knowledge. The use of Yahoo Local and GDACS for mining spatial patterns is only a simple example of what could eventually be possible. Another example using the GKD model could be to analyze numerous blogs and social networks of finance investors at certain geographic locations in order to determine even more detailed market trends. If people are worried about specific sectors then that could imply a bull or bear market. Another possible use of the framework is for human geography analysis. Hypothesis could be made about human development by using the framework and applying classification and prediction techniques on the Web 2.0 information. The possibilities are literally endless.

References

1. Chang, N.B., Wei, Y.L., Tseng, C.C., Kao, C.Y.J.: The Design of a GIS-based Decision Support System for Chemical Emergency Preparedness and Response in an Urban Environment. Computers, Environment and Urban Systems 21(1), 67–94 (1997)
2. Dymon, U.J., Winter, N.L.: Evacuation Mapping: The Utility of Guidelines. Disasters 17(1), 12–24 (1993)
3. Kevany, M.J.: GIS in the World Trade Center Attack - Trial by Fire. Computers, Environment and Urban Systems 27(6), 571–583 (2003)
4. Montoya, L.: Geo-data Acquisition through Mobile GIS and Digital Video: An Urban Disaster Management Perspective. Environmental Modelling & Software 18(10), 869–876 (2003)
5. Salt, C.A., Dunsmore, M.C.: Development of a Spatial Decision Support System for Post-Emergency Management of Radioactively Contaminated Land. Journal of Environmental Management 58(3), 169–178 (2000)
6. Zerger, A., Smith, D.I.: Impediments to Using GIS for Real-time Disaster Decision Support. Computers, Environment and Urban Systems 27(2), 123–141 (2003)
7. Lee, I., Pershouse, R., Phillips, P., Christensen, C.: What-if Emergency Management System: A Generalized Voronoi Diagram Approach. In: Yang, C.C., Zeng, D., Chau, M., Chang, K., Yang, Q., Cheng, X., Wang, J., Wang, F.-Y., Chen, H. (eds.) PAISI 2007. LNCS, vol. 4430, pp. 58–69. Springer, Heidelberg (2007)
8. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, San Francisco (2000)
9. Krause, E.F.: Taxicab Geometry. Addison-Wesley, California (1975)
10. Okabe, A., Boots, B.N., Sugihara, K., Chiu, S.N.: Spatial Tessellations: Concepts and Applications of Voronoi Diagrams, 2nd edn. John Wiley & Sons, West Sussex (2000)
11. Shih, F.Y., Wu, Y.T.: Fast Euclidean Distance Transformation in Two Scans Using a 3 × 3 Neighborhood. Computer Vision and Image Understanding 93(2), 195–205 (2004)
12. Torpelund-Bruin, C.: An Efficient Generalized Voronoi Diagram Based Computational Model for Geographic Knowledge Discovery for Use in Emergency Management Support Systems. Honours's thesis, James Cook University, Townsville, Australia (2008)

E^3TP: A Novel Trajectory Prediction Algorithm in Moving Objects Databases*

Teng Long[1], Shaojie Qiao[2,1,3,**], Changjie Tang[1], Liangxu Liu[4,3],
Taiyong Li[1, 5], and Jiang Wu[5]

[1] School of Computer Science, Sichuan University, Chengdu 610065, China
[2] School of Information Science and Technology, Southwest JiaoTong University,
Chengdu 610031, China
qiaoshaojie@gmail.com
[3] School of Computing, National University of Singapore, Singapore, 117590, Singapore
[4] School of Electronic and Information Engineering, Ningbo University of Technology,
Ningbo, 315016, China
[5] School of Economic Information Engineering, Southwest University of Finance and
Economics, Chengdu 610074, China

Abstract. Prediction of uncertain trajectories in moving objects databases has recently become a new paradigm for tracking wireless and mobile devices in an accurate and efficient manner, and is critical in law enforcement applications such as criminal tracking analysis. However, existing approaches for prediction in spatio-temporal databases focus on either mining frequent sequential patterns at a certain geographical position, or constructing kinematical models to approximate real-world routes. The former overlooks the fact that movement patterns of objects are most likely to be local, and constrained in some certain region, while the later fails to take into consideration some important factors, e.g., population distribution, and the structure of traffic networks. To cope with those problems, we propose a general trajectory prediction algorithm called E^3TP (an **E**ffective, **E**fficient, and **E**asy **T**rajectory **P**rediction algorithm), which contains four main phases: (*i*) mining "hotspot" regions from moving objects databases; (*ii*) discovering frequent sequential routes in hotspot areas; (*iii*) computing the speed of a variety of moving objects; and (*iv*) predicting the dynamic motion behaviors of objects. Experimental results demonstrate that E^3TP is an efficient and effective algorithm for trajectory prediction, and the prediction accuracy is about 30% higher than the naive approach. In addition, it is easy-to-use in real-world scenarios.

Keywords: trajectory prediction, moving objects databases, criminal tracking analysis, hotspot regions, frequent sequential routes.

* This work is supported by the National Natural Science Foundation of China under Grant No. 60773169, the 11th Five Years Key Programs for Science and Technology Development of China under Grant No. 2006BAI05A01, the Youth Software Innovation Project of Sichuan Province under Grant No. 2007AA0032 and 2007AA0028.

** Corresponding author.

H. Chen et al. (Eds.): PAISI 2009, LNCS 5477, pp. 76–88, 2009.
© Springer-Verlag Berlin Heidelberg 2009

1 Introduction

With the rapid development in the wireless and mobile techniques, law enforcement agencies are provided with a large volume of trajectory data in terms of “moving objects” in a crime database [1], such as the movement of vehicles and criminals. From these data, we can easily obtain the instant information of these objects, i.e., the current location and moving direction. Such information can be very helpful for law enforcement agencies in various applications such as criminal location analysis and border safety control. In general, there exist no apparent rules of moving objects that are stored and never used for research purpose. However, these data contain a lot of useful movement behaviors [2]. So, how to discover useful information from moving objects databases relevant to crime cases and use them to predict trajectories of fleeing criminals has become a hot research topic.

It is important to accurately predict the current position of a moving object at a given time instant. For instance, it can help law enforcement agencies accurately track criminals who have committed burglaries during a particular time interval. There exist several characteristics in terms of trajectory data. Firstly, the trajectories are often constrained by streets and highways, so it is impossible for them to move dynamically like hurricanes [3]. Secondly, the speed of objects is mainly determined by the current road condition and other natural factors (e.g., the weather). Thirdly, there are a variety of “frequent routes” which are used more frequently than others.

Recently, many manufactories equip their mobile products with GPS devices, which increases the demand of developing new trajectory prediction technologies [2]. Particularly, law enforcement agents can use such technologies to trace criminals. In addition, it can be applied to handling traffic planning problems. These requirements motivate us to develop an effective, efficient and easy trajectory prediction approach.

However, it is difficult to accurately and efficiently predict trajectories of moving objects due to the following reasons. Firstly, the position information is periodically delivered to the central databases, and the periodicity of position acknowledgement is apt to be affected by several factors, e.g., signal congestions. Whenever the position of an object is unknown, an effective trajectory prediction approach is necessarily required [2]. Secondly, the data of moving objects cannot precisely depict its real location due to continuous motions or network delays [4].

The original contributions of this paper include:

- We proposed to discover hotspot regions where frequent trajectory patterns clustered by processing historical trajectory data of moving objects.
- We proposed an FP-growth [5] based approach to discover frequent routes of moving objects within a certain range.
- We performed extensive experiments to demonstrate the effectiveness and efficiency of our purposed algorithm by comparing it with the naive approach.

2 Problem Statement

The 3D cylindrical model [4, 6] of trajectories is defined beyond temporally annotated sequences (*TAS*) [7, 8] which extended the sequential patterns by adding temporal information of locations. Formally, the definition is presented as follows.

Definition 1 (*Trajectory*). A trajectory of objects is composed of a series of triples:

$$S=\{(x_1, y_1, t_1), \ldots, (x_i, y_i, t_i), \ldots, (x_n, y_n, t_n)\} \tag{1}$$

where t_i represents time-stamps, $\forall i \in [1, n-1]$, $t_i < t_{i+1}$. (x_i, y_i) represents 2D coordinates.

However, it is difficult to accurately locate moving objects in real-life scenarios. So we use a disk area to replace the $\{xy\}$-coordinate in Equation 1. In Fig. 1, T_1 and T_2 are treated as the same trajectory if the following equation holds.

$$(x_i - x_i')^2 + (y_i - y_i')^2 \leq r^2 \tag{2}$$

where $i \in \{1,2,\ldots, n\}$, and r is the radius of the disk.

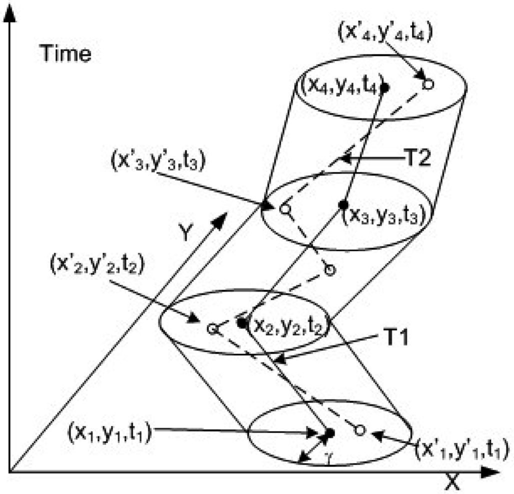


Fig. 1. The 3D cylindrical model of trajectories

When a trajectory is given, we can perform trajectory prediction. Generally, there are two kinds of information contained in a trajectory, i.e., spatial information and temporal information. The former is displayed by coordinates or disks, and the later is related to time-stamps. Our prediction algorithm takes into account these two aspects.

As mentioned in Section 1, the movement of objects is embraced by streets or highways, in other words, it is restricted within a map [9]. Once the geographical information of items (i.e, streets, and buildings) in a map is given, the position of a moving object can be predicted by exploring the streets and highways that can be visited in the future. Here, we firstly give two definitions.

Definition 2 (*Map*). A map is composed of streets, and each street is depicted by the following attributes:

- ID: the identifier number of a street;
- Str: a polygonal line between two ending points, which consists of several line-segments.

Figure 2 is an example of a map, where the number beside each street is the street ID, and each street is denoted as the path between two capital letters.

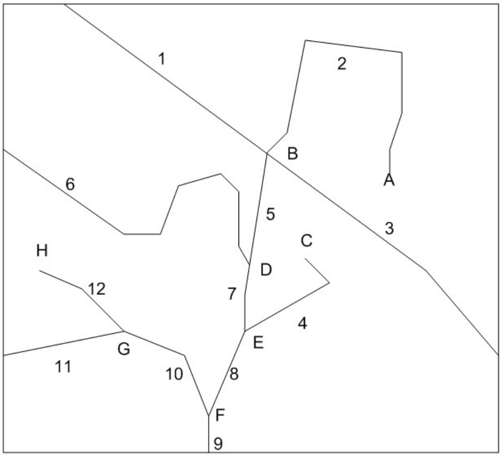


Fig. 2. Example of a map

Definition 3 (*Route*). A route is a sequence of streets:

$$R = \{s_1, s_2, \ldots, s_i, \ldots, s_n\} \tag{3}$$

where $i \in [1, n]$, and s_i has at least one common ending point with s_{i+1}.

For example, in Fig. 2, a route can be indicated as {11, 10, 8, 7, 5}, which means some object can move from street 11 to 5 by way of street 10, 8, 7.

The goal of trajectory prediction is to obtain trajectories as defined in Equation 1, so we should find the possible routes of objects and extract its temporal information as well. When a moving object arrives at a crossing road, it needs to choose which street to go next. Basically, the decision is made on where the object comes from, and where it is going. In this model, there exist some hidden rules in this map, i.e., traffic rules, population distributions. These rules can be mined from historical traffic data.

As for timestamps, they can be calculated as we know the approximate speed of moving objects. The speed is determined by several factors that can be divided in to two categories: external factors (i.e., weather, road condition) and internal factors. We should consider both categories of factors when designing prediction methods.

In addition, we should take into account another question as "*do the previous mentioned rules exist everywhere in a map*?" It is straight-forward that regions with heavy traffic most likely contain internal rules. So, in our algorithm, we firstly find hotspot areas in a map, and perform predicting in those hotspot areas.

In general, there are four phases to predict trajectories of moving objects:

1) Mining hotspot regions in a map from moving objects databases;
2) Discovering frequent routes from hotspot areas;
3) Computing the speed of a variety of moving objects in distinct streets;
4) Predicting the dynamic motion behaviors of moving objects.

3 Mining Hotspot Regions

Mining hotspot regions is an essential step for trajectory prediction, because the rules of choosing routes exist in those regions rather than the ones that are rarely visited. Historical traffic data in terms of these regions are processed in order to find rules. In addition, frequent routes are discovered in those regions. In summary, moving objects in hotspot areas are more predictable than the ones outside them.

Hotspot region mining is treated as a preprocessing phase in our approach. By calculating the visiting frequency of each street, we can find such streets that are frequently visited, i.e., they have higher visiting frequencies than a given threshold. Those connected streets are grouped into a hotspot region. Then, all historical data are scanned again, and segments of trajectories in those hotspot regions are stored.

The details of mining hotspot regions from trajectory data are shown in Algorithm 1. In Algorithm 1, we firstly calculate the visiting frequency of each street in E, and sort them in descending order (lines 1-3). Then, we compute the specified frequency threshold f (line 4), and delete such streets whose frequency is lower than f (lines 5-7). Next, the group of remaining streets forms a hotspot region (lines 9-12). After that, we iteratively scan each hotspot region, if two hotspot regions are connected, combine them into one region until no more hotspot regions can be amalgamated (lines 13-17). Finally, output all hotspot regions (line 18).

In Algorithm 1, p is the percentage of frequently visited streets, the Sort(·) function is used to sort all streets by frequency in descending order. The work mechanism of lines 9-18 is grouping all of the connected high-frequent streets.

Algorithm 1. Mining hotspot regions

Input: a trajectory data set $\boldsymbol{D}$, a set of streets $\boldsymbol{E}$, a threshold $\boldsymbol{p}$
Output: a set of hotspot region $O = \{O_1, \ldots, O_{num}\}$

```
1.  for each street $e \in E$ do
2.      $e.freq \leftarrow$ ComputeFreq($e$);
3.  Sort($E$);
4.  $f \leftarrow E.size * p$;
5.  for each street $e \in E$ do
6.      if $e.freq < f$ then
7.          delete $e$;
8.  $ID \leftarrow 0$;
9.  for each street $e \in E$ do
10.     create a new hotspot region $O_{ID;}$
11.     $O$.add($O_{ID}$);
12.     $ID$++;
13. while ($O.size$ changes)
14.     for each $O_i$, $O_j$ in $O$ do
15.         if ($O_i$ and $O_j$ are connected) then
16.             $O_i$.add($O_j$);
17.             delete($O_j$);
18. output $O$;
```

After constructing the hotspot regions, we process historical trajectories by clearing out trajectory segments that are outside hotspot regions, while storing the segments in hotspot regions. In the following sections, we will introduce the approach of discovering moving rules from hotspot regions.

4 Mining Frequent Routes Based on FP-Tree

The main idea of our prediction approach is to discover moving rules of moving objects in those hotspot regions and use these rules to predict possible trajectories.

The FP-growth algorithm [5] is an efficient and scalable frequent itemset mining method. When treating each street in a hotspot region as an item, these items can be organized as a sequence by connecting their ending points. So, the problem of discovering frequent routes is equivalent to mining frequent trajectory patterns that are composed of streets in hotspot regions.

We modify FP-growth approach to suit the trajectory prediction problem, and we name the new method FR_mining (Frequent Route Mining), which includes FP-tree construction and mining phases. The FP-tree construction step is analogous to that applied in FP-growth, and is given in Algorithm 2.

Algorithm 2. FR_mining

Function 1: FP-tree construction

Input: ***R***, a hotspot region contains a set of streets and a set of trajectory segments; ***min_sup***, the minimum support count

```
for each street $s_i$ in $R$ do
  count $s_i.freq$;
  if ($s_i.freq<min\_sup$) then
    delete $s_i$;
sort $R$ by $freq$ in descending order as $L$;
create a root node $t$ of the FP-tree and label it as "null";
for each trajectory segment $S$ in $R$ do
    select and sort streets in $S$ and put them into $S'$ by the order of nodes in $L$, $S'=[s|S^*]$, where $s$ is the first street and $S^*$ is the remaining nodes in $L$;
    insert_tree([$s|S^*$], $t$);
output $t$, $L$;
```

Procedure insert_tree([$s|S^*$], t)

```
if ($t$ has a child $c$ and $c.sid=s.sid$)
  $c.freq \leftarrow c.freq+1$;
else
  create a new node $n$;
  $n.freq \leftarrow 1$;
  $n.parent \leftarrow t$;
  find $n'$ in $L$, where $n'.sid=n.sid$;
  if $n'.next=null$ then
    $n'.next \leftarrow n$;
  else
    do
      $n' \leftarrow n'.next$;
    while ($n'.next \neq null$)
    $n'.next \leftarrow n$;
if ($S^*$ is not empty) then
  insert_tree([$S^*$], $n$);
```

Function 2: Mining frequent routes from FP-tree

Input: ***t***, the root node of an FP-tree; ***L***, a list of streets in a hotspot region corresponding to *t*.

Output: a set of frequent routes in a hotspot region *R*.

Method:

```
Create a set of frequent routes $R$;
```

Procedure FP_mining(t, α)

```
if ($t$ contains a single path $P$) then
    generate a pattern $p=P\cup\alpha$ with $support\_count$ =the minimum $freq$ of nodes in $P$;
   $R.add(p)$;
else
```

```
2.  **FP_mining**($t$, $null$);
3.  Reorganize frequent routes in $R$;
5.  **for** each $a_i$ in $L$ **do**
6.      generate pattern $\beta = a_i \cup \alpha$ with
        support_count = the minimum
        freq of nodes in $\beta$;
7.      construct $\beta$'s conditional
        pattern base and conditional
        FP-tree $T'$;
8.      **if** ($T' \neq \emptyset$ ) **then**
9.          FP_mining(T', $\beta$);
10.     **if** (R remains unchanged) **then**
11.         $R$.add($\beta$);
```

In Algorithm 2, we firstly scan all streets in *R* to calculate how frequently they have been visited based on trajectory segments. If the frequency of a street is lower than *min_sup*, remove this street from the hotspot region (lines 1-4). Then, sort the remaining streets by frequency in descending order, save nodes' order in a head list *L* (line 5). Next, create a root node *t* of the FP-tree, label it as "*null*" (line 6), and insert all trajectory segments in *R* into *t* by calling the procedure of "**insert_tree**(·)" (lines 7-9). Finally, we output *t* and *L* (line 10).

When mining a FP-tree, however, the process is distinct from the FP-growth mining approach, as presented in Function 2. Basically, our objective is to find the longest frequent route whose *support_count* is higher than the given threshold. The difference lies in Step 2 and Step 6 of the procedure FP_mining(·), it simply generates the largest frequent pattern without obtaining their combinations. The purpose of Step 3 in Function 2 is to reorganize streets in each frequent pattern into a route.

By processing each hotspot region using Algorithm 2, we can obtain a set of frequent routes for hotspot regions. Given a previous moving route of an object, we can compare it with discovered frequent routes and find the most matching one, and then use this frequent route to predict trajectories of this object.

5 Time-Stamped Trajectory Prediction

In this section, we present and analyze the work mechanism of E^{3}TP, especially for the approach of extracting temporal information from trajectory data and obtaining timestamps of frequent routes. The E^{3}TP algorithm is detailed in Algorithm 3.

```
Algorithm 3. E^3TP- Effective&Efficient&Easy Trajectory Prediction

Input: a historical trajectory data set D, the set of all streets E in a map m, hot-
spot mining threshold p, the minimum support count s for FR_mining, and the
previous trajectory t of an aiming object
Output: a time-stamped trajectory t' of o
Hotspot_mining(D, E, p);
FR_mining(m, s);
SpeedCalculation(D, E);
TrajPredict(t, m, E);
```

In Algorithm 3, Steps 1 and 3 are proposed to discover the movement rules from a certain map based on historical data. Steps 1 and 2 have already been introduced in Section 3 and Section 4, respectively. In Step 3, we obtain the possible speed for moving objects in each street by calculating their average speed. Ordinarily, an object moves at a constant speed, so we use the average speed to approximate it. Here, the category of moving objects represents its internal characters, and the *ID* of a street indicates its individual features, such as road condition and traffics. The obtained rules can be quickly retrieved when needed.

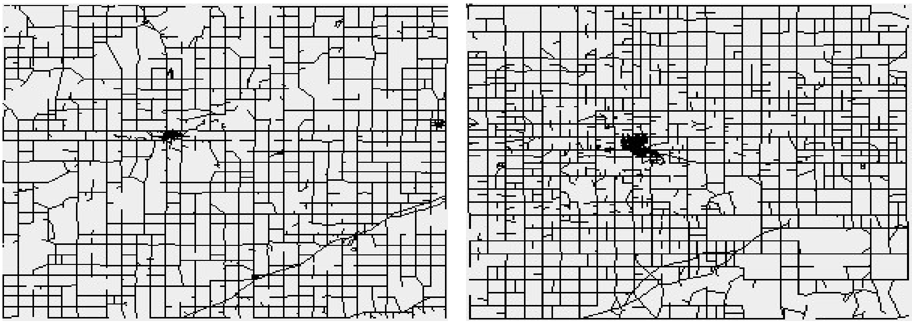
For predicting trajectories of moving objects, we firstly check whether an object is currently in a hotspot region. If so, we extract its previous trajectory, transform it into a route, and compare it with existing frequent routes in the hotspot region in order to find the most matched one based on Equation 2. Then, we employ kinematical formulations [10] to compute time intervals of visiting a street, and obtain the accurate position of objects at given timestamps. In this phase, we suppose that an object moves in a uniform speed, with the speed obtained by Step 3. Finally, we output timestamps and locations of the object in the form as given in Equation 1.

6 Experiments and Discussions

6.1 Experimental Setting and Definition

In this section, we perform experiments by comparing E^{3}TP with the naïve prediction method (called Naïve for short). In general, Naïve does not consider the hotspot mining and frequent pattern mining phases to predict possible routes. It only calculates the frequency of visiting each street and chooses the most frequently visited street to go when it arrives at a crossroad. Both algorithms were implemented in Java and the experiments were conducted on an AMD Athlon X2 5000+, 2.6GHz CPU with 2GB of main memory, running on Windows XP professional system.

All experiments were running on two distinct data sets generated by Brinkhoff's network-based generator [9]. They were generated by the network-based spatio-temporal data generating approach [9]. The moving objects are represented by its $\{xy\}$-coordinate in a real-world map. These two maps are shown in Fig. 3.



(a) The map of Kansas (b) The map of New York

Fig. 3. Maps for experiments

- A part area of Kansas State (denoted as Kansas).
- A part area of New York State (denoted as NY).

The parameter r in Equation 2 plays an essential role in determining whether two trajectories are equivalent. It has a great impact on the experimental results. Another important parameter is t, which decides the visiting time interval of an object's trajectory. In order to tune these two parameters, we gradually increase the value of r in a specified t value, aiming to find a proper r value.

In this section, we use *accuracy* defined in Equation 4 to estimate the effectiveness of prediction algorithms, and it is given as follows.

$$accuracy(N) = \frac{n_{hit}}{N} \tag{4}$$

In Equation 4, n_{hit} represents the number of predicted trajectories which match the real-world situation, and N is the number of all predicted trajectories.

6.2 Parameter Settings

As introduced in the previous sections, there are four important parameters that need to be specified first, including p (introduced in Algorithm 1), *min_sup* (see Algorithm 2), r (as shown in Equation 2) and t (the time interval during which prediction accuracy is evaluated). Due to the differences between these two data sets, these parameters should be specified separately. The properties and parameter settings of these two data sets are presented in Table 1**.**

Table 1. Properties of data sets and parameter settings

Parameter	*Kansas*	*NY*
Map width (pixel)	437,998	563,287
Map height (pixel)	348,693	435,186
Number of moving objects	8,000	8,000
p	0.2	0.2
min_sup	28	32
r (pixel)	2,000	2,000
Time interval t (time unit)	15	15

Table 1 indicates that each data set contains 8,000 moving objects. We use the historical data of 4,000 objects to build movement rules for each map, and the other 4,000 objects are employed to estimate the accuracy of our proposed algorithms. By trade-off, p is set to 0.2 for both data sets. When p equals 0.2, the minimum values of *min_sup* for both data sets are determined by f, because it is the minimum visiting frequency in terms of all streets in hotspot regions. We choose to use this minimum value in order to perform more accurate predictions. The setting of r depends on the scale of both maps (it is set 2,000 pixel based on these two maps), and t is set to 15 time units, which is large enough to compare both algorithms, i.e., Naïve and E^{3}TP.

6.3 Performance Analysis of Movement Rules Extraction

This section analyzes the time trend and memory cost in terms of hotspots region construction and frequent pattern mining phases as the number of trajectories grows.

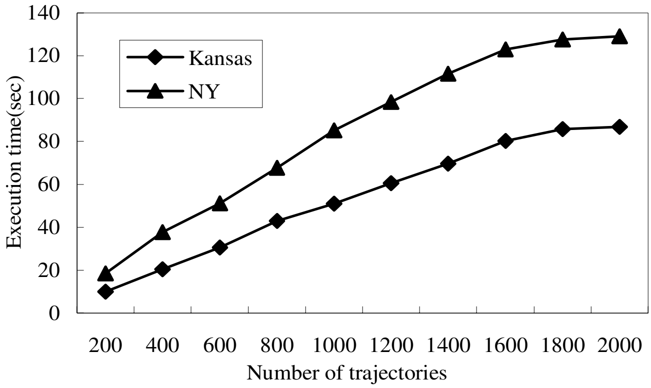


Fig. 4. Runtime comparison of hotspots region construction and frequent routes mining

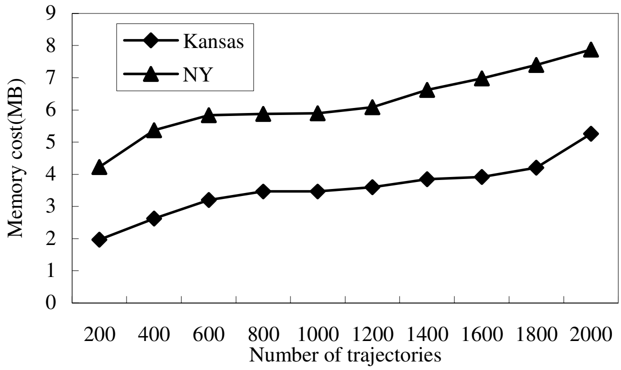


Fig. 5. Memory cost of hotspots region construction and frequent routes mining phases

Fig. 4-5 show the time performance comparison and memory cost of movement rules extraction and frequent pattern mining between Naïve and E^{3}TP, respectively, where the *x* axis is the number of trajectories, and the *y* axes represent the runtime and memory cost, respectively. According to Fig. 4, the execution time of hotspots region construction and frequent routes mining phases increases in an approximate linear manner with the number of trajectories growing. The reason is that most prediction time lies in trajectory processing and route searching, and the time complex approximates $O(n)$. As for Fig. 5, the memory cost does not increase drastically as the number of trajectories increases. Because redundant trajectory information is cut off when transforming historical trajectory data into frequent routes.

6.4 Prediction Accuracy Estimation

Prediction accuracy is used to estimate the performance of both Naive and E^{3}TP.

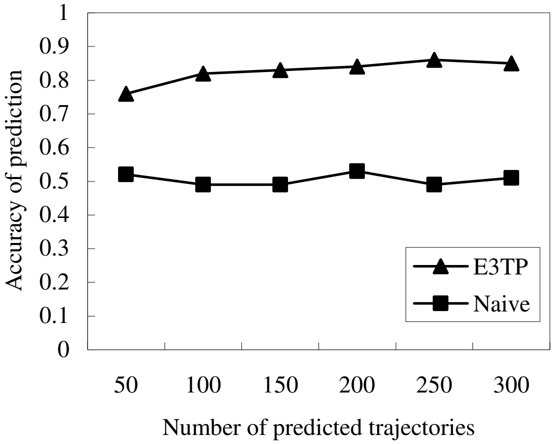


Fig. 6. Prediction accuracy comparison between E^{3}TP and Naïve on Kansas data set

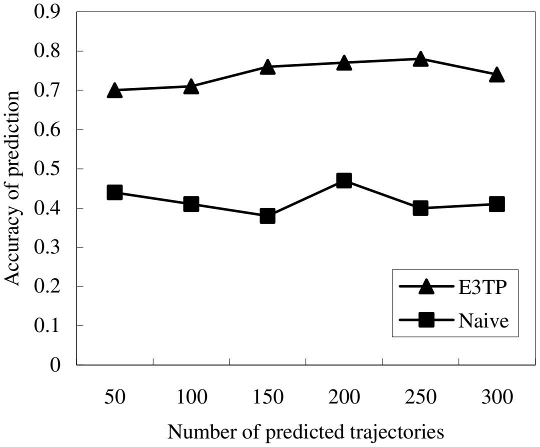


Fig. 7. Prediction accuracy comparison between E^{3}TP and Naïve on NY data set

In Fig. 6 and Fig. 7, the *x* axis represents the number of predicted trajectories, and the *y* axis is the percentage of trajectories that "hit" (match) the real-world situation by setting parameters as introduced in Section 6.2. By Fig. 6 and Fig. 7, E^{3}TP outperforms the naive method with a big gap of 32% on Kansas data set and 30% on NY data set, respectively. This is because E^{3}TP provides a better solution of finding possible routes of moving objects with high- frequent patterns in a certain region. In addition, we predict the motion curves of moving objects by taking into account both the road condition and the internal characteristics of moving objects, which helps improve the accuracy of calculating timestamps.

7 Conclusion and Further Work

Traditional trajectory prediction algorithms mainly focus on discovering frequent movement patterns of moving objects without constraints, which is far from the real-world situations. In this paper, we addressed the characteristics of motion behaviors of moving objects, including the key factors in route selection and the distribution rules of frequent trajectory patterns. By experiments, we compare the proposed trajectory prediction algorithm based on FP-tree, called E^{3}TP, with probability-based naïve method, and show the advantages of our solution. Most importantly, E^{3}TP can be used to prediction the fleeing criminals in order to help security agencies trace criminal suspects.

For further research, we will improve the proposed frequent routes discovery algorithm to be more suitable and effective for mining large scale of spatial and temporal data, and develop other data mining approaches, i.e., Genetic Algorithms [11, 12], neural networks [13], immune algorithms [14], to accurately find the most possible route. In addition, stochastic theories can be utilized to depict the distribution rules of moving objects' speed. Finally, trajectory prediction of moving objects in non-hotspot region will be seriously considered in our future study.

References

[1] Qiao, S., Tang, C., Jin, H., Dai, S., Chen, X.: Constrained K-Closest Pairs Query Processing Based on Growing Window in Crime Databases. In: 2008 IEEE International Conference on Intelligence and Security Informatics, ISI 2008, Taipei, pp. 58–63 (2008)

[2] Morzy, M.: Mining frequent trajectories of moving objects for location prediction. In: Perner, P. (ed.) MLDM 2007. LNCS, vol. 4571, pp. 667–680. Springer, Heidelberg (2007)

[3] Lee, J., Han, J., Whang, K.: Trajectory Clustering: A Partition-and-Group Framework. In: SIGMOD 2007, Beijing, China, pp. 593–604. ACM, New York (2007)

[4] Trajcevski, G., Wolfson, O., Zhang, F., Chamberlain, S.: The geometry of uncertainty in moving objects databases. In: Jensen, C.S., Jeffery, K., Pokorný, J., Šaltenis, S., Bertino, E., Böhm, K., Jarke, M. (eds.) EDBT 2002. LNCS, vol. 2287, pp. 233–250. Springer, Heidelberg (2002)

[5] Han, J., Pei, J., Yin, Y.: Mining frequent patterns without candidate generation. In: SIGMOD 2000: Proceedings of the 2000 ACM SIGMOD international conference on Management of data, pp. 1–12. ACM, New York (2000)

[6] Trajcevski, G., Wolfson, O., Hinrichs, K., Chamberlain, S.: Managing uncertainty in moving objects databases. ACM Trans. Database Syst. 29(3), 463–507 (2004)

[7] Giannotti, F., Nanni, M., Pedreschi, D.: Efficient mining of temporally annotated sequences. In: SDM 2006: Proceedings of the 6th SIAM International Conference on Data Mining, pp. 346–357. SIAM, Bethesda (2006)

[8] Giannotti, F., Nanni, M., Pedreschi, D., Pinelli, F.: Mining sequences with temporal annotations. In: SAC 2006: Proceedings of the 2006 ACM symposium on Applied computing, pp. 593–597. ACM, New York (2006)

[9] Brinkhoff, T.: A framework for generating network-based moving objects. Geoinformatica 6(2), 153–180 (2002)

[10] Halliday, D., Resnick, R., Walker, J.: Fundamentals of Physics, 8th edn. Wiley, Chichester (2007)

[11] Qiao, S., Tang, C., Peng, J., Fan, H., Xiang, Y.: VCCM Mining: Mining Virtual Community Core Members Based on Gene Expression Programming. In: Chen, H., Wang, F.-Y., Yang, C.C., Zeng, D., Chau, M., Chang, K. (eds.) WISI 2006. LNCS, vol. 3917, pp. 133–138. Springer, Heidelberg (2006)

[12] Qiao, S., Tang, C., Peng, J., Hu, J., Zhang, H.: BPGEP: Robot Path Planning based on Backtracking Parallel-Chromosome GEP. In: Proceedings of the International Conference on Sensing, Computing and Automation, ICSCA 2006, DCDIS series B: Application and Algorithm, vol. 13(e), pp. 439–444. Watam Press (2006)

[13] Qiao, S., Tang, C., Peng, J., Yu, Z., Jiang, Y., Han, N.: A Novel Prescription Function Reduction Algorithm based on Neural Network. In: Proceedings of the International Conference on Sensing, Computing and Automation, ICSCA 2006, DCDIS series B: Application and Algorithm, vol. 13(e), pp. 939–944. Watam Press (2006)

[14] Shao-jie, Q., Chang-jie, T., Shu-cheng, D., Chuan, L., Yu, C., Jiang-tao, Q.: SIGA: A novel self-adaptive immune genetic algorithm. Acta Scientiarum Natralium Universitatis Sunyatseni 47(3), 6–9 (2008)

A User-Centered Framework for Adaptive Fingerprint Identification

Paul W.H. Kwan[1,*], Junbin Gao[2], and Graham Leedham[1]

[1] School of Science and Technology, University of New England,
Armidale NSW 2351, Australia
Phone: +61-2-6773-2034; Fax: +61-2-6773-3312
{paul.kwan,graham.leedham}@une.edu.au
[2] School of Accounting and Computer Science, Charles Sturt University,
Bathurst NSW 2795, Australia
jbgao@csu.edu.au

Abstract. In recent years, law enforcement personnel have been greatly aided by the deployment of automated fingerprint identification systems (AFIS). These "black-box" systems largely operate by matching distinctive features automatically extracted from fingerprint images for their decisions. However, current systems have two major shortcomings. First, the identification result depends solely on the chosen features and the algorithm that matches them. Second, these systems cannot improve their results by benefiting from interactions with expert examiners who often can identify small differences between fingerprints. In this paper, we demonstrate by incorporating *Relevance Feedback* in a fingerprint identification system as an add-on module, a persistent semantic space over the database of fingerprints for an expert user can be incrementally learned. Here, the learning module makes use of a *Dimensionality Reduction* process that returns both a low-dimensional semantic space and an out-of-sample mapping function, achieving a two-fold benefits of data compression and the ability to project novel fingerprints directly onto the semantic space for identification. Experimental results demonstrated the potential of this user-centered framework for adaptive fingerprint identification.

Keywords: User-centered, Biometrics, Fingerprint identification, Adaptive information processing, Relevance feedback, Dimensionality reduction.

1 Introduction

Biometric authentication based on a person's physiological and behavioral traits is gaining acceptance as a method for uniquely verifying one's real identity [1]. Among these biometric traits: fingerprint, face, speech, iris and hand geometry are the most commonly used. Biometric authentication systems have been applied with successes in a number of real world applications in law enforcement, border control, welfare services, etc. An early example of this technology was the Automated Fingerprint Identification System (AFIS).

[*] Correspondence author.

H. Chen et al. (Eds.): PAISI 2009, LNCS 5477, pp. 89–100, 2009.
© Springer-Verlag Berlin Heidelberg 2009

However, current systems have two major shortcomings. First, the result of identification depends solely on the features selected and the algorithm that matches them. Second, there is no way of having these systems adapt their outcomes to seasoned examiners, who often can identify minute differences between fingerprints beyond what is capable of by current systems. In other words, most AFIS have a static processing architecture that lacks a functionality to capture and reuse knowledge of expert examiners in constructing the identification outcome. As an illustration, a simplified model of current generation systems is shown in Figure 1.

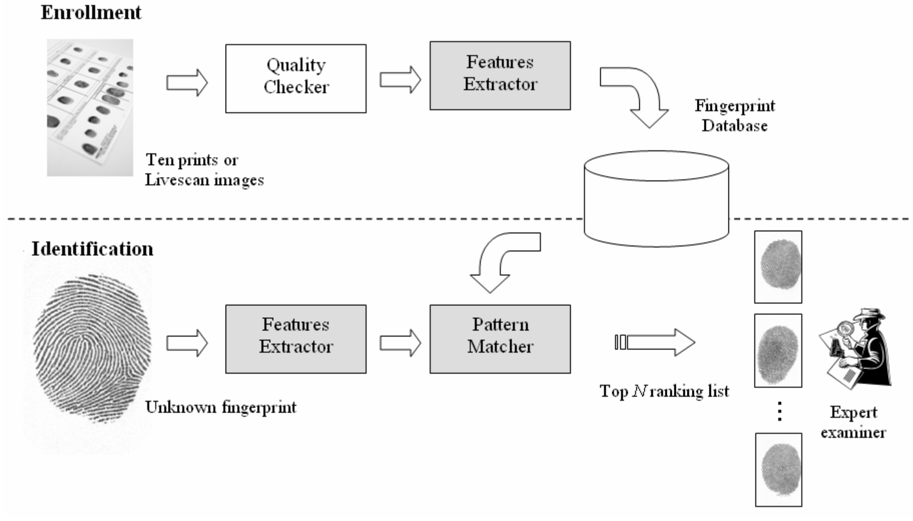


Fig. 1. A simplified model of current Automated Fingerprint Identification System (AFIS)

Due to both the Features Extractor and the Pattern Matcher being fixed, there is no way for improving the identification result even if impressions of the same finger as the unknown fingerprint did not turn up initially in the top *N* images of the ranking list. The user would be misled in judging that such a finger/identity does not exist in the database based only on the direct outcome of the system. The impact of such problem could potentially be minimized if feedbacks from an expert examiner on the relevance or irrelevance of certain fingerprints were captured, enabling the system to recalculate the ranking list accordingly. Here, we emphasize that the power to accept or reject the outcome of relevance feedback lies with the expert user.

In this paper, we demonstrate by incorporating *Relevance Feedback* in a fingerprint identification system as an add-on module, a persistent semantic space over the database of fingerprints for an expert user can be incrementally learned. Whereas relevance feedback has been extensively researched and applied in document retrieval and more recently in content-based image retrieval [2]; however, not much has been reported on integrating relevance feedback into biometric authentication systems both in research and in practice. One reason could be that in order for biometric authentication to benefit from relevance feedback, a supervised setting is necessary which is not possible in many deployment scenarios. However, in the case of an AFIS, the operating requirement makes it a suitable application for novel integration of relevance feedback and biometric authentication.

The remainder of this paper is organized as follows. In Section 2, an overview of the user-centered framework will be given. In Section 3, the fingerprint features used in this research will be briefly described. In Section 4, the major components of the proposed framework will be explained. In Section 5, experimental evaluation of the user-centered framework will be presented. Lastly, in Section 6, we will conclude and mention future directions.

2 Overview of Proposed Framework

The User-centered framework is made up of three main components including: *Input Space Transformation*, *Relevance Feedback*, and *Semantic Space Learning*. The framework is designed to be loosely rather than tightly coupled with other modules of the host AFIS as shown in Figure 2. As a result, it could be integrated as an add-on module in an existing system with some customizations.

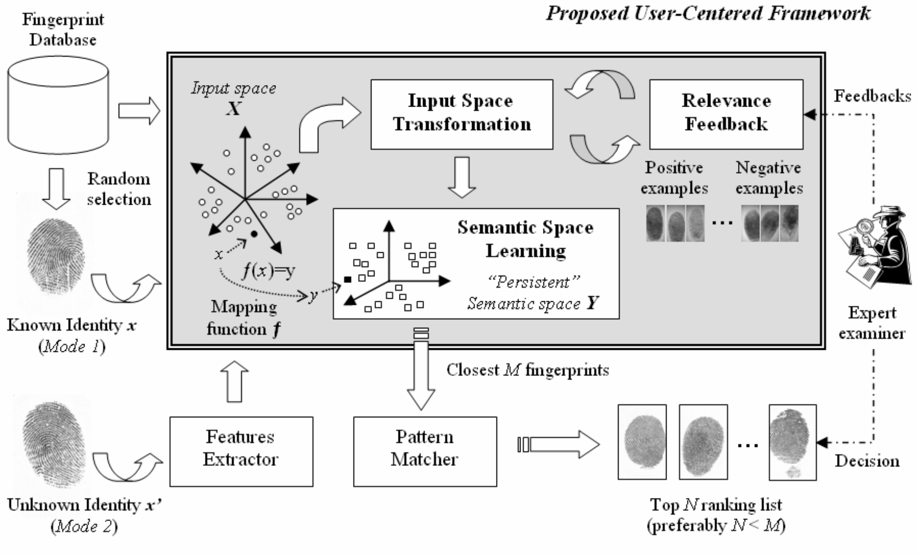


Fig. 2. Relationship of the proposed framework with modules of the host AFIS

Besides these three components, an important element of the learning framework is the input space *X* from which the persistent semantic space will be learnt. By a sequence of relevance feedback from an expert examiner the state of the input space will be updated. The source of data that forms the basis for constructing the initial input space is the set of extracted features, one for each fingerprint in the database. They are obtained by processing the corresponding fingerprint images by the Features Extractor module of the host system. Inside such module, normally a sequence of image processing steps are performed such as image enhancement, segmentation of fingerprint regions, detection and extraction of fingerprint features, and possibly mapping features to numeric values [3].

The proposed framework provides a novel mechanism by which an examiner can choose to incorporate his or her subjective knowledge into the construction of the persistent semantic space over the fingerprint database. There are four distinct steps in the execution of this mechanism, which are summarized as follows.

First, input to the framework is in the form of a fingerprint *x*, either by random selection from the database or taken from an unknown identity. These are denoted as Mode 1 and Mode 2 in Figure 2, respectively. Mode 1 is useful if the examiner chooses to incorporate additional knowledge into the formation of his or her persistent semantic space without being presented with an unknown identity. In such case, no features extraction is needed as the fingerprint is drawn from the database. In Mode 2, the examiner is being presented with a fingerprint from an unknown identity. In this case, features of the unknown fingerprint will be extracted by the host system before being passed to the framework.

Second, the examiner interacts with the framework via the *Relevance Feedback* component. Regardless of whether it is operating in Mode 1 or Mode 2, based on the fingerprint selected, the framework returns a subset of fingerprints (excluding *x*) that are similar based on the nearest neighbor criterion. The examiner selects as positive those fingerprints that are judged similar based on detailed observations. The negative selections are those that are judged dissimilar. Given the positive and negative selections, the corresponding entries in the distance matrix (the representation we used in this work) will be adjusted by the *Input Space Transformation* component, thereby transforming the input space *X*. The relevance feedback loop repeats until the user decides to exit. The outcome is a distance matrix that has learnt the semantic judgment of the expert examiner.

Third, based on the transformed distance matrix, the *Semantic Space Learning* component will either construct (if for the first time) or update the persistent semantic space of the expert examiner. To accomplish this, we modeled the learning process as a *Dimensionality Reduction* (*DR*) problem in which the input space corresponds to the *D*-dimensional features space while the semantic space to a lower-dimensional embedding space of dimension *d* ($d<<D$) [4]. Two advantages can be achieved by this modeling. First, the amount of computation that is required to operate in the high-dimensional features space can be significantly reduced by the data compression gained from *DR*. This makes both the learning and use of the semantic space more efficient. Second, by utilizing a suitable *DR* method that supports additionally an out-of-sample extension [5], a mapping function *f* that projects an unknown fingerprint onto the semantic space without repeating the learning process can be obtained. These advantages enable the learning framework to achieve the required efficiency.

Fourth, based on the state of the semantic space, a list of *M* fingerprints that are closest to the input can be identified. In the case of Mode 2 where the main objective is to decide if the unknown fingerprint is similar to any fingerprints stored in the database, this interim list will be passed to the Pattern Matcher module of the host AFIS. By the pattern matching process, a top *N* ranking list will be returned as the identification result to the expert examiner for his or her acceptance or rejection.

3 Fingerprint Features

Here, we summarize the steps of the features extraction algorithm used in the paper [3]. It employs both global and local ridge characteristics to construct a fixed length vector of size $D = 512$ for every fingerprint called FingerCode. Each FingerCode is comprised of an ordered enumeration of the features extracted from the local ridge characteristics contained in each sub-image or sector specified by a tessellation. As a result, each sector captures the local information and the ordered enumeration of the tessellation captures the invariant global relationships among these local patterns. Finally, Gabor filters are applied to decompose the local discriminatory characteristics in each sector into bi-orthogonal components based on their spatial frequencies. Figure 3 visualizes the process of features extraction carried out by [3].

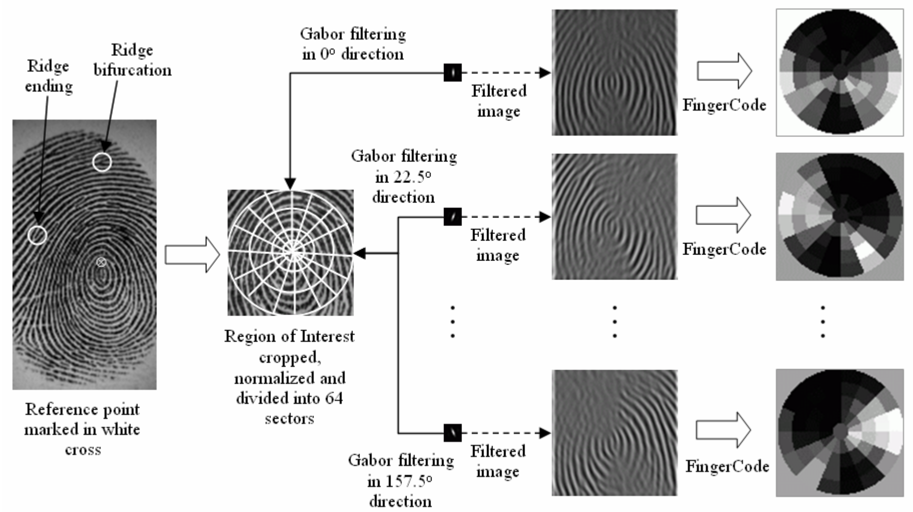

Ridge ending
Ridge bifurcation
Reference point marked in white cross
Region of Interest cropped, normalized and divided into 64 sectors
Gabor filtering in 0° direction
Gabor filtering in 22.5° direction
Gabor filtering in 157.5° direction
Filtered image
Filtered image
Filtered image
FingerCode
FingerCode
FingerCode

Fig. 3. Fingerprint features are extracted by [3] using a bank of Gabor filters aligned in eight different directions including {0°, 22.5°, 45°, 67.5°, 90°, 112.5°, 135°, 157.5°}

4 Framework Components

In this section, each of the three major components of the user-centered framework depicted in Figure 2 will be explained.

4.1 Input Space Transformation

The main function of this component is to transform the topology among objects of the input space X through iteratively updating the distance matrix. The ij-th element of this matrix, denoted M_{ij}, measures the Euclidean distance between the FingerCode vectors of fingerprints i and j in R^D as:

$$M_{ij} = \sqrt{\sum_{k=1}^{D} \left(i_k - j_k\right)^2} \qquad (1)$$

Based on these pair-wise distances, the semantics of the input space can be encapsulated by the n-by-n real-valued matrix M, where n equals the number of fingerprints that are involved.

The amount of updating is determined by the subjective knowledge input by the expert examiner as captured through the *Relevance Feedback* component, which will be described in Section 4.2. Against an input fingerprint x, after one iteration of relevance feedback a list of *nret* most similar fingerprints in the database is retrieved and shown to the examiner. From the list, the examiner can indicate as positive selections those that he or she considers similar and negative selections those that are dissimilar. The sets of positive and negative selections are denoted by $P = \{p_1, p_2, \ldots, p_i\}$and $N = \{n_1, n_2, \ldots, n_j\}$ respectively, with $nret = i + j$. In addition, there is an adjustable parameter $\beta \in (0,1]$ that controls the amount of increase or decrease made to the entries of the distance matrix after each iteration. The initial values for *nret* and β used in our experiments were 10 and 0.8, respectively.

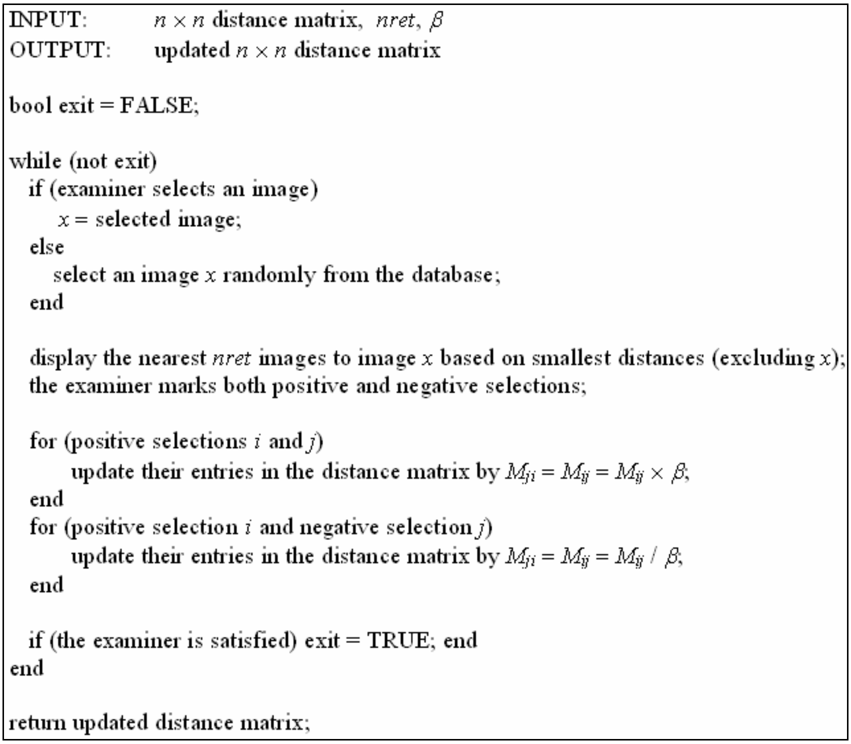
Here, it is worthwhile to emphasize that a distance matrix is only one method of encapsulating the semantics of an input space. While it is not a goal in this paper to compare the relative performances of different representations, we note that in recent years a number of research efforts have proposed alternate representations that are more suitable in certain situations. One of these is by using a kernel function as similarity measure, thereby resulting in a kernel Gram matrix that captures the pair-wise similarity between objects in a potentially very high-dimensional features space [6]. Another representation would be the use of a pure metric space where only the pair-wise distances are known, while the geometrical properties of a Euclidean space is not required.

4.2 Relevance Feedback

In the user-centered framework, an expert examiner interacts with the fingerprint identification system via the *Relevance Feedback* component. Relevance feedback, an *adaptive information processing* technique, was first applied in document retrieval in the 1960s. It was later adapted and used in content-based image retrieval (CBIR) that has a strong link to this research. In its most common form, relevance feedback involves polling the user for feedback on the relevancy of the current retrieval results. Based on the feedback, the system learns and improves its performance in the next round, iteratively if necessary.

As described in Section 2, the proposed framework has two modes of operation at present. For the initial formation and subsequent updating of the semantic space, *Mode* 1 is used. In this mode, a fingerprint x in the database can either be picked randomly or chosen by the examiner. Based on fingerprint x, a subset of fingerprints (excluding x) that are similar based on closest distances is returned by the *Input Space Transformation* component. Through the graphical user interface (GUI), the examiner marks as positive selections those fingerprints that are judged similar according to his or her subjective knowledge. The negative selections are those that are judged dissimilar. These feedbacks are passed back to the *Input Space Transformation* component where the corresponding entries in the distance matrix will be decreased or increased accordingly. The relevance feedback loop repeats until the examiner decides to exit the current "learning" session. The outcome is a transformed distance matrix that has incorporated the subjective knowledge of the fingerprint examiner.

Below, we summarize the relevance feedback process by the pseudo code given in Figure 4. The inputs include the *n*-by-*n* distance matrix, the parameter *nret* indicating the number of fingerprints included in the feedback, and *β* that determines how much the entries of the distance matrix will be increased or decreased.



```
INPUT:      $n \times n$ distance matrix, $nret$, $\beta$
OUTPUT:     updated $n \times n$ distance matrix

bool exit = FALSE;

while (not exit)
  if (examiner selects an image)
      $x$ = selected image;
  else
      select an image $x$ randomly from the database;
  end

  display the nearest $nret$ images to image $x$ based on smallest distances (excluding $x$);
  the examiner marks both positive and negative selections;

  for (positive selections $i$ and $j$)
      update their entries in the distance matrix by $M_{ji} = M_{ij} = M_{ij} \times \beta$;
  end
  for (positive selection $i$ and negative selection $j$)
      update their entries in the distance matrix by $M_{ji} = M_{ij} = M_{ij} / \beta$;
  end

  if (the examiner is satisfied) exit = TRUE; end
end

return updated distance matrix;
```

Fig. 4. Pseudo code of the relevance feedback process

4.3 Semantic Space Learning

In this research, the "learning" of the persistent semantic space is modeled as a dimensionality reduction process that projects the higher-dimensional features space onto a lower-dimensional semantic space. The two-fold benefits are data compression and a mapping mechanism that can project an unknown fingerprint onto the semantic space without repeating the entire learning process. This is used when the proposed framework is operating under *Mode* 2 in which the fingerprint examiner is being presented a fingerprint of an unknown identity *x*'. Taking its feature vector as input, the framework projects it onto the semantic space by using the learnt mapping function. From this, the closest *m* fingerprints can be identified and passed as input to the Pattern Matcher module of the host system to obtain the top *n* ranking list for the examiner's evaluation.

Whereas traditional linear methods like Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) have been used in dimensionality reduction, a number of new techniques have been proposed recently for dealing with the inherent non-linearity that exists in the relational structure among complex objects like those of biometric data [6]. In addition to the lower-dimensional representation for the input data, some of the DR methods can return a direct *out-of-sample* mapping function by which novel input can be projected onto the latent space while others make use of estimation techniques that are more universally applicable.

In this paper, due to our choice of representing the input space as a distance matrix, we have selected two representative methods, namely Multi-Dimensional Scaling (MDS) and Laplacian Eigenmaps (LE), as candidates for the learning of the persistent semantic space because they make use of the distance matrix in their DR process. The former is a global non-linear method while the latter is a local non-linear method, according to the taxonomy given in [4]. However, as both of these methods do not return an out-of-sample mapping function directly, we resort to using an *estimation technique* to achieve the similar objective.

In order to assess the improvement in identification accuracy due to relevance feedback, we will compare the results obtained by MDS and LE (both employing relevance feedback) with PCA and Locality Preserving Projections (LPP) (both not employing relevance feedback) in our empirical experiments. First, the reason for selecting PCA in our comparison is that it is often used as a benchmark while being able to return a linear mapping function for projecting novel input onto the semantic space. Second, the reason for including LPP in our comparison is that while it employs a distance matrix (an extension of LE) in its DR process, it is not required to have the matrix updated. Furthermore, it can return a linear mapping function for out-of-sample extension directly.

5 Experimental Evaluation

To demonstrate the potential of the proposed framework for improving identification accuracy, several experiments were conducted on a subset of the MCYT-Fingerprint-100 (Ministerio de Ciencia y Tecnología, Spanish Ministry of Science and Technology) sub-corpus collected by the Biometric Research Laboratory - ATVS of the Universidad Politecnica de Madrid under the MCYT project [7]. The MCYT-Fingerprint-100 sub-corpus consists of ten prints, each having 12 impressions, of 100 people taken using two different acquisition devices, making a total of 24,000 (100 × 10 × 12 × 2) fingerprints.

For our experiments, we randomly chose 50 fingers out of the sub-corpus, resulting in a database of 1,200 (50 × 12 × 2) fingerprints. In these experiments, 1,100 fingerprints (i.e., 11 × 2 impressions from each finger) comprised the training set while the remaining 100 fingerprints as the test set. The test fingerprints will be used as query in our experiments. We used three parameters $nc = 50$, $ns = 24$, and $tns = 22$ to denote the number of fingers (or classes), the total number of impressions for each finger, and the number of impressions for each finger in the training set, respectively.

In performing our experiments, we addressed the limitation of not being able to involve actual fingerprint examiners at this stage by developing a software module to simulate the quality of relevance feedback (right versus wrong decisions) that would have been given by either a *normal* or a *strong* expert. To accomplish this, in our simulation we made use of a random number r generated from a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$ that obeys the 68-95-99.7% rule (Figure 5). For each fingerprint that appears in the list of most similar m fingerprints after each iteration of relevance feedback, we decide if the expert would make a right or wrong decision based on the following two simple rules:

$$\text{Normal expert:} \quad |r| \le 1 \quad \text{(right)}, \quad |r| > 1 \quad \text{(wrong)} \tag{2}$$

$$\text{Strong expert:} \quad |r| \le 2 \quad \text{(right)}, \quad |r| > 2 \quad \text{(wrong)} \tag{3}$$

In other words, for similar fingerprints that were wrongly judged as dissimilar, their distances from the novel input will be increased (divide by β) while dissimilar fingerprints that were incorrectly judged as similar will be decreased (multiply by β).

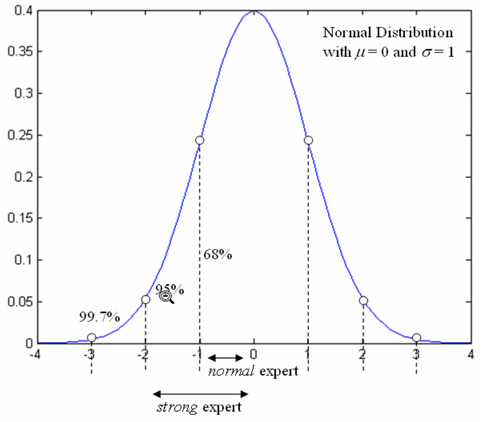


Fig. 5. Simulate *normal* and *strong* experts in the experiments using the 68-95-99.7% rule of normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$

Furthermore, to ensure that our experimental results are repeatable, the following three conditions were adhered consistently in our experiments:

1. A sequence of fingerprints (each identified by a unique number in the database) was generated beforehand, and used in the experiments that involve relevance feedback;
2. The dimensionality of the semantic space (d = 6) is estimated by a Maximum Likelihood Estimator based on the training set; and
3. k = 12 as the number of nearest neighbors used in MDS, LE and LPP to construct the neighborhood graph in their DR process.

5.1 Experiment #1

This experiment compares visually the mapping of the set of FingerCode vectors from the input space (D = 512) to the semantic space (d = 6) by different DR methods. In Figure 6, the left column plots the initial semantic space in the first two dimensions for 7 of the 50 fingers used in experiments for sake of illustration. The right column shows the updated semantic space after projecting the test set using OOS extension.

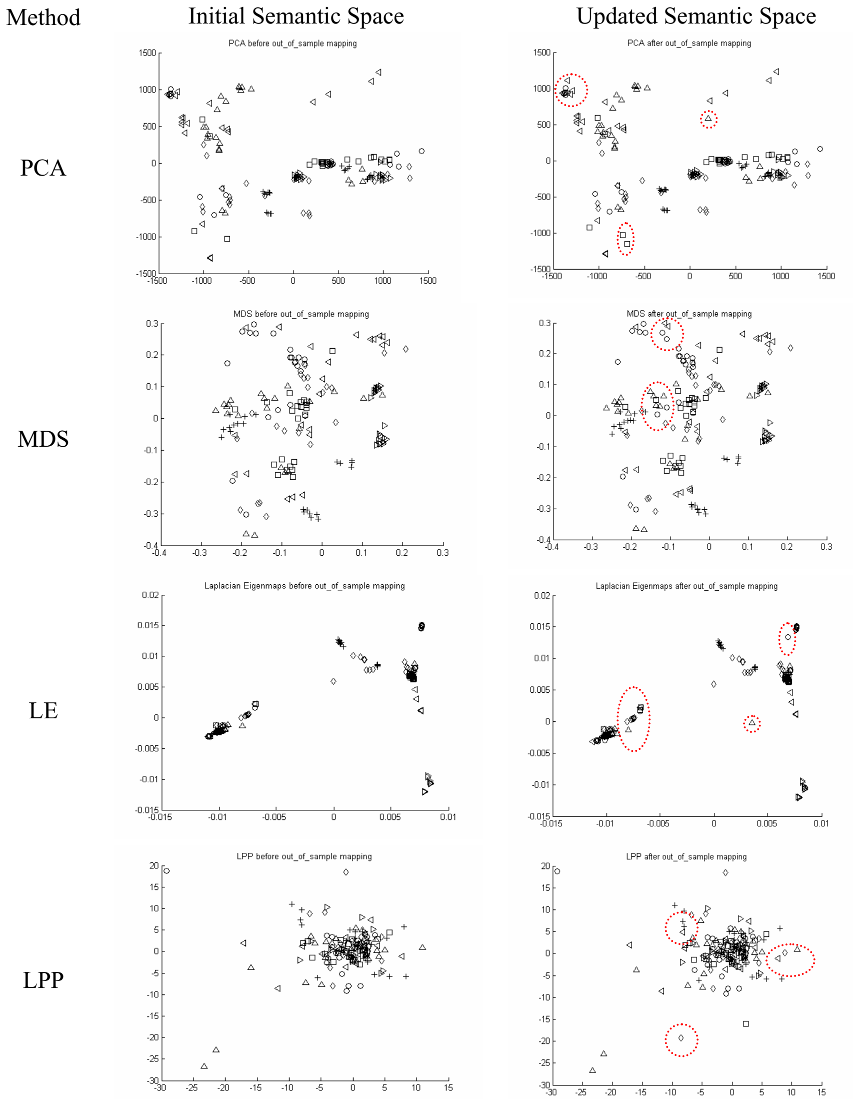


Fig. 6. Left column plots the initial semantic space in the first two dimensions for 7 of the 50 fingers used in experiments for sake of illustration. Right column plots the updated semantic space after projecting the test set (some highlighted in dotted circles) using OOS extension.

5.2 Experiment #2

The second experiment attempts to compare the difference between a *normal* and a *strong* expert by their effects on the identification accuracy measured using the *k*-NN classification errors. The left sub-figure of Figure 7 shows the result by the normal expert while the right sub-figure the result by the strong expert, respectively. In these figures, only MDS and LE that incorporated relevance feedback into their DR process are shown. The baseline refers to the *k*-NN classification errors obtained by using the default Euclidean distance in the initial high-dimensional features space. Note that, a suffix like “rf_10” meant the result obtained after 10 iterations of relevance feedback.

From the plots of Figure 7, it is reasonable to conclude that there is no significant difference in identification accuracy between a normal and a strong expert (as defined earlier in this section) based on our experimental setup. Based on this comparison, we have therefore decided to simulate a normal expert in the final experiment as this would be more representative of the real world situation.

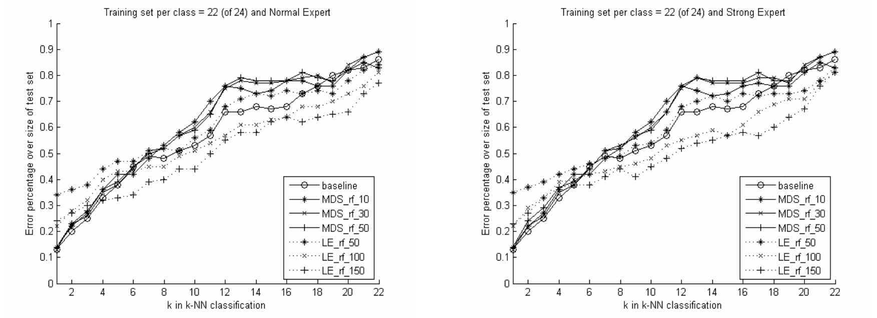


Fig. 7. Comparison of a *normal* expert (left plot) and a *strong* expert (right plot) on their identification accuracy based on *k*-NN classification errors for “relevance feedback” enabled MDS and LE

5.3 Experiment #3

In the final experiment, we compare the identification accuracy obtained by applying PCA, LPP, MDS, and LE. One might recall that both PCA and LPP return a linear mapping for out-of-sample extension directly albeit without incorporating relevance feedback, while both MDS and LE do in this experiment.

In Figure 8, one can observe that PCA performed worse than the baseline for all values of *k* while LPP did significantly worse. For MDS, one can notice that there is no significant improvement in identification accuracy even after going through 30 or 50 iterations of relevance feedback. On the other hand, while LE started out having worse performance than the baseline, PCA and MDS; after 100 iterations of relevance feedback, it has already achieved better accuracy than the baseline for $k > 6$; while after 150 iterations, it has better accuracy for k > 4. A conclusion based on the results of this empirical experiment could be drawn here. That is, certain DR methods like LE is capable of obtaining more improvement in accuracy than others such as MDS, which also exploited relevance feedback.

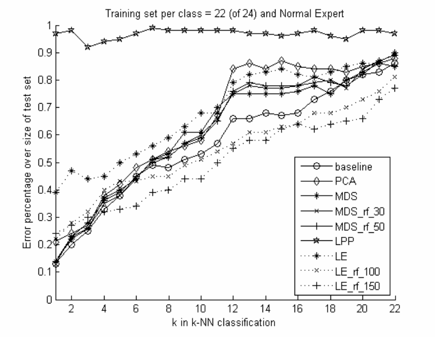


Fig. 8. Comparison of identification accuracy of MDS, LE and their "relevance feedback" enabled extensions with PCA and LPP

6 Conclusions

In this paper, we have introduced a user-centered framework for adaptive fingerprint identification that can be incorporated as an add-on module in a host AFIS. This is achieved by exploiting relevance feedback to capture an expert examiner's subjective knowledge into the formation of a persistent semantic space over the fingerprint database in which the accuracy of identification might be potentially improved.

Several experiments were conducted on a subset of the MCYT-Fingerprint-100 sub-corpus to simulate the performance of the proposed framework. The experimental results demonstrated the framework's potential for adaptive fingerprint identification.

Future works include potential collaboration with Australian Federal Government's CrimTrac Agency in implementing the proposed framework in their AFIS for actual testing and developing approaches for adaptation to larger fingerprint databases.

References

1. Wayman, J., Jain, A.K., Maltoni, D., Maio, D.: Biometric Systems Technology, Design and Performance Evaluation. Springer, London (2005)
2. Zhou, X.S., Huang, T.S.: Relevance feedback in image retrieval: A comprehensive review. Multimedia Systems 8, 536–544 (2003)
3. Jain, A.K., Prabhakar, S., Hong, L., Pankanti, S.: Filterbank-Based Fingerprint Matching. IEEE Trans. Image Processing 9(5), 846–859 (2000)
4. van der Maaten, L.J.P., Postma, E.O., van den Herik, H.J.: Dimensionality Reduction: A Comparative Review. Neurocognition (2008) (submitted)
5. Bengio, Y., Paiement, J.F., Vincent, P., Delalleau, O., Le Roux, N., Ouimet, M.: Out-of-sample extensions for LLE, Isomap, MDS, eigenmaps, and spectral clustering. In: Advances in Neural Information Processing Systems, vol. 16, pp. 177–184. MIT Press, Cambridge (2004)
6. Guo, Y., Gao, J., Kwan, P.: Twin Kernel Embedding. IEEE Trans. Pattern Analysis and Machine Intelligence 30(8), 1490–1495 (2008)
7. Ortega-Garcia, J., et al.: MCYT baseline corpus: A bimodal biometric database. IEE Proceedings Vision, Image and Signal Processing 150(6), 395–401 (2003)

Design of a Passport Anti-forgery System Based on Digital Signature Schemes

Lei Shi[1,*], Shenghui Su[1], and Zhengrong Xiang[2]

[1] College of Computer Science, Beijing University of Technology,
100 Pingleyuan Chaoyang District, 100124 Beijing, P.R.China
{stony,shsu}@bjut.edu.cn
[2] School of Economics & Management, Beijing University of Posts & Telecoms,
10 Xitucheng Road Haidian District, 100876 Beijing, P.R.China
zrxiang@bupt.edu.cn

Abstract. The terrorism threat which is conducted by international terrorists who possess forged identities and passports becomes more and more serious and imminent. Traditional methods and techniques of preventing passport from forgery are not effectual enough. This paper proposes a passport anti-forgery system based on the public key signature scheme RSA. The anti-forgery system contains four modules: the passport dada edit module, passport key production module, passport signature module and passport verification module. A private key used for signing is preserved separately by an authorized issue official, and a public key used for verifying is stored into a database which a web server can access. In this way, the anti-forgery property of a passport is equivalent to the security of an adopted public key signature scheme, and the passport is bound with its issue authority tightly.

Keywords: Passport anti-forgery, Digital signature scheme, IC chip, Algorithm, Client / server.

1 Introduction

In the world of today, people are faced with the dangers from terrorism raids by terrorists who entered attacked countries through the customhouses resorting to forged passports. For example, on September 11, 2001, the 19 Islamist terrorists who were affiliated with al-Qaeda organization and seemed to own legal identities crashed the two hijacked airliners into Twin Towers of the World Trade Center in New York City, which resulted in the death of more than 2900 people, and caused the United States government to launch the War on Terrorism.

Naturally, the anti-forgery of a passport, which is a formal document issued by an authorized official of a country to one of its citizens that is usually necessary for exit from and reentry into the country, that allows the citizen to travel in a foreign country in accordance with visa requirements, and requests protection for the citizen while abroad, should be one part of the anti-terrorism movement.

[*] Project 2007CB311100 founded by MOST.

H. Chen et al. (Eds.): PAISI 2009, LNCS 5477, pp. 101–111, 2009.

© Springer-Verlag Berlin Heidelberg 2009

Traditional methods or techniques of preventing passports from forgery are mainly based principles of physics or optics. Practices showed that they were undependable to some extent. For instance, before the 9/11 attack, the terrorists with the forged passports went through the customhouse and entered into US.

This paper suggests a new thread of solving the problem —— the passport anti-forgery system based on digital signature scheme [1].

The passport anti-forgery system is a managemant information system, including four modules: the passport dada edit module, passport key production module, passport signature module and passport verification module. Additionally, an IC chip used to store the passport number and signing-code is embedded into a page of a passport.

The function of the passport dada edit module is to input, modify, delete and print the data of the holder of a passport. Generally, these data are displayed on the pages of the passport.

The function of the passport key production module is to produce a private key and a public key, preserve the private key into a flash memory file which is off-line, and store the public key into a relevant database file which is on-line.

The function of the passport signature module is to convert the passport data into a message digest through a hash function [2][3], sign the message digest with a private key, and store the signing-code and passport number into the IC chip of a passport through a writing IC function.

The function of the passport verification module based on client / server frame is to read the signing-code and passport number from the IC chip of a passport, obtain the corresponding passport data and public key from the relevant on-line databases, verify the signing-code with the public key, and return the verification result to the client.

Section 2 of the paper analyzes the requirements of the passport anti-forgery system, including the transaction flow, the data flow, and the relevant performance. Section 3 elaborates the preliminary design of the passport anti-forgery system, including the frame of the system and the four modules of the system. Section 4 describes the three main algorithms of the passport anti-forgery system, including the key generation algorithm, the digital signature algorithm and the identity verification algorithm. Section 5 makes the effectiveness and advantage analysis of the passport anti-forgery system.

2 Requirement Analysis of the Passport Anti-forgery System

In our country, the citizen exit-reenter administration is the Public Security Authority (shortly, PSA) and its subsections. A related requirement analysis is made.

2.1 Transaction Flow of Applying a Passport

If a citizen wants to obtain a passport, he first submits an application form to PSA, then PSA reviews his application information, and last PSA issues or does not issue a passport to the citizen according to the reviewing result.

The brief transaction flow of passport application is referred to Figure 1.

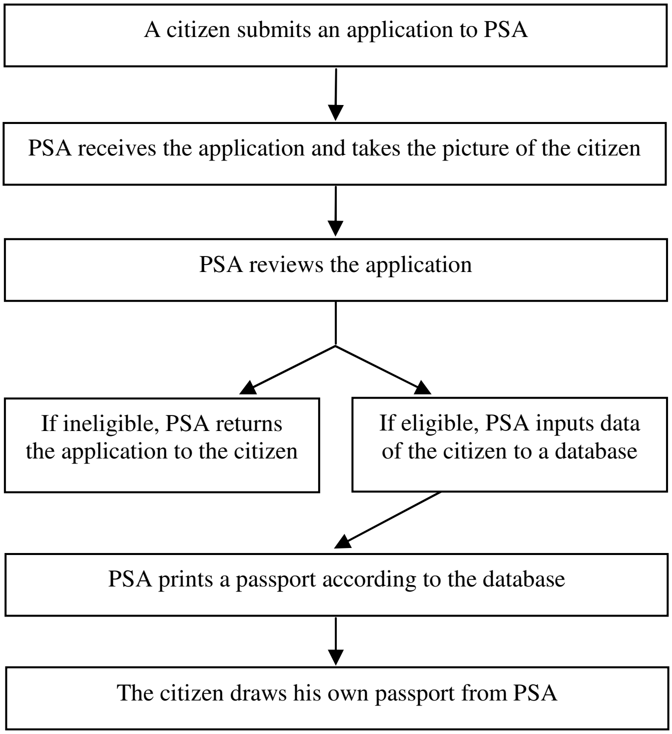

A citizen submits an application to PSA
PSA receives the application and takes the picture of the citizen
PSA reviews the application
If ineligible, PSA returns the application to the citizen
If eligible, PSA inputs data of the citizen to a database
PSA prints a passport according to the database
The citizen draws his own passport from PSA

Fig. 1. Transaction flow of a passport application

The data of the applicant citizen contains the name, birthday, identity number (ID number), nationality (country code), gender, face photo etc. These items addition to the passport number, issue authority, issue date and expiration date compose a passport data or a record. The passport date must be printed on a page of a passport conforming to the national regulations.

It should be noted that if a signing-code is produced afterwards, the signing-code should be printed on a page of a passport.

2.2 Data Flow of the System

The data flow through which the designers and users dialogue can be derived from the application, issue and verification of a passport. It consists of outer entities, processes and data storages. A box with a shadow represents an outer entity, a box with round angles does a process, and a cylinder does a data storage.

The brief data flow of the anti-forgery system is shown as figure 2.

The word ‘Condition’ means that PSA staffer should select a specific passport record for signature from the passport database.

The transmitted data between the server and the client are the same as those between the customs and the client.

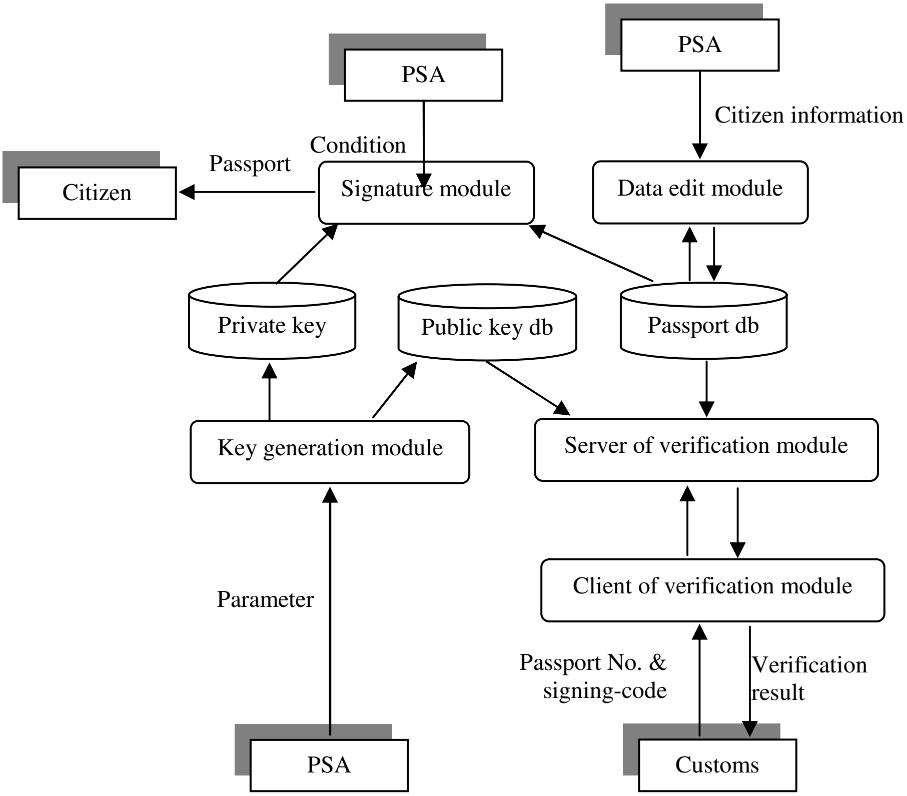


Fig. 2. Data flow of the anti-forgery system

2.3 Desire for the Passport Anti-forgery System

A passport is an identity and nationality certificate. It is issued to native citizens regulation by PSA according to legal rules, and is used to exit-reenter country and oversea trip. The information of holder is printed on the passport.

For convenience and expeditiousness, the system should satisfies the following requirements:

- Platform independence;
- Data being integrated highly;
- Modularization;
- Reliability and security.

Besides, the hard disk space of the relevant web servers should be large enough, and the operation speed of computers should be fast enough.

3 Preliminary Design of the Passport Anti-forgery System

The passport anti-forgery system is essentially a management information system, and composed of hardware and software two parts. Hardware contains client computers, database server computers, web server computers and IC (Integrated Circuit) chip read-write machines etc. Software contains Unix or Windows OS, Oracle database system, and the application modules: the passport data edit module, passport key production module, the passport signature module, and passport verification module.

3.1 Frame of the Anti-forgery System

Management information systems have mainly C/S (client/server) frame and B/S (browser /server) frame two types. C/S frame should be adopted in our fourth module. Development and upgrade of an application with C/S frame is not complex, and data access is not limited to a region or network.

The design of the system should consider that there may not be a direct connection between the Internet and the Intranet. All data must pass a firewall when applications on the Intranet access the Internet, or applications on the Internet access the Intranet. In parallel, a console process takes charge of copying the data on the Intranet to the Internet everyday. Inner computers can link to the Internet web server while outer computers may not link to the Intranet.

3.2 Module Desigh of the Anti-forgery System

The passport anti-forgery system consists of four modules: the passport data edit module, the key generation module, the passport signature module, and the passport verification module.

3.2.1 Passport Data Edit Module

The passport edit module runs on a general computer off-internet, and is managed by the Public Security Authority of an exit-reenter country. The module can input, save, modify and delete the data to be printed on a passport. It should perform the following functions:

(1) Create a database file for storing passport data. This file should contain these fields: ID number, name, face photo, nation code, passport type, passport number, issue date, issue authority, and public key number. The public key number comes from the key generation module. The date of birth, place of birth and gender are got from ID number. The date of expiration can be calculated by the date of issue plus 5 years or so. The database records should be sorted according to the ascent of the passport number.

(2) Exit-reenter country administrative department looks through the information of a passport applicant. If passed then write the information of the regular passport applicant into the database file. Each passport data has a corresponding record in the database.

(3) If the information of a passport is modified, then modify the record in the database file. The operation should be done via an operator authorized.

(4) If the passport is cancelled or out of date, then delete the record in the database file. The operation should be done via an operator authorized.

The database table structure of this module is designed as table 1.

Table 1. Fields of the passport database file

Number	Field	Full name	Type	Null or not
1	Idn	ID number	Char(18)	Not null
2	Name	Name	Char(20)	Not null
3	Cc	Country code	Char(3)	Not null
4	Fp	Face photo	Memo	Not null
5	Tc	Passport type code	Char(1)	Not null
6	Isd	Issue date	Date	Not null
7	Isp	Issue authority	Char(2)	Not null
8	Pn	Passport number	Char(9)	Not null
9	Pbkn	Public key number	Char(6)	Not null

The primary key is passport number, and the secondary keys are public key number and ID number.

3.2.2 Passport Key Production Module

This module runs on a computer off-internet, is only used by PSA and has a high secrecy. This module produces a public key and a private key by the RSA public key scheme. Its functions should be as follows:

(1) Set the parameters of the selected public key cryptosystem according to the stipulations — the length of a modulus and / or of a hash digest for example.

(2) Produce a public key and a private key by calling the key generation algorithm.

(3) Output the private key to an USB flash disk, which is kept by a special kernel official and must not be divulged.

(4) Store the public key to a database file which is provided for customs or other inspection organizations to check up.

The module designed a database table of key management for store the information about key.

The database table structure is shown as table 2.

Table 2. Fields of the public key database file

Number	Field	Full name	Type	Null or not
1	Pbkn	Public key number	Char(6)	Not null
2	Pbk	Public key	Varchar(256)	Not null
3	Pbka	Public key cryptosystem	Char(10)	Not null
4	Ked	Key expiration date	Char(8)	Not null
5	Pvkg	Private key guarder	Char(20)	Not null
6	Pvkn	Private key number	Char(2)	Not null

The primary key is the field named public key number.

Because many people are allowed to inquire about database information, the content of a private key is not stored in the database file directly. A private key guarder takes charge of an USB disk storing a private key. If the private key guarder mode is of multilevel, the field private key guarder deposits the name of the root guarder.

3.2.3 Passport Signature Module

Before a signing-code is injected into a passport, the IC chip embedding work should be finished in the special plant, and the passport printing work should also be done. The selected IC chips are non-touch-sensitive.

This module is used by country's exit-reenter administrations, has a high secrecy, and runs on a computer off-internet. It connects with an IC chip read-write machine.

Let *hash* be a hash function, and the concatenation of the binary strings of the items ID number, passport number, country code, name, issue authority and issue date be a message *m*. Notice that an IC chip number is not recognized as a part of the message since an ID number exists on a passport.

The module should performs the following functions:

(1) Calculate the digest of a passport message by calling *hash*(*m*).

(2) Read a private key from an USB disk.

(3) Regard the message digest and private key as the parameters, and produce the signing-code of a passport by the digital signature algorithm.

(4) Write the signing-code and passport number into the IC chip embedded in the passport via the IC read-write machine.

When writing the signing-code and passport number, we should avoid revealing the private key while it is used.

Because the name of a citizen is consistent with his ID number, and the country code is consistent with the issue authority, to enhance the operation speed, we may substitute the irreducible fields for the passport message, which means we regard the output of *hash*(ID number, passport number, issue authority, issue date) as the message digest. Correspondingly, when verifying a signing-code, we need to substitute the irreducible fields for a whole passport message. This will not affect the anti-forgery effectiveness of the system.

3.2.4 Passport Verification Module

This module has a client / server frame, and need to access databases placed in a web server. Customhouses or other inspection organizations use signing-codes to verify passports through the Internet.

Suppose that definitions of *hash* and *m* are the same as section 3.2.3.

The module should have functions as follows:

(1) Capture the signing-code and passport number via an IC read-write machine. The two data are submitted the web server from the client.

(2) Obtain a passport message according to the passport number, and calculate the digest by calling *hash*(*m*).

(3) Get the public key number from the database according to the passport number, and further find the corresponding public key.

(4) Calling the identity verification algorithm with the digest, public key and signing-code as the input parameters, and return the output result to the client.

(5) If the output result is ‘true’, then the passport is genuine. Otherwise, the passport is forged, and an alarm sound should ring.

The information stored in an IC chip embedded in a passport is sent to the web server through the Internet, and verification speed rests with the chosen identity verification algorithm.

The passport database file and the public key database file are placed on the Intranet, and they may be accessed by the verification module on the web server after the access is examined. Alternatively, for higher security and efficiency, the copies of the two database files may be placed directly on the web server.

4 Core Algorithms of the Modules

The RSA scheme [4][5] is chosen as a public key signature prototype in the passport anti-forgery system. Of course, other signature schemes are also considered when the design is implemented.

RSA includes three algorithms which are the core algorithms of the system. In what follows, they are optimized and described.

The high-powered REESSE recursive algorithm is selected to seek modular multiplicative inverses [6].

The symbol ‘gcd’ represents the greatest common divisor, and ‘%’ does a modular operation, and ‘&’ does a bitwise logical operation ‘and’.

4.1 REESSE Recursive Algorithm for Seeking Modular Inverses

Assume that a simple congruence is $ax \equiv b \pmod{m}$, that is, $ax - my = b$.

Input: a, b and m which pass values with $m \in \mathbb{Z}^+$, $a \in [1, m - 1]$, and $b \in [1, m - 1]$ or $b = 0$.

Output: x, y and d which pass names and satisfy $ax - my = b$ and $d = \gcd(a, m)$.

If $b = 1$ and $\gcd(a, m) = 1$, x is the multiplicative inverse. If $b \neq 1$ and $\gcd(a, m) \mid b$, x is a general solution. If $\gcd(a, m) \nmid b$, the congruence has no solutions, and the algorithm returns $x = \infty$.

```
S1: If ($a$ = 1 or $a$ = 0) then
    S1.1 Set $y \leftarrow 0$;
    S1.2 If ($a$=0 and $b \neq 0$) then $x \leftarrow \infty$, or else $x \leftarrow b$;
    S1.3 If ($a$ = 0) then $d \leftarrow m$, or else $d \leftarrow 1$;
    S1.4 Go to S6.
S2: Calculate $q_1$, $r_1$, $q_2$ and $r_2$ satisfying $m = q_1 a + r_1$ and $b = q_2 a + r_2$.
S3: Let $m \leftarrow a$, $a \leftarrow r_1$ and $b \leftarrow - r_2$.
S4: Call the function itself with $a$, $m$, $b$, $x$, $y$ and $d$.
S5: If ($x \neq \infty$) then $t \leftarrow x$, $x \leftarrow y + q_1 x + q_2$ and $y \leftarrow t$.
S6: Return to S5 of the superior level or the main.
```

Note that in the main function we should judge whether x and y are negative or not. If x is negative, we substitute $(x + m)$ for x, and if y negative, do $(y + a)$ for y. When the congruence has several solutions, the main function should compute the other solutions $(x + km / d) \bmod m$, where $k = 1, \ldots, d - 1$ from the known d and x.

4.2 Key Generation Algorithm

In this algorithm, need to call the Rabin-Miller algorithm [2], and the REESSE recursive algorithm [6].

Suppose that u and v are the bit lengths of two prime numbers. They are the input parameters of the algorithm.

```
S1: Obtain two large strong prime numbers $p$ and $q$ by calling the Rabin-Miller
    algorithm with $u$ or $v$ respectively.
S2: Compute $N \leftarrow pq$, $\varphi(N) \leftarrow (p–1)(q–1)$.
S3: Select $e$ making gcd $(e, \varphi(N)) = 1$.
S4: Compute $d$ satisfying $e\,d \equiv 1$ (% $\varphi(N)$) by calling the REESSE recursive
    algorithm.
S5: Return $(N, d)$ as a private key and $(N, e)$ as a public key to the main procedure.
```

Notice that the above algorithm is called by the module in section 3.2.2.

4.3 Digital Signature Algorithm

In the passport anti-forgery system, we substitute the SHA-1 algorithm for *hash*. SHA-1 is a security hash function by now, and outputs 160 bits message digest [3][7].

Suppose that (N, d) is the private key, and m is a passport message as section 3.2.3. They are the input parameters of the algorithm.

```
S1: Compute $h \leftarrow$ SHA-1$(m)$.
S2: Compute $S \equiv h^d$ (% $N$) through the following steps:
    S2.1: Set $S \leftarrow 1$;
    S2.2: If $d$ & $1 = 1$, compute $S \leftarrow S\,h$ % $N$;
    S2.3: Shift $d$ right by 1 bit;
    S2.4: If $d > 0$, compute $h \leftarrow h^2$ % $N$;
    S2.5: If $d > 0$, go to S2.2.
S3: Return the signing-code $S$ to the main procedure.
```

Notice that the above algorithm is called by the module in section 3.2.3.

4.4 Identity verification Algorithm

Let SHA-1 be the same hash function as section 4.2.

Suppose that (N, e) is the public key, m is the passport message, and S is the signing-code. They are the input parameters of the algorithm.

```
S1: Compute $h \leftarrow$ SHA-1$(m)$.
S2: Compute $h' \equiv S^e$ (% $N$) through the following steps:
    S2.1: Set $h' \leftarrow 1$;
    S2.2: If $e$ & $1 = 1$, compute $h' \leftarrow S\,h'$ % $N$;
    S2.3: Shift $e$ right by 1 bit;
    S2.4: If $e > 0$, compute $S \leftarrow S^2$ % $N$;
    S2.5: If $e > 0$, go to S2.2.
S3: If $h = h'$, return ‘true’; or else return ‘false’ to the main procedure.
```

Notice that the above algorithm is called by the module in section 3.2.4.

5 Effectiveness and Advantage of the System

If a symmetric encryption scheme is adopted, the symmetric key appears in both a signature process and a verification process, and is disclosed potentially at each work stage. It is very difficult for the exit-reenter administration of the country to supervise and protect the symmetric key effectually. So public key cryptosystem is more secure than symmetric key cryptosystems in our system. Although the execution speed of a public key cryptosystem is comparatively slow with a symmetric key cryptosystem, if we select a suitable public key cryptosystem and its parameters, the execution speed of the passport anti-forgery system will possibly satisfy the practicable requirement of users [8][9].

A public key cryptosystem has the following advantages:

(1) The signing key and verifying key are two different keys, it is convenient to distribute and management them. The security of the system is upgraded to the greatest degree.

(2) The public key is got from the Internet, and irrelevant to an IC passport read-write machine. Even if the public key is changed according to the private key by the administration, all IC passport read-write machines do not need to be changed or maintained.

In addition, the existing physical and optical anti-forgery manners can be imitated by attackers, and thus, existing passports can be counterfeited. If we select a fit public key scheme for the passport anti-forgery system, the security of the system will be warranted. This effectiveness is incomparable with that of the physical and optical anti-forgery manners. Furthermore, the passport anti-forgery system can counteract the potential threat from enhancement of computer speed by setting greater parameter values renewedly.

Currently, IC chip techniques are pretty mature. The chips micron-scaled have already come into a product line, and the chips nanometer-scaled are being in the process of research and development. The speed of reading IC chips is very fast [10][11]. When customs and other inspect organizations work, first read the signing-code and the passport number in an IC chip embedded in a passport, then submit these data to an administrative web server to calculate, last return verification results to the clients. The time spent during the whole process is only a few of seconds. Compared to security, the effective advantage is very obvious.

If the administration adopts the IC passport and passport anti-forgery system, it will stamp out passport forgery completely and will keep down the social costs of issuing passports by the administration.

6 Conclusion

In this paper, we suggest a new anti-forgery method based on digital signature scheme, and design the passport anti-forgery system which contains the four modules.

The passport anti-forgery system binds passports with the issue authority tightly. An identity card and its number bind the holder of a passport with the passport. Ordinarily we believe that the photo of a person binds an identity card with its holder according to present regulations although this binding is relaxed.

Of course, the biologic characteristic of a person may be employed to bind the person with a passport tightly.

If the country PSA combines biological recognition techniques such as fingerprint recognition, iris recognition, and facial recognition with digital signature techniques to carry out double anti-forgery, it will implement binding of a passport and his holder effectively, which not only can stop passport forgery but also stamp out the case of using other people's passports.

Acknowledgment

The authors would like to thank the Academician Changxiang Shen, Professors Shuwang Lü, Jian Li, and Yongquan Cai for their important advice, suggestions and helps.

References

1. Diffie, W., Hellman, M.E.: New Directions in Cryptography. IEEE Transactions on Information Theory 22(6), 644–654 (1976)
2. Schneier, B.: Applied Cryptography: Protocols, algorithms, and source code in C, 2nd edn. John Wiley & Sons, New York (1996)
3. Menezes, A., van Oorschot, P., Vanstone, S.: Handbook of Applied Cryptography. CRC Press, London (1997)
4. Rivest, R.L., Shamir, A., Adleman, L.M.: A Method for Obtaining Digital Signatures and Public-key Cryptosystems. Communications of the ACM 21(2), 12–126 (1978)
5. Yan, S.Y.: Number Theory for Computing, 2nd edn. Springer, New York (2002)
6. Su, S., Yang, B.: The REESSE Unified Recursive Algorithm for Solving Three Computational Problems. Wuhan University Journal of Natural Sciences 12(1), 172–176 (2007)
7. Stallings, W.: Cryptography and Network Security: Principles and Practice, 2nd edn. Prentice-Hall, New Jersey (1999)
8. Goldreich, O.: Foundations of Cryptography: Basic Tools. Cambridge University Press, Cambridge (2001)
9. Hemspaandra, L.A., Ogihara, M.: The Complexity Theory Companion. Springer, Heidelberg (2002)
10. Rabaey, J.M., Chandrakasan, A., Nikolic, B.: Digital Integrated Circuits, 2nd edn. Prentice-Hall, New Jersey (2002)
11. Martin, K.: Digital Integrated Circuit Design. Oxford University Press, Oxford (1999)

A Chronological Evaluation of Unknown Malcode Detection

Robert Moskovitch, Clint Feher, and Yuval Elovici

Deutsche Telekom Laboratories at Ben Gurion University
Ben Gurion Univsersity of the negev, Beer Sheva 84105, Israel
{robertmo,clint,elovici}@bgu.ac.il

Abstract. Signature-based anti-viruses are very accurate, but are limited in detecting new malicious code. Dozens of new malicious codes are created every day, and the rate is expected to increase in coming years. To extend the generalization to detect unknown malicious code, heuristic methods are used; however, these are not successful enough. Recently, classification algorithms were used successfully for the detection of unknown malicious code. In this paper we describe the methodology of detection of malicious code based on static analysis and a chronological evaluation, in which a classifier is trained on files till year k and tested on the following years. The evaluation was performed in two setups, in which the percentage of the malicious files in the training set was 50% and 16%. Using 16% malicious files in the training set for some classifiers showed a trend, in which the performance improves as the training set is more updated.

Keywords: Unknown Malicious File Detection, Classification.

1 Introduction

The term malicious code (malcode) commonly refers to pieces of code, not necessarily executable files, which are intended to harm, generally or in particular, the specific owner of the host. Malcodes are classified, based mainly on their transport mechanism, into five main categories: worms, viruses, Trojans, and a new group that is becoming more common, which comprises remote access Trojans and backdoors. The recent growth in high-speed internet connections has led to an increase in the creation of new malicious codes for various purposes, based on economic, political, criminal or terrorist motives (among others). A recent survey by McAfee indicates that about 4% of search results from the major search engines on the web contain malicious code. Additionally, Shin et al. [12] found that above 15% of the files in the KaZaA network contained malicious code. Thus, we assume that the proportion of malicious files in real life is about or less than 10%, but we also consider other options.

Current anti-virus technology is primarily based on signature-based methods, which rely on the identification of unique strings in the binary code, while being very precise, are useless against unknown malicious code. The second approach involves heuristic-based methods, which are based on rules defined by experts,

H. Chen et al. (Eds.): PAISI 2009, LNCS 5477, pp. 112–117, 2009.
© Springer-Verlag Berlin Heidelberg 2009

which define a malicious behavior, or a benign behavior, in order to enable the detection of unknown malcodes [4]. The generalization of the detection methods, so that unknown malcodes can be detected, is therefore crucial. Recently, classification algorithms were employed to automate and extend the idea of heuristic-based methods. As we will describe in more detail shortly, the binary code of a file is represented by n-grams, and classifiers are applied to learn patterns in the code and classify large amounts of data. A classifier is a rule set which is learnt from a given training-set, including examples of classes, both malicious and benign files in our case.

Over the past five years, several studies have investigated the option of detecting unknown malcode based on its binary code. Schultz et al. [11] were the first to introduce the idea of applying machine learning (ML) methods for the detection of different malcodes based on their respective binary codes. This study found that all the ML methods were more accurate than the signature-based algorithm. The ML methods were more than twice as accurate, with the out-performing method being Naïve Bayes, using strings, or Multi-Naïve Bayes using byte sequences. Abou-Assaleh et al. [1] introduced a framework that used the common n-gram (CNG) method and the k nearest neighbor (KNN) classifier for the detection of malcodes. The best results were achieved using 3-6 n-grams and a profile of 500-5000 features.

Kolter and Maloof [6] presented a collection that included 1971 benign and 1651 malicious executables files. N-grams were extracted and 500 were selected using the information gain measure [8]. The authors indicated that the results of their n-gram study were better than those presented by Schultz and Eskin [11]. Recently, Kolter and Maloof [7] reported an extension of their work, in which they classified malcodes into families (classes) based on the functions in their respective payloads.

Henchiri and Japkowicz [5] presented a hierarchical feature selection approach which makes possible the selection of n-gram features that appear at rates above a specified threshold in a specific virus family, as well as in more than a minimal amount of virus classes (families). Moskovitch et al [9], who are the authors of this study, presented a test collection consisting of more than 30,000 executable files, which is the largest known to us. A wide evaluation consisting on five types of classifiers, focused on the imbalance problem in real life conditions, in which the percentage of malicious files is less than 10%, based on recent surveys. After evaluating the classifiers on varying percentages of malicious files in the training set and test sets, it was shown to achieve the optimal results when having similar proportions in the training set as expected in the test set.

In this paper we investigate the need in updating the training set, through a rigorous chronological evaluation, in which we examine the influence of the updates of the training set on the detection accuracy. We start with a survey of previous relevant studies. We describe the methods we used to represent the executable files. We present our approach of detecting new malcodes and perform a rigorous evaluation. Finally, we present our results and discuss them.

2 Methods

2.1 Data Set Creation

We created a data set of malicious and benign executables for the Windows operating system. After removing obfuscated and compressed files, we had 7688 malicious files, which were acquired from the VX Heaven website. The benign files set contained 22,735, including executable and DLL files, were gathered from machines running Windows XP operating system on our campus. The Kaspersky anti-virus program was used to verify that these files indeed contain malicious code, or don't for the benign files.

2.2 Data Preparation and Feature Selection

We parsed the files using several *n-gram* lengths moving windows, denoted by *n*. Vocabularies of 16,777,216, 1,084,793,035, 1,575,804,954 and 1,936,342,220, for 3-gram, 4-gram, 5-gram and 6-gram respectively were extracted. Later each n-gram term was represented using its Term Frequency (TF), which is the number of its appearances in the file, divided by the term with the maximal appearances. Thus, each term was represented by a value in the range [0,1]. To reduce the size of the vocabularies we first extracted the top features based on the Document Frequency (DF) measure. We selected the top 5,500 features which appear in most of the files, (those with high DF scores). Later, three feature selection methods: *Gain Ratio (GR)* [8] and *Fisher Score (FS)* [3], were applied to each of these two sets. We selected the top 50, 100, 200 and 300 features based on each of the feature selection techniques. More details on this procedure and results can be found in [9], in which we found that the optimal settings were top 300 features selected by Fisher score where each feature is 5-gram represented by TF from the top 5500 features, which we used in this study.

3 Evaluation

We employed four commonly used classification algorithms: *Artificial Neural Networks* (ANN) [Bishop, 1995], *Decision Trees* (DT) [10], *Naïve Bayes* (NB) [2]. We used the Weka [13] implementation for the Decision Trees and the Naïve Bayes and the ANN tool box in Matlab.

To evaluate the importance of and need for updating the training set, we divided the entire test collection into the years from 2000 to 2007, in which the files were created. Thus, we had 6 training sets, in which we had samples from year 2000 till year 2006. Each training set was evaluated separately on each following year from 200k+1 till 2007. Obviously the files in the test were not presented in the training set. We present two experiments which vary in the Malicious Files Percentage (MFP) in the training set, having 50% which is commonly used and 16% which is expected to maximize the performance, which was the same in the test set (16%) to reflect real life conditions (in both cases).

Decision Trees

Figure 1 presents the results of the chronological evaluation, for the 50% MFP in the training set. Training on 2000 results below 0.9 accuracy, while training on the next years improved the accuracy. However, generally a significant decrease in performance was seen when testing on 2007.

Figure 2 presents the results of the chronological evaluation, in which the MFP in the training set was 16%. In Figure 2 we see generally a higher performance than in figure 1. 2004 introduced a significant challenge for the training sets of until 2000 and 2001. In this set of results there is a clear trend which shows that the more the training set is updated the higher the accuracy in the following years, and even when testing on 2007 the accuracy was above 0.9, when trained on till 2004, 2005 and 2006.

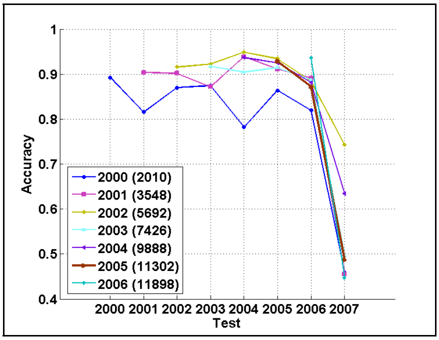


Fig. 1. For the 50% MFP training set the more updated the higher the accuracy. A significant decrease is seen in 2007, while training on 2006 outperforms.

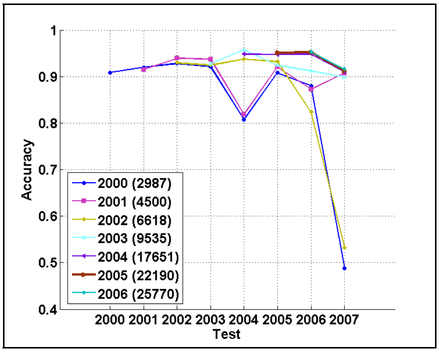


Fig. 2. For the 16% MFP training set, the more updated the higher the accuracy. Testing on year 2004 presented a challenge for the 2000 and 2001 training sets.

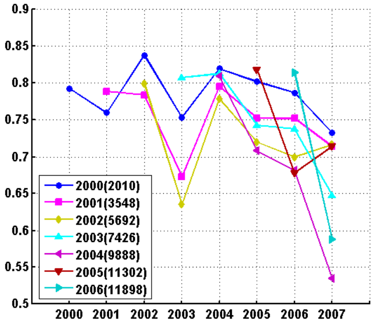


Fig. 3. For the 50% MFP a decrease is in 2003 and generally the results are low

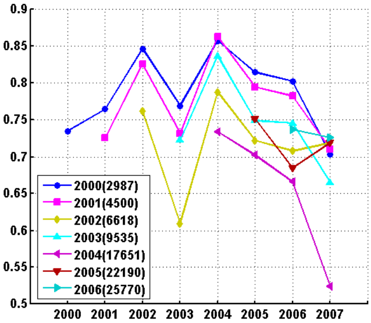


Fig. 4. For the 16% MFP there is a slight improvement in comparison to the 50% MFP

Naïve Bayes

Figure 3 and 4 present the results of the chronological evaluation using the Naïve Bayes classifier, for the 50% MFP (fig 3) and 16% MFP (fig 4) training sets. In 2003 there is a significant drop in the accuracy in both MFPs, which appears only with this classifier. The results in general are lower than the other classifiers. The results with the 16% MFP are slightly better. However, in both figures the accuracy drops for the last years, especially for 2007.

Artificial Neural Networks

Figures 5 and 6 present the chronological results for the ANN classifier. The results seem better than the Naïve Bayes, especially for the 16% MFP results. In Figure 6 the results seem to perform very well along most of the years, out of a significant drop for the training set from 2005, especially with the test set of 2007.

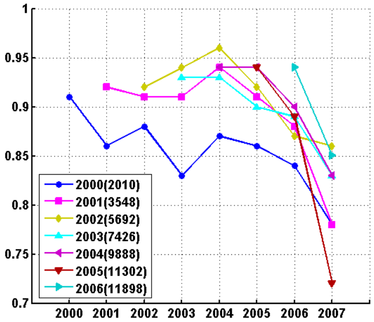


Fig. 5. For the 50% MFP training set, training on 2000 performed very low, unlike the others

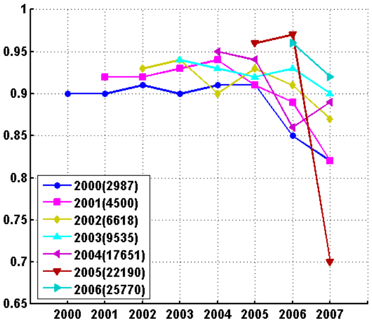


Fig. 6. For the 16% MFP there is a significant improvement in comparison to the 50% MFP

4 Discussion and Conclusions

We presented the problem of unknown malicious code detection using classification algorithms. We described the use of n-grams for the representation where feature selection methods are used to reduce the amount of features. We presented the creation of our test collection, which is 10 times larger than any previously presented. In a previous study [9], we investigated the aspects of the percentage of malicious files in the training set to maximize the accuracy in real life conditions.

In this study we referred to the question of the importance of updating the training set with the new malicious codes in a yearly time granularity and whether it is important to keep samples of old files in the training set from few years ago. Our results indicate that when having 16% MFP in the training set which corresponds to the test set we achieve a higher level of accuracy, and also a relatively clear trend that as the training set is more updated the accuracy is higher. However, this varies according to the classifier and one should be aware of this influence in deployment, as sometimes it decreases the accuracy. Moreover, it seems to be better to have also files which are from several years earlier and to incrementally update the database.

References

[1] Abou-Assaleh, T., Cercone, N., Keselj, V., Sweidan, R.: N-gram Based Detection of New Malicious Code. In: Proceedings of the International Computer Software and Applications Conference (COMPSAC 2004) (2004)

[2] Domingos, P., Pazzani, M.: On the optimality of simple Bayesian classifier under zero-one loss. Machine Learning 29, 103–130 (1997)

[3] Golub, T., Slonim, D., Tamaya, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C., Lander, E.: Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science 286, 531–537 (1999)

[4] Gryaznov, D.: Scanners of the Year 2000: Heuristics. In: Proceedings of the 5th International Virus Bulletin (1999)

[5] Henchiri, O., Japkowicz, N.: A Feature Selection and Evaluation Scheme for Computer Virus Detection. In: Proceedings of ICDM 2006, Hong Kong, pp. 891–895 (2006)

[6] Kolter, J.Z., Maloof, M.A.: Learning to detect malicious executables in the wild. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 470–478. ACM Press, New York (2004)

[7] Kolter, J., Maloof, M.: Learning to Detect and Classify Malicious Executables in the Wild. Journal of Machine Learning Research 7, 2721–2744 (2006)

[8] Mitchell, T.: Machine Learning. McGraw-Hill, New York (1997)

[9] Moskovitch, R., Stopel, D., Feher, C., Nissim, N., Elovici, Y.: Unknown Malcode Detection via Text Categorization and the Imbalance Problem. In: IEEE Intelligence and Security Informatics (ISI 2008), Taiwan (2008)

[10] Quinlan, J.R.: C4.5: programs for machine learning. Morgan Kaufmann Publishers, Inc., San Francisco (1993)

[11] Schultz, M., Eskin, E., Zadok, E., Stolfo, S.: Data mining methods for detection of new malicious executables. In: Proceedings of the IEEE Symposium on Security and Privacy, pp. 178–184 (2001)

[12] Shin, S., Jung, J., Balakrishnan, H.: Malware Prevalence in the KaZaA File-Sharing Network. In: Internet Measurement Conference (IMC), Brazil (October 2006)

[13] Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann Publishers, Inc., San Francisco (2005)

Relation Discovery from Thai News Articles Using Association Rule Mining

Nichnan Kittiphattanabawon and Thanaruk Theeramunkong

Sirindhorn International Institute of Technology, Thammasat University, Thailand
knichnan@gmail.com, thanaruk@siit.tu.ac.th

Abstract. Association among news articles is useful information for us to track situation related to events, persons, organizations and other concerned issues as well as to detect inconsistency among news. In this paper, we propose an association-based approach towards mining relations in Thai news articles by exploiting coincident terms. This approach first transforms news documents into term-document representation, applies term weighting techniques and generates association by means of statistics. In the work, either unigram or bigram is used for term representation, term frequency, boolean frequency and their modification with inverse document frequency are alternatively applied for term weighting, and confidence or conviction is in turn selected for association measure. Due to this combination, sixteen possible methods are investigated using approximately 811 Thai news of three categories, i.e., politics, economics, and crime. The ranked relations obtained by each method are compared with evaluation done by human. As the result, the method using bigram, term frequency, and conviction achieves the best performance with a rank-order mismatch of 0.84% on the top-50 mined relations. For the top-300 mined relations, the method with bigram, term frequency with inverse document frequency and conviction performs the best with 6.98% rank-order mismatch.

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

1 Introduction

With rapid growth of the Internet, there have been several news providers publishing online newspapers. While it becomes easy for readers to access information about an event, a large pile of these online newspapers trigger a burden in tracking an event. To alleviate this difficulty, many news providers try to manually organize their news articles, under a set of criteria, into some kinds of relationship structures, such as grouping them by category, by recency, or by popularity. Recently several methods have been proposed to support automatic organization of news articles, such as topic identification [1], event extraction [2], news summarization [3], classification of news articles [4] and temporal mining in news [5]. However, most techniques applied in these previous works may not be adequate for the readers to track a specific event or a situation related

H. Chen et al. (Eds.): PAISI 2009, LNCS 5477, pp. 118–129, 2009.
© Springer-Verlag Berlin Heidelberg 2009

to a person and other concerned issues. Several examples of applications about news relations were demonstrated in various languages. An approach using the misclassification information was proposed to find relationships among the categories of English news [6]. In Chinese news, an event identification, a main storyline construction and a storyline-based summarization were proposed to relate happened events in a same topic [7]. In Thai language, even there have been several works towards extraction of information on online document, most of them still have limitation. An approach to Thai news for extracting the most relevant paragraphs from the original document to form a summary was given in [8]. Recently, an approach to mine relations in scientific research publications [9] was proposed by using association rule minings (ARM) with support-confidence framework by extending a concept of traditional ARM to mine frequent itemsets on the database with real-valued instead of only item existences. This method could discover a set of topically similar documents as high quality relations. In this work, a novel evaluation method is presented based on an evaluation by benchmark citation matrix, using ACM database[1]. Although there have been a number of works on mining relation on English texts, there has been no research work on mining relations from Thai news articles.

In this paper, we propose an association-based approach towards mining relations using ARM and present its application on Thai news articles. We compare our newly proposed support-conviction framework against the support-confidence framework previously proposed in [9]. Using evaluation results by human as the reference, we investigate several performance obtained from several combinations of term representations, term weightings and association measures when compared with the ranked relations. In the rest, Sect. 2 gives description of relations occurred among news. In Sect. 3, news relation discovery are described under the formation of association rules. Moreover, the factors in relation discovery are given. Section 4 displays our implementation and evaluation including Thai news collection and preprocessing. Construction of evaluation dataset and evaluation criteria are also explained. A number of experimental results and discussion are presented in Sect. 5. Finally, a conclusion and future works are made in Sect. 6.

2 Relations among News

News articles can be related to each other with several relation types. The connection among them are useful for us to gain more detailed information. In this work, the meaningful relations among news documents can be divided to three relation types, i.e., (1) “completely related”, (2) “somehow related” and (3) “unrelated”. For the relation of “somehow related”, three subtypes can be considered, i.e., “series”, “subnews” and “similar them e”. The first relation type, “completely related”, refers to a relation between two news that mentions an exactly same event but may be presented in different headline, different styles, or different publishers. They are always published in a same date or very close time.

[1] http://www.portal.acm.org

For “somehow related” relation, two news are related under the same event but not exactly the same since one news may be series, subnews, or similar theme with another news. Two news with the relation of ‘series” seem to have a content with consecutive story. If some content of one news is a part of another news, this relation is considered as “subnews”. For “similar theme” relation, two news are addressed on the same outstanding topic. For example, two news which are related by a theme of “Thai Elephants” show that they are related to each other by a topic of Thai elephants. However, their details are different, for example, one news may be about the conservation of Thai elephants while another news may show how Thai elephants contribute to a tour in Thailand. The publishing time of “somehow related” relation is in the same period. The last type is “unrelated”, it indicates that two news have nothing to be related. By considering a source of publishing, “completely related” relations come from difference sources because same publishers don’t publish the same news more than one time. For “somehow related” relation, two news can be published by both same publisher and different publisher since this relation type is not exactly the same news. For “unrelated” relation, it can be published in any publishers due to unrelated event.

3 News Relation Discovery Using Association Rule Mining

3.1 Generalized Association Rules

The association among items in data mining approach is the relation of a co-occurrence item as a rule of the general form “if antecedent, then consequent” generated by ARM [10]. ARM is a process consisting two steps. The first step called frequent itemset mining (FIM) is to find frequent patterns. Such patterns are used as input to the second step where the association rules are finally extracted. Concerning with this process, minimum support (MINSUP) is used in the first step in order to discover the itemsets of which their supports exceed at least this threshold, considered as frequent itemsets. Based on these frequent itemsets, minimum confidence (MINCONF) is used to generate and filter out the rules of which their confidences lower than this threshold, considered as association rules. In this work, we generalize the traditional support to calculate the support of itemsets not only concerning the term existence, but also concerning the term weight. To this end, assume that an itemset $X = \{x_1, x_2, ..., x_k\} \subset I$ with k items, so-called k-itemset, where $\{i_1, i_2, ..., i_m\}$ is a set of items I and $\{t_1, t_2, ..., t_n\}$ is a set of transactions identifiers T, the generalized support of X is defined as shown in (1) [9].

$$sup(X) = \frac{\sum_{b=1}^{n} min_{a=1}^{k} w(x_a, t_b)}{\sum_{b=1}^{n} max_{a=1}^{m} w(i_a, t_b)}, \tag{1}$$

where $w(i_a, t_b)$ represents a weight between an item i_a and a transaction t_b. A subset of I is called an itemset. For association rule discovery, the confidence can be used for measuring association strength using the generalized support as $conf(X \rightarrow Y) = sup(X \cup Y)/sup(X)$. Here, X → Y is the association rules, specifying that item Y is the co-occurrences of item X. Assume that an items $X = \{x_1, x_2, ..., x_k\} \subset I$ with k items and an items $Y = \{y_1, y_2, ..., y_l\} \subset I$ with l items. To this end, the generalized confidence of X → Y is defined as (2).

$$conf(X \rightarrow Y) = \frac{\sum_{b=1}^{n} min_{a=1}^{k+l} w(z_a, t_b)}{\sum_{b=1}^{n} min_{a=1}^{k} w(x_a, t_b)}, \tag{2}$$

where $Z = \{z_1, z_2, ..., z_{k+l}\} \subset I$ with $k+l$ items since they are the co-occurrences between k-itemset of X and l-itemset of Y.

As an alternative to confidence, a relatively new measure called conviction is applied in our work. The definition of conviction, as acknowledged by ARM approach, is defined as $(1 - sup(Y))/(1 - conf(X \rightarrow Y))$. To this end, by substituting $sup(Y)$ and $conf(X \rightarrow Y)$ by (1) and (2) consequently, the definition of generalized conviction can be explored in (3). Note that all variables are same as the variables of (1) and (2).

$$conv(X \rightarrow Y) = \frac{1 - \frac{\sum_{b=1}^{n} min_{a=1}^{l} w(y_a, t_b)}{\sum_{b=1}^{n} max_{a=1}^{m} w(i_a, t_b)}}{1 - \frac{\sum_{b=1}^{n} min_{a=1}^{k+l} w(z_a, t_b)}{\sum_{b=1}^{n} min_{a=1}^{k} w(x_a, t_b)}}, \tag{3}$$

Due to the equation of confidence and conviction, it can be observed that conf($X \rightarrow Y$) is not equal to conf($Y \rightarrow X$) and conv($X \rightarrow Y$) is also different from conv($Y \rightarrow X$) because they are bidirection functions. In this work, for simplicity, we ignore the direction of the rule. Any X and Y will have a strong association if both $X \rightarrow Y$ and $Y \rightarrow X$ have a large confidence or large conviction. To express this constraint, the following equations are occupied as the association level between X and Y.

$$conf(X, Y) = min(conf(X \rightarrow Y), conf(Y \rightarrow X)) \tag{4}$$

$$conv(X, Y) = min(conv(X \rightarrow Y), conv(Y \rightarrow X)) \tag{5}$$

This constraint is established according to the following reason. The small value can conform to the judgment of human when they have overlooked the direction of news relation. For example, if $news1 \rightarrow news2$ has 95% of confidence while $news2 \rightarrow news1$ has 50% of confidence, due to size of $news1$ and $news2$ is unequal, human preferably agrees with "somehow related" type more than "completely related" type (see Sect. 2) since news relation is not related by the whole content among them.

3.2 Term-Document Representation for News Documents

In this paper, we apply association rule mining to find relation among news documents. Our approach first transforms news documents into term-document

Table 1. The weight derived from BF (left), TF (middle), IDF (right) weightings

	d_1	d_2	d_3	d_4	d_5	d_1	d_2	d_3	d_4	d_5	DF	IDF
$term_1$	1	0	0	0	0	2	0	0	0	0	1	log(5/1)
$term_2$	1	1	0	0	0	4	1	0	0	0	2	log(5/2)
$term_3$	1	1	1	1	1	6	2	2	1	4	5	log(5/5)

representation since the goal is to investigate relations among news documents, hence term-document orientation is used instead of transaction-term orientation [9]. For transaction-term orientation, the discovered frequent itemset is a set of greatly co-occurred items in the transaction while a set of documents which share a high number of terms is called a frequent docset in term-document orientation [9]. In this orientation, a transaction corresponds to a term while an item is mapped to a document. Such a document-term database is an input for FIM.

3.3 Three Factors in Relation Discovery

This approach considers three factors for discovering the news relations, i.e., (1) term representation, (2) term weighting and (3) association measure, to compare the quality of mined association rules. After transforming news documents into term-document format as described in previous section, either unigram or bigram is used for term representation, boolean frequency, term frequency and their modification with inverse document frequency are alternatively applied for term weighting, and confidence or conviction is in turn selected for association measure. As all possible combinations of these three factors, sixteen methods can be considered.

Term Representations. As mentioned above, the relation among news as an association rule will arise when it has a set of highly co-occurred terms. After segmenting which shall be elaborated1 in Sect. 4.1, the words are defined as unique terms separated by spaces. Then stopwords are applied to eliminate useless words which may frequently appear in documents. Finally, a list of remaining terms is represented in news document. However a number of text processing approaches were showed that individual words alone is not good enough for representing the content of a document. N-gram is selected to be term representation to lighten this case. According to this reason, unigram (UG), a single isolated term in a document, may not good enough to represent semantics of the text due to its ambiguity. Consequently, bigram (BG), two neighboring terms in the document, will be applied to handle compound words and reduces semantic ambiguity of words. Note that, besides unigram and bigram, a higher n-gram can be used for representing the terms, but this work are preliminary addressed only unigram and bigram. Also, because of the limitation of time, which is typically harmful from exponential growth of the number of terms, in mining process.

Table 2. The characteristics of news collection

Sources	Politics	Economics	Crime
Dailynews	92	107	115
Komchadluek	86	41	79
Manager online	88	102	101
Total	266	250	295

Term Weightings. As term weighting, boolean frequency (BF) weighting, term frequency (TF) weighting and their combinations with inverse document frequency (IDF) weighting are explored into the consideration. Totally four conditions are considered as a weight for the experiment, i.e., BF, BFIDF, TF and TFIDF. Hence they will be calculated by substituting to the weight in (1) as stated in Sect. 3.1. Table 1 shows the weight derived from BF, TF and IDF weightings. As BF is a boolean frequency, it simply indicates the term existence in a document (1: presence, 0: absence) while TF indicates the number of terms found in the document. For IDF, it is a measure of how rare a term is in a collection of documents, calculated by the total number of documents in the collection (N) divided by the number of documents containing the term (DF). IDF shows that very common terms, commonly referred to as stopwords, will have very low IDF and are therefore often excluded from considered results. Alternatively, the frequencies can be transformed in the standard ways by using standard logarithm. To this end, 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 as a means of frequency which is modulated by a factor that depends on how commonly the term is used in other documents.

Association Measures. To measure the quality and utility of the discovered association rules, quantitative measures have been studied. In the past, the support-confidence framework was well-known for ARM which can point out the frequent patterns and high confident association rules. In this work, the support-conviction framework is proposed to improve the relevant of the news relations since it was proposed that its result tend to give better strength rules than confidence [11] [12]. Thus, either confidence (CONF) or conviction (CONV) is compared as an association measure in our experiment.

4 Implementation and Evaluation Dataset

4.1 Thai News Collection and Preprocessing

To evaluate our approach, we have collected 811 Thai news articles from three news online sources (Dailynews, Komchadluek and Manager online)[2] during

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

August 14-31, 2007. Three categories of news, politics, economics and crime, are considered. Table 2 shows the characteristics of news collection.

As our approach concerns over news articles in Thai. Unlike English language which has explicit word boundary markers, Thai language has no any word and sentence boundaries. Thai words may be written with different boundaries depend on the individual judgment. The first task of preprocessing module, therefore, is term segmentation process to separate the stream of characters into individual terms. To this end, a word separator program, namely CTTEX[3] is used for segmenting words in news documents, combining with a suitable dictionary. Such a dictionary includes 74,751 terms from a general dictionary and a named entity dictionary. Moreover, a stopword list of 1,809 terms is used for eliminating trivial terms. Finally, the news documents are then sent to the mining process.

4.2 Construction of the Evaluation Dataset

To evaluate the quality of proposed methods in discovering of news relation, a dataset for evaluation is manually constructed by selecting a number of news relations from our news collection and testing human subjects to specify their relation types due to no standard dataset available as a benchmark. Three evaluators are chosen to evaluate a set of selected news relations. Due to labor intensive task, thus the selected relations for evaluation are ones in top-300 output by each of the sixteen methods mentioned in Sect. 3.3. In totals, approximately 1,200 news relations are considered and used as the evaluation dataset for our experiments. Three evaluators are asked to give an evaluation result for each relation by a three level score, i.e., 1 point for "completely related" relation, 0.5 point for "somehow related" relation, and 0 point for "unrelated" relation. Three evaluation results from the three evaluators are voted for the final decision. However, sometimes the three score may not be consensus, for example, the first evaluator gives 1 point but the second and the third evaluator give 0.5 and 0 point respectively. To solve this situation, the iteration process is done by asking the evaluators to confirm their decision.

4.3 Evaluation Criterion

To compare the qualitative results between our methods using ARM and the results from an evaluation of human. Rank-order mismatch (ROM) shown in (6) is explored by the calculation of dividing the mismatch score with the mismatch score of the worst case which all news relations in one method are arranged in the reverse order compared to the other method. Note that ROM is equal to 0 when all news relations are investigated corresponding to human suggestion.

$$ROM(X, Y) = \frac{2 \times M(X, Y)}{N(N-1)} \tag{6}$$

[3] arthit.googlepages.com/bazaar

$$M(X,Y) = \sum_{i=1}^{N} \sum_{j=i+1}^{N} |\delta(r_X(i), r_X(j)) - \delta(r_Y(i), r_Y(j))| \tag{7}$$

For (7), $M(X,Y)$ is a mismatch score which means the number of rank mismatches between two methods, say X and Y, which rank a set of N objects. Here, $r_X(k)$ and $r_Y(k)$ are the respective rank of the k-th objects based on method X and Y repectively. $\delta(a,b)$ is a mismatch function which returns 1 when a less than b, otherwise 0. This method refers to paired-wise comparisons techniques [13] which was used for counting the mismatch between their ranking.

5 Experimental Settings and Results

5.1 Experimental Setting

In this work, 811 Thai news articles of three categories, politics, economics, and crime, are gathered from a collection of Thai online newspapers as the experimental dataset. As stated in Sect. 4.2, the evaluation dataset was constructed by three evaluators with an iteration process for conflict resolutions. As mentioned in Sect. 3.3, three factors are examined to compare the quality of mined association rules. Totally we have 16 methods on consideration. They are the combinations of two term representations {UG, BG}, four term weightings {BF, BFIDF, TF, TFIDF} and two association measure {CONF, CONV}. For those combinations, their minimum support settings may not be the same since they occupy different support specifications. The minimum support of each combination is set to generate top-K frequent itemsets. In this work, K is set to 300. MINCONF and MINCONV are set by their smallest values as possible in order to reach our consideration on top-300 relations. Eventually, the ranked news relations are compared with the evaluation results by using ROM. In the preliminary stage, we focus on only an association rule with a single antecedent and a single consequent, i.e., the rule which indicates the relation of two news.

5.2 Experimental Results

In this section, a number of experimental results are given in order to evaluate our proposed method To analyze the quality of top-K news relations when K is varied from 50 to 300.

News Relations in the Evaluation Dataset. As described in Sect. 2 about the types of relations, Table 3 shows sources and timelines. The number of "completely related", somehow related", and "unrelated" relations are 65, 571 and 496 respectively. For "completely related" relation, many of them come from different sources of publishing, i.e., 45, and most of the publishing time of two news is a same date, i.e., 0. For "somehow related" relation, "series" and "similar theme" are discovered from both the same publisher and different publisher with almost the same figure, i.e., 95 vs. 104 and 149 vs. 148. For "subnews", they are

Table 3. News relations in the evaluation dataset: their sources and time intervals based on evaluation by human

News Relation Types	Total	Sources (relations)		Time Interval (days)		
	Number of Relations	Same	Different	Min	Max	Mode
Completely Related	65	20	45	0	7	0
Somehow Related	571	263	308	0	17	1
- series		95	104	0	16	1
- subnews		19	56	0	4	0
- similar theme		149	148	0	17	3
Unrelated	496	245	251	0	15	2

published in more different publisher than same publisher, i.e., 56 vs. 19. Most relations of "somehow related" are emerged in very close date, i.e., within 3 days. However, the interval of publishing time of "series" relations is closer than another two relations, i.e., 4 days. For "unrelated" relation, two news can be published in same publisher as well as different publisher and any time since they don't be related to each other. Note that the maximum interval of publishing time is 17 days due to our news collection.

Evaluation on Three Factors

Unigram vs. Bigram. The results in Fig. 1 show that bigram yields more number of positive values than the negative values for varied K's. This result indicates that bigram gives lower *ROM* than unigram. In other words, bigram gives that matching human results.

Conviction vs. Confidence. Similar to previous experiment except confidence and conviction are taken into account instead of unigram and bigram. Because bigram outperforms unigram in almost cases, comparing the results of confidence and conviction in the case of bigram, Fig. 2 shows that conviction performs better than confidence.

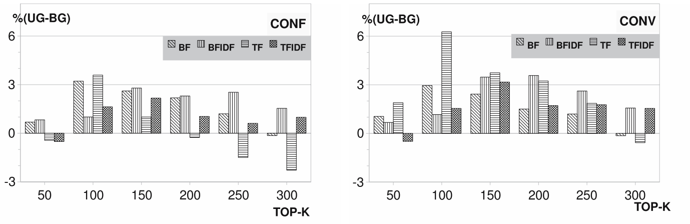


Fig. 1. Percentages of *ROM* difference between UG and BG (%UG–%BG) in the cases of CONF (left) and CONV (right)

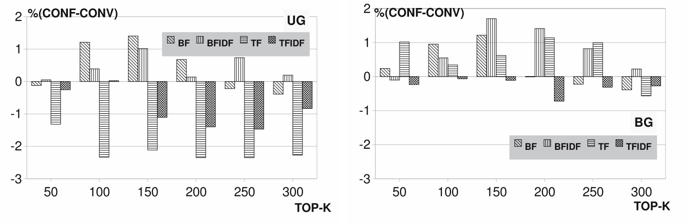


Fig. 2. Percentages of *ROM* difference between CONF and CONV (%CONF–%CONV) in the cases of UG (left) and BG (right)

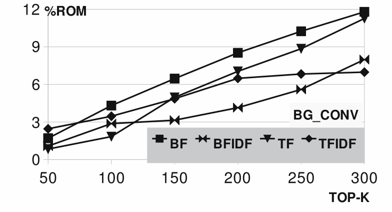


Fig. 3. Percentages of *ROM* between BF, BFIDF, TF and TFIDF in the cases of BG-CONV

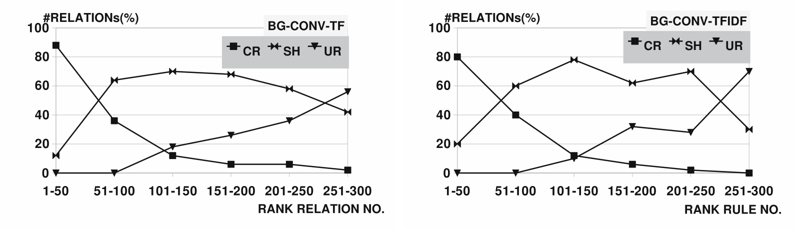


Fig. 4. Percentages of the number of relations group by relation types

BF vs. TF vs. BFIDF vs. TFIDF. Due to the potential factors from previous results, i.e., bigram and conviction, Fig. 3 where the lines show the percentage of *ROM* based on them displays that the method using bigram, conviction and term frequency achieves the best performance with *ROM* of 0.84% on top-50 mined relations. For top-300, the method with bigram, conviction and term frequency with inverse document frequency performs the best with 6.98% *ROM*.

Analysis by Relation Types. For more details, the news relations based on relation type, "completely related" (CR), "somehow related" (SR) and "unrelated" (UR) are made to investigate how good news relations perform in various top-K ranks. Note that types of "somehow related" are left since they are planned on a future work. Fig. 4 shows the results of two cases from the potential methods of previous experiments using bigram and conviction combined by either term frequency or term frequency with inverse document frequency. It can be observed from every case that the higher rank it provides, the smaller number of relations on CR will be, while, for UR, the higher rank it is, the larger number of relations will be produced. For SH, the number of relations increase when locate in the middle rank and decrease when present in the higher rank.

5.3 Discussion

By experiments, Table 3 showed the number of relation types that agree with their characteristics as described in Sect.2. Moreover, we have figured out the statistics of the number of relations occurred in each relation types as well as publishing time of two news. However, larger news collection should be considered to confirm these results. The rest of experimental results presented the performance of our methods. It was observed that bigram outperforms unigram in almost cases (term weightings and association measures). Comparing the results of conviction and confidence in the case of bigram, we found out that conviction is superior to confidence. For term weightings in the case of bigram and conviction, term frequency weighting and term frequency with inverse document frequency weighting are suitable. Term frequency weighting performs well in the lower rank while term frequency with inverse document frequency weighting works well in the higher rank. More details were explored to study the distribution of "completely related", "somehow related" and "unrelated" in each rank period. We found out that our methods with bigram, conviction and term frequency (BG-CONV-TF) and that with bigram, conviction and term frequency with inverse document frequency (BG-CONV-TFIDF) output top-100 relations that were judged to either "completely related" or "somehow related" (without "unrelated" cases). With higher ranks, say 251-300, the extracted relations were not suitable since 88% and 80% of relations were just to be "unrelated" for these two cases, respectively. It can be concluded that conviction obtained better potential relations.

To improve quality of relations, two possible approaches can be considered. The first approach is to improve the quality of dictionary and the named entity set. The second is to exploit a resource of synonym in order to mapping two or more words with the same meaning. However, there are also errors triggered by human in the evaluation dataset. Although, our iteration process can alleviate them, but it may not be completely correct due to individual judgment. To solve this problem, we may need to increase the number of evaluators. Moreover, it is also worth investigating the performance of the top-K in a higher K.

6 Conclusion and Future Works

This paper proposed an association-based method to find relationships among Thai news documents. A number of term representations, term weightings and association measures were explored. The experimental results showed that the method with bigram representation, conviction measure and term frequency weighting is the best combination to achieve finding semantic relation with high quality in lower ranks (0.84 $\%ROM$) while the same method but adding inverse document frequency into its term frequency performs better in higher ranks (6.98 $\%ROM$). The experimental results imply that a support-conviction framework outperforms the support-confidence framework. As a future work, deeper analysis on the characteristics of relations corresponding to directions of rule will be explored. We will investigate incremental mining when news articles are gradually input into the mining process.

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. Fukumoto, F., Suzuki, Y.: Topic tracking based on bilingual comparable corpora and semisupervised clustering. ACM TALIP 6(3) (2007)
2. Kumaran, G., Allan, J.: Using names and topics for new event detection. In: Proc. of HLT-EMNLP, pp. 121–128 (2005)
3. Kuo, J., Chen, H.: Multidocument summary generation: Using informative and event words. ACM TALIP 7(1), 1–23 (2008)
4. Montalvo, S., Martnez, R., Casillas, A., Fresno, V.: Multilingual news clustering: Feature translation vs. identification of cognate named entities. Pattern Recognition Letters 28(16), 2305–2311 (2007)
5. Kalczynski, P.J., Chou, A.: Temporal document retrieval model for business news archives. IPM 41(3), 635–650 (2005)
6. Mengle, S., Goharian, N., Platt, A.: Discovering relationships among categories using misclassification information. In: Proc. of ACM SAC, pp. 932–937 (2008)
7. Lin, F., Liang, C.: Storyline-based summarization for news topic retrospection. Decision Support Systems 45(3), 473–490 (2008)
8. Jaruskulchai, C., Kruengkrai, C.: A practical text summarizer by paragraph extraction for thai. In: Proc. of IRAL, pp. 9–16 (2003)
9. Sriphaew, K., Theeramunkong, T.: Quality evaluation for document relation discovery using citation information. IEICE Trans. Inf. Syst. E90-D(8), 1225–1234 (2007)
10. Larose, D.T.: Discovering Knowledge in Data: an Introduction to Data Mining. John Wiley and Sons, Inc., Hoboken (2005)
11. Brin, S., Motwani, R., Ullman, J., Tsur, S.: Dynamic itemset counting and implication rules for market basket data. In: Proc. of ACM SIGMOD, pp. 255–264 (1997)
12. Jorge, A., Azevedo, P.: An experiment with association rules and classification: Post-bagging and conviction. In: Hoffmann, A., Motoda, H., Scheffer, T. (eds.) DS 2005. LNCS, vol. 3735, pp. 137–149. Springer, Heidelberg (2005)
13. David, H.: The Method of Paired Comparisons. Oxford University Press, Oxford (1988)

Discovering Compatible Top-K Theme Patterns from Text Based on Users' Preferences

Yongxin Tong[1], Shilong Ma[1], Dan Yu[1], Yuanyuan Zhang[2], Li Zhao[1], and Ke Xu[1]

[1] State Key Lab. of Software Development Environment
Beihang University, Beijing 100191, China
{yxtong,slma,yudan,lzh,kexu}@nlsde.buaa.edu.cn
[2] China Academy of Telecommunication Technology, Beijing 100191, China
yyzhang@catt.ac.cn

Abstract. Discovering a representative set of theme patterns from a large amount of text for interpreting their meaning has always been concerned by researches of both data mining and information retrieval. Recent studies of theme pattern mining have paid close attention to the problem of discovering a set of compatible top-k theme patterns with both high-interestingness and low-redundancy. Since different users have different preferences on interestingness and redundancy, how to measure the attributes of the users' preferences, and thereby to discover "preferred compatible top-k theme patterns" (PCTTP) is urgent in the field of text mining. In this paper, a novel strategy of discovering PCTTP based on users' preferences in text mining is proposed. Firstly, an evaluation function of the preferred compatibility between every two theme patterns is presented. Then the preferred compatibilities are archived into a data structure called theme compatibility graph, and a problem called MWSP based on the compatibility graph is proposed to formulate the problem how to discover the PCTTP. Secondly, since MWSP is proved to be a NP-Hard problem, a greedy algorithm, DPCTG, is designed to approximate the optimal solution of MWSP. Thirdly, a quality evaluation model is introduced to measure the compatibility of discovering theme patterns. Empirical studies indicate that a high quality set of PCTTP on four different sub text sets can be obtained from DBLP.

1 Introduction

Usually a large amount of text information is encountered in the application of text processing. Thus, how to discover a representative set of theme patterns automatically from a large amount of text to interpret their meaning is still concerned by researches of both data mining and information retrieval. Since recent researches in the domain of data mining show that a set of closed frequent patterns have many overlaps and redundant patterns, it is difficult for users to understand the meaning of huge number of patterns directly. The traditional frequent theme pattern mining from texts also confronts the same difficulties. Therefore, we should discover a small scale of theme patterns, with high-interestingness and low-redundancy simultaneously, from much text information. However, different users have different preferences on

H. Chen et al. (Eds.): PAISI 2009, LNCS 5477, pp. 130–142, 2009.
© Springer-Verlag Berlin Heidelberg 2009

interestingness and redundancy, so it is hard to use one unified criterion to find patterns with both high-interestingness and low-redundancy. We give an example to make a further explanation that discovering "preferred compatible top-k theme patterns" (*PCTTP* for short) from a large amount of text is the urgent and interesting challenge in text mining post-processing.

An interesting example is to discover the hot research topics by analyzing the titles of papers in DBLP database (The DBLP is a well known and very popular search engine which is used to index papers in the computer science area.) Given the periods and the ranges of some conferences and journals, we can discover PPCTP with high-interestingness and few-overlap research topics based on users' preferences. For example, a researcher, especially a beginner, often wants to know what are the hottest research topics lately, that is, likely to pay more attention to the level of popularity, even there may be some overlap of terms in these topics. However, other researchers want to obtain many different topics even not the hottest ones, because the overlap among these topics is very low. Hence, discovering PCTTP from a bulk of titles of papers can precisely satisfy these requirements from different users.

It is important to discover PCTTP from the above example, however, mining compatible top-k theme patterns based on users' preferences has not been well addressed. Although recent studies of text pattern mining have mined redundancy-aware top-k patterns [8] and excluded the redundancy among theme patterns [1, 7, 9], they could not mine a set of patterns according to users' preferences of interestingness and redundancy. A detailed discussion of the related work is given in Section 6.

The rest of the paper is organized as follows. Section 2 introduces the problem formulation, including the evaluation function of preferred compatibility of every theme pattern, theme compatibility graph and MWSP (Maximal Weight Sum Problem). The DPCTG (Discovering Preferred Compatible Theme based Greedy) algorithm is proposed to approximate the MWSP in Section 3. A quality evaluation function is introduced in Section 4. The experimental results, the related work and the conclusion are given in Section 5, 6 and 7 respectively.

2 Problem Formulation

In this section, we firstly introduce some basic concepts of interestingness and redundancy of theme pattern, then, define the problem of discovering PCTTP.

2.1 Preliminaries

Given a sequence database D , $D=\{s_1,s_2,...,s_n\}$ is a set of sequences. Each sequence is associated with an ID. The symbol $|D|$ represents the number of sequences in D . We define α is a sequential pattern. The support of a sequence $\sup(\alpha)$ in D is the number of sequences in D which contains α . Given a minimum support threshold, denoted as *min_sup*, the set of **frequent sequential patterns** is a set of all the sequences whose support is no less than *min_sup*. The set of **closed frequent sequential pattern** is a set of sequences which have no *super-sequence* with the same support.

According to the frequent sequential patterns and the closed frequent sequential patterns, a text collection C is recognized as the above sequence database D, each sentence in C as a sequence in D, and the theme pattern and the closed theme pattern is defined as follows.

Definition 1 (Theme Pattern). Given a text collection C equal to a sequence database, a theme pattern is a frequent sequential pattern in C.

Definition 2 (Closed Theme Pattern). Given a text collection C equal to a sequence database, a closed theme pattern is a closed frequent sequential pattern in C.

Definition 3 (Theme Pattern Interestingness)[8]. Given a set of patterns P, there is a function which maps any pattern $p \in P$ to a real value and is used to measure interestingness about pattern p, denoted as *I(p).* In this paper, the interestingness of theme pattern is weighted by a *tf-idf* scoring function, denoted as follows:

$$I(p)=\sum_{i=1}^{t}\frac{1+\ln(1+\ln(tf_i))}{(1-s)+s\dfrac{dl}{avdl}}\times\ln\frac{N+1}{df_i} \tag{1}$$

where tf_i equals the support of the pattern p, df_i is the inverse sentence frequency of a word, dl is the average sentence length associated with P, $avdl$ is the overall average sentence length, N is the number of sentences in the text collection, and s is an empirical parameter (usually 0.20).

Definition 4 (Jaccard Distance)[7]. A distance measure $Dis : P \times P \rightarrow [0,1]$ is a function mapping two patterns $p_m, p_n \in P$ to a value in [0,1]. In this paper, we use a Jaccard Distance to measure the pattern distance between p_m and p_n:

$$Dis(p_{m,} p_n) = 1 - | D_m \cap D_n | / | D_m \cup D_n | \tag{2}$$

Since the ideal redundancy measure $R(p_m, p_n)$ of any two theme patterns is generally difficult to obtain, we use the above Jaccard Distance to approximate the redundancy.

Definition 5 (Theme Pattern Redundancy). Given a set of theme patterns P, there is a function which maps any two theme pattern $p_m, p_n \in P$ to a real value and is used to measure redundancy between any two pattern p_m, p_n, denoted as $R(p_m, p_n)$

$$R(p_m, p_n) = 1/D(p_m, p_n) \tag{3}$$

According to definition (4) and (5), the redundancy ∞ (Jaccard Distance *0*) means two patterns are completely relevant, and redundancy *1* (Jaccard Distance *1*) means two patterns are completely independent.

2.2 Function of Preferred Compatibility

In this subsection, we describe how to measure the compatibility between interestingness and redundancy of theme patterns based on users' preferences (In what follows, we brief it as the preferred compatibility). In this paper, for simplicity of the problem, we assume that the users' preferences are evaluated by two categories: interestingness and redundancy.

Definition 6 (Function of Preferred Compatibility between Two Theme Patterns). Given a theme pattern set P, *I(x)* measures the interestingness of a pattern belong to the set *P*. *R(x)* measures the redundancy of every two patterns. l represents a proportion value between redundancy and interestingness in the users' preferences. $C(p_m, p_n)$**,** an evaluation function**,** denotes the value of the preferred compatibility between two patterns, which maps any two patterns $p_m, p_n \in P$ to a real value, shown as:

$$C(p_m, p_n) = I(p_m) + I(p_n) / R(p_m, p_n)^l \tag{4}$$

In formula (4), $I(p_m)+I(p_n)$ is the sum of interestingness of the two patterns and $R(p_m, p_n)$ is the redundancy of them. The function satisfies the feature that interestingness is inversely proportional to redundancy, namely, both increasing the interestingness and decreasing the redundancy will lead to the increase of $C(p_m, p_n)$, vice versa. In the followed Theorem 1, we will explain why l can measure the users' preferences. To prove the Theorem 1, we will firstly introduce the concept of elasticity.

Definition 7 (Elasticity) [6]. Elasticity of a differentiable function f at point x is the ratio of the incremental change of the logarithm of a function with respect to an incremental change of the logarithm of the argument, it is defined as:

$$Ef(x) = x / f(x) * f'(x) = d \ln f(x) / d \ln x \tag{5}$$

Theorem 1. The variable l in the formula (4) is the users' preference proportion both interestingness and redundancy of pattern.

Proof of Theorem 1. Let $I = I(p_m)+I(p_n)$, $R = R(p_m, p_n)$, then the formula (4) is substituted by a new function only including I and R. The new function is shown as:

$$G(I,R) = I / R^l \tag{6}$$

According to the concept of elasticity, we can get elasticity of S and R respectively:

$$\begin{cases} \dfrac{\partial G}{\partial I} \Big/ \dfrac{G(I,R)}{I} = \dfrac{1}{R^l} \dfrac{I}{I/R^l} = 1 \\ \dfrac{\partial G}{\partial R} \Big/ \dfrac{G(I,R)}{R} = -l \dfrac{I}{R^{l+1}} \dfrac{R}{I/R^l} = -l \end{cases} \tag{7}$$

From formula (7) we can see that the proportion of the elasticity of R to that of I with function $G(I,R)$ is just $|l|$. According to the concept of elasticity, when I and R change 1% respectively, the relative changes of $G(I,R)$ with them are just $|l|$ times different. Hence, $|l|$ represent the proportion of the relative changes of $G(I,R)$ influenced by I and R respectively. The single influence just denote the users' preferences proportion of interestingness and redundancy, so we get Theorem 1. □

2.3 Compatibility Graph

From the evaluation function of the preferred compatibility defined in the previous subsection, it is natural to think about how to measure compatibilities among n patterns. Since the redundancies of patterns are influenced by their interestingness and themselves, we should take interestingness and redundancy into account simultaneously rather than compute them respectively.

Previous researches have employed the redundancy graph to archive all information about interestingness and redundancy of patterns. A redundancy graph of a set of patterns is a weighted complete graph where every vertex corresponds to a pattern. The weight of vertex is the interestingness of pattern and the weight on the edge (m, n) is the redundancy of p_m and p_n. However, such redundancy graph may leads to separately considering the interestingness and the redundancy.

Since each theme pattern has a compatibility with any other theme patterns, in addition, the compatibility of two theme patterns are influenced by their interestingness, it is a crucial problem that how to distribute interestingness of every pattern into their compatibility. A reasonable solution is to partition the interestingness of every pattern by n-1 parts on average if the power of the set of patterns P is n. The distributed interestingness, $ID(P_i)$, is shown as:

$$ID(P_i) = 1/n - 1 * I(P_i) \tag{8}$$

According to formula (4) and formula (8), given a set of n theme patterns P, the preferred compatibility between two theme patterns can be redefined as follows:

$$CD(p_m, p_n) = [ID(p_m) + ID(p_n)] / R(p_m, p_n)^l \tag{9}$$

By formula (9), we can propose a novel structure, called compatibility graph, which is used to archive the preferred compatibilities among n patterns.

Definition 8 (Theme Compatibility Graph). Given a set of theme pattern P, the power of the set is |P|, a compatibility graph of P is a weight complete graph $G(P) = G(V, E)$ where each vertex m in the vertex set V corresponds to a pattern P_m. The edge set of $G(P)$ is $E = \{e_{uv} = CD(p_u, p_v) \mid (u,v) : u, v \in V, u \neq v\}$.

2.4 Maximal Weight Sum Problem

According to the above introductions of the evaluation function and the compatibility graph, we have stored the preferred compatibilities of a set of patterns into the compatibility graph. The next crucial step is to make a reasonable problem formulation based on the compatibility graph.

Since we aim at discovering PCTTP, the result set ought to have K patterns. Let the total compatibility of K patterns be written as TC. In general, there are the redundancies with every two patterns. With the compatibility graph, a general total compatibility of K theme patterns is shown as:

$$TC == \sum_{i=1}^{k}\sum_{j=i}^{k} CD(p_i, p_j) = \sum_{i=1}^{k}\sum_{j=i}^{k} ID(p_i) + ID(p_j) \big/ R(p_i, p_j)^l \tag{10}$$

In formula (10), $ID(p_i)$ represents the interestingness fused into the compatibility. The goal of discovering PCTTP is to maximize the result of formula (10). Hence, we firstly define what the maximal weight sum problem is as follows:

Definition 9 (Maximal Weight Sum Problem). Given a weighted complete graph G with n vertices, selecting K vertices from n vertices to make the sum of weight on every edge is maximal.

Hence, given a set of theme patterns P whose power is |P|, formulating the problem of discovering PCTTP is to the MWSP in the compatibility graph whose number of node is |P| correspondingly. However, we can obviously find it impossible to solve MWSP by enumerating. Actually, it is a NP-Hard problem. *Why the MWSP is a NP-Hard problem? The proof of it will be given in the next section.*

3 NP-Hardness

In this section, we show that the MWSP defined above is NP-Hard.

Theorem 2. The Maximal Weight Sum Problem is NP-Hard.

It is well-known that the optimization of one problem must be NP-Hard if its decision problem is NP-Complete. Hence, we firstly transform the MWSP to its corresponding decision problem, and then prove the decision problem is NP-Complete. In order to do it, we need the following definition.

Definition 10 (Decision problem of Maximal Weight Sum Problem). Given a weighted complete graph G which has n vertices, whether there exist k vertices in n vertices with the sum, no less than a given value M, of all edges for the k vertices.

The decision problem is written as WSP (Weight Sum Problem)

Proof of Theorem 2. We firstly show that WSP can be verified in polynomial time. Suppose we are given a graph G=(V, E), an integer k and a given value M. We use the result set of vertices V' ⊆ V as a certificate for G. The verification algorithm affirms that |V'| = k, and then it checks the total weighted sum of all edges between every two vertices in V' is no less than M. This verification can be performed straightforwardly in polynomial time.

Then, we prove that the MWSP is NP-hard by showing that CLIQUE $\leq_P$ WSP. This reduction is based on the concept of the "complement" of a graph. Given an undirected complete graph G = (V, E), we define the complement of G as $\overline{G} = (V, \overline{E})$, $\overline{E} = \{(u,v) : u, v \in V, u \neq v, (u,v) \notin E\}$. Let the edges in E be weighted 1 and the edges in $\overline{E}$ is weighted 0, thus we get an undirected complete graph $G^{'} = G + \overline{G}$.

Based on the complete weighted graph $G^{'}$ defined above, if there is a clique of k vertices, the weighted sum of all edges in the clique would be k(k-1)/2, namely, the weighted sum of all edges between every two vertices of k vertices is maximal. Whereas, the graph including k vertices must be a clique, if the weighted sum of all edges between every two of k vertices is k(k-1)/2. Hence, the clique problem, a well-known NP-Complete problem, can be reduced to WSP. This implies that the decision problem of MWSP is NP-Complete and so we finish the proof of Theorem 2.

4 Discovering Preferred Compatible Theme based Greedy

In this section, we describe an algorithm for mining PCTTP. Since MWSP is a NP-Hard problem, it's natural to get the idea of developing an approximate algorithm to solve MWSP. We design a greedy algorithm, called DPCTG (Discovering Preferred Compatible Theme based Greedy), which approximates the optimal solution. The pseudo-code of DPCTG is shown in Algorithm 1.

The DPCTG algorithm contains two steps. The first one is to select a edge whose weight is maximal out of n(n-1)/2 edges of the compatibility graph with n vertices and then the edge will be archived into the result set. The second one is to iteratively select the k-2 vertices from remaining n-2 vertices. Each time choose one vertex from those remained that has the maximal sum of weight between itself and all vertexes in the present result set.

Algorithm 1. Discovering Preferred Compatible Theme based Greedy

```
Input: A set of n closed frequent theme patterns TP
    A compatibility graph contained n vertices CG
    Number of output closed frequent theme, k
Output: A result set of k patterns. RS
Selecting two theme patterns, $p_m, p_n$, which are contained in a edge
  whose weight is maximal in all edges of CG.
while (The size of RS is no more than k)
 do
   search for a vertex which maximize the sum of edge weight between it
      and all vertexes in the present result set
   $RS \leftarrow RS \cup P_i$
return RS
```

5 Quality Evaluation Model

We have transformed the post-processing of mining frequent sequential pattern from the discovering PCTTP. However, how can we measure that the result set discovered by MPCTG is the best PCTTP based on users' preferences ? Since traditional evaluation approaches of mining frequent pattern can no longer apply to the interestingness and redundancy measuring of patterns simultaneously, we propose a quality evaluation model that is able to measure the compatibility of the discovering PCTTP.

Given a set of closed theme pattern, we are able to discover k theme patterns with the top-k interestingness, and discover other k theme patterns which have the minimal redundancy between any patterns. The prefect case is that the above two set including k theme patterns are the same, however it is impossible generally. Thus, we define the extreme interestingness of k theme patterns by summing individual interestingness of the top-k interesting theme patterns, and define the extreme redundancy of k theme patterns by summing all redundancies between of k theme patterns with the minimal redundancy of every two patterns. Then, "the extreme average compatibility of k theme patterns" is proposed, which is the ratio between extreme interestingness and extreme redundancy.

Definition 11 (Extreme Interestingness of K Theme Patterns). Given a set of the closed theme patterns $P = \{t_1, t_2, ..., t_n,\}$, the top-k interesting theme patterns among P is $IK = \{t_1, t_2, ..., t_k,\}$, the extreme interestingness of k theme patterns of S is the sum of individual interestingness of theme patterns in IK, denoted as EIK.

$$EIK = \sum_{i=1}^{k} I(t_i) \tag{11}$$

Definition 12 (Extreme Redundancy of K Theme Patterns). Given a set of the closed theme patterns $P = \{t_1, t_2, ..., t_n,\}$, the k theme patterns with minimal redundancy between any patterns among S is $RK = \{t_1, t_2, ..., t_k,\}$. The extreme redundancy of k theme patterns of S is the sum of individual redundancies between any theme patterns in RK, denoted as ERK.

$$ERK = \sum_{i=1}^{k} \sum_{j=i+1}^{k} R(t_i, t_j) \tag{12}$$

Definition 13 (Extreme Average Compatibility of K Theme Patterns). Given a set of the closed theme patterns $P = \{t_1, t_2, ..., t_n,\}$ .The extreme interestingness of k theme patterns in S is EIK and the extreme redundancy of k theme patterns in S is ERK. An extreme average compatibility of k theme patterns among S is denoted as: $AECK = EIK / ERK$.

According to the above definitions, for a set of the closed theme patterns P, extreme interestingness of k theme patterns and extreme redundancy of k theme patterns measure the most ideal interestingness and redundancy among the set S respectively. Thus, they can be regarded as two measuring criterions. Thus, the extreme average compatibility of k theme patterns is naturally regarded as the criterion measuring the compatibility of k theme patterns. In order to quantify the quality of the compatibility of k theme patterns, it is necessary to compute an approximation ratio with the extreme average compatibility of k theme patterns. Therefore, the average compatibility of k theme patterns and the approximation ratio of k theme patterns will be defined as follows.

Definition 14 (Average Compatibility of K Theme Patterns). Given a set including k theme patterns $KT = \{t_1, t_2, ..., t_n,\}$. An average compatibility of k theme patterns among KT is denoted as:

$$ACK = \sum_{i=1}^{k} I(t_i) \Big/ \sum_{i=1}^{k} \sum_{j=i+1}^{k} R(t_i, t_j) \tag{13}$$

Definition 15 (Approximation Ratio of K Theme Patterns). Given a set of the closed theme patterns $P = \{t_1, t_2, ..., t_n,\}$ and a set of the discovering k theme patterns $KT = \{t_1, t_2, ..., t_n,\}$ .The $AECK$ and ACK can be computed respectively. An approximation ratio of k theme patterns is denoted as: $AR = ACK / AECK$

To sum up, the quality evaluation model compute the approximation ratio of k theme patterns to quantify the quality of the compatibility of the discovering k theme patterns from text. The approximation ratio ranges in [0,1], and The more the ratio closer to 1,the better the compatibility of discovering k patterns is. The further discussion of the quality evaluation model will be shown in the next section.

6 Experimental Study

In this section we will provide the empirical evaluation for the effectiveness of our strategy and EPCEG algorithm for real-world tasks. The EPCEG algorithm is implemented in Java. The version of JDK is JDK1.52. All experiments are performed on a 2.0GHZ, 2GB-memory, Intel PC running Windows XP.

The four text datasets used in our experiment all come from subsets of the DBLP dataset. They contain papers from the proceedings of 25 major conferences, such as SIGMOD, SIGKDD, VLDB, SIGIR and etc., in Data Mining, Database and Information Retrieval. Considering the relations and the differences between the above four subjects, we classify these titles of papers into four text datasets, including database, data mining, database and data mining, data mining and information retrieval respectively. The detail information of datasets is shown in Table 1. In those experiments, we firstly use the tool Clospan [10] to generate closed sequential patterns, namely closed theme patterns, and the title terms are stemmed by Krovertz stemmer [2]. Then we discover the PCTTP based on the result set from the Clospan.

According to the quality evaluation model, we use TSP (mining closed top-k sequential pattern algorithm [5]) and DPCTG by different support thresholds, at the users' preferences proportion $l = 5/5 = 1$, on the four datasets respectively, and the approximation ratios are illustrated as curves in the figures. For the sparse datasets, the chosen support thresholds are low than 5%. From the figures, we can see that the compatibility of the set of discovering PCTTP is obviously better than that of the result set of TSP as the min_support is decreasing. Results of four datasets are shown respectively with the support threshold set 5 as a representative value. Four columns of each represent the 10 top results of TSP, and DPCTG with different l respectively. In ordering to explore the generalization of results, we assume l as 1/9, 5/5 and 9/1 in table2, l as 2/8, 5/5 and 8/2 in table 3, l as 3/7, 5/5 and 7/3 in table 4 and l as 1/9, 5/5 and 9/1 in table 5. From these tables, the italic represents the inclusion relations among this theme, and the bold represents the significant and focus theme. These tables indicate that our proposed strategy can meet the different demands when we adjust the users' preferences proportion.

To sum up, DPCTG can find effective result set of PCTTP from a large set of text. In addition, different users can obtain different result sets to satisfy the users' preferences with interestingness and redundancy by DPCTG.

Table 1. Information of four data sets

Table ID	Topic of data set	Number of Transaction	Conferences
Table 2	Data mining	984	AAAI ,SIGKDD, CIKM,SDM, ICDM, PKDD, PAKDD, ADMA
Table 3	Database	15280	SIGMOD, PODS, VLDB, ICDE, ICDT, CIDR, ER, EDBT SSDBM, DASFAA, WAIM
Table 4	Data Mining and Database	24404	AAAI ,SIGKDD, SIGMOD, PODS, VLDB, ICDE, ICDT, CIDR, ER, CIKM, SDM, ICDM, EDBT SSDBM, PKDD, PAKDD, DASFAA, WAIM, ADMA
Table 5	Data Mining and Information Retrieval	13531	AAAI, SIGKDD, SIGIR,WWW,WISE,ECIR , CIKM,SDM, ICDM, PKDD, PAKDD, APWeb, DMA

Table 2. Top-10 Theme Pattern on the Dataset of DM

Top-k	TSP	DPCTG (R/I=1/9)	DPCTG (R/I=5/5)	DPCTG (R/I=9/1)
1	*data*	*data*	data mining	**association rule**
2	*mining*	*mining*	**association rule**	data clustering
3	clustering	clustering	**frequent pattern**	**support vector machine**
4	*data mining*	*data mining*	database	knowledge discovery
5	database	database	classification	information retrieval
6	pattern	pattern	data clustering	**feature selection**
7	classification	classification	search	**text classification**
8	learning	learning	**time series**	**data streams**
9	discovery	association rule	**support vector machine**	mining database
10	rule	discovery	information retrieval	**neural network**

Table 3. Top-10 Theme Pattern on the Dataset of DB

Top-k	TSP	DPCTG (R/I=2/8)	DPCTG (R/I=5/5)	DPCTG (R/I=8/2)
1	data	data	database system	**database management system**
2	*database*	*database*	database management	**distributed database**
3	query	query	**distributed database**	**relational database system**
4	*system*	management	**query processing**	database design
5	management	*database system*	database design	data streams
6	processing	relational	data streams	**query optimization**
7	relational	query processing	xml	data model
8	model	data model	data model	data mining
9	*database system*	xml	data mining	query processing
10	xml	data stream	relational database	**concurrency control**

Table 4. Top-10 Theme Pattern on the Dataset of DM&DB

Top-k	TSP	DPCTG (R/I=3/7)	DPCTG (R/I=5/5)	DPCTG (R/I=7/3)
1	*data*	*data*	data	**association rule mining**
2	database	database	data mining	distributed database systems
3	*mining*	*query*	database	query processing
4	query	*data mining*	data clustering	**Support vector machine**
5	system	system	association rule	data clustering
6	clustering	information	Data stream	Data stream
7	information	*query processing*	query processing	time series
8	*data mining*	data clustering	information	knowledge discovery
9	management	data management	**distributed database systems**	**database management systems**
10	processing	relational	time series	mining pattern

Table 5. Top-10 Theme Pattern on the Dataset of DM&IR

Top-k	TSP	DPCTG (R/I=1/9)	DPCTG (R/I=5/5)	DPCTG (R/I=9/1)
1	data	data	information retrieval	**information retrieval system**
2	web	web	**data**	**association rule mining**
3	mining	mining	**data mining**	web search
4	*retrieval*	*retrieval*	database	knowledge discovery
5	*information*	*information*	classification	**support vector machine**
6	search	search	clustering	**web semantic**
7	clustering	clustering	web search	feature selection
8	*Information retrieval*	*information retrieval*	text	web mining
9	text	data mining	**web sematic**	text classification
10	learning	web search	**association rule mining**	**time series data**

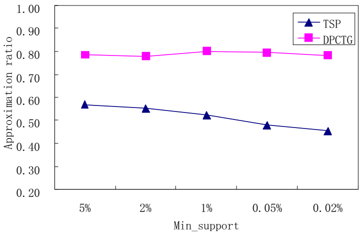


Fig. 1. Approximation Ratio on DM

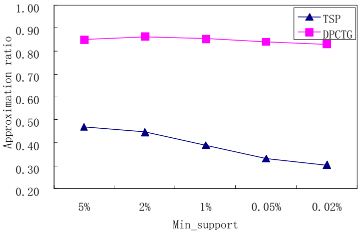


Fig. 2. Approximation Ratio on DB

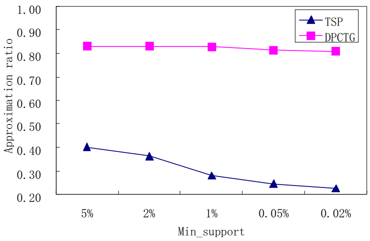


Fig. 3. Approximation Ratio on DM&DB

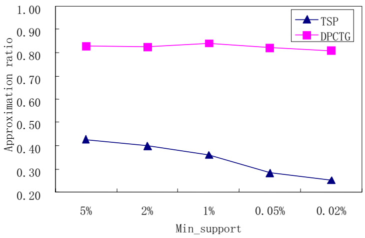


Fig. 4. Approximation Ratio on DM&IR

7 Relate Work

To the best of our knowledge, the problem of discovering PCTTP has not been well studied in current work. Most studies of frequent pattern mining still focus on the fast mining result set from the database and dose not pay attention to the post-processing of KDD. Recent studies in this field have paid close attention to mining compatible top-k patterns with both high-interestingness and low-redundancy, instead of finding top-k closed frequent patterns or compressing the size of frequent patterns to exclude redundancy among them separately, such as mining closed frequent pattern, compressing patterns, summarizing patterns and so on[1, 7, 9, 10].

References [5] [9] mine top-k closed frequent sequences and summarizing patterns respectively. But these papers only study the single factor, either interestingness or redundancy, not evaluate compatibility of the both to the effect of the results. Reference [8] mines top-k theme patterns considering interestingness and redundancy simultaneously, however, ignores to study the influences of different users on results, leading to the limits of the model.

In the field of text mining, semantic analysis is still a hotspot topic [3, 4]. However, there are few focuses on the problem of discovering theme patterns based on users' preferences considering interestingness and redundancy simultaneously.

8 Conclusion

Existing work of discovering a representative set of theme patterns from a large amount of text by frequent pattern mining usually generates a huge amount of theme patterns for the downward closure property. Thus, these discriminative theme patterns will be drowned in a large number of redundant patterns. Some recent post-processing studies of KDD introduced the technique of mining a set of top-k frequent patterns with both high-interestingness and low-redundancy. However, since different users have different preferences with interestingness and redundancy, it is hard to use one unified criterion to discover theme patterns with both high-interestingness and low-redundancy. The problem of discovering PCTTP has not been well addressed so far.

In this paper, the novel strategy of the post-processing in text mining is proposed, that is, discovering PCTTP based on users' preferences. For the problem discovering PCTTP, the evaluation function of preferred compatibility between every two theme patterns is presented firstly. Then the preferred compatibilities are archived into a data structure called theme compatibility graph, and a problem called MWSP based on the compatibility graph is proposed to formulate the problem how to discover the PCTTP. In addition, since MWSP is proved to be a NP-Hard problem, a greedy algorithm, DPCTG, is designed to approximate the optimal solution of the MWSP. Empirical studies indicate that a high quality set of PCTTP can be obtained on the different text datasets from DBLP.

The proposed strategy and algorithm of this paper is general, however, we only study the theme patterns from a huge amount of text datasets, and use tf-idf approach to weight the interestingness of theme pattern. In the future work, we will study the text datasets which have the time attributes, and extend our strategy to other information retrieval models, such as the probabilistic model.

Acknowledgments

We would like to thank Zhiyuan Cheng and Yuhan Song for discussions regarding our work. This research was supported by National 973 Project of China under Grant No.2005CB321902.

References

1. Afrati, F.N., Gionis, A., Mannila, H.: Approximating a collection of frequent sets. In: KDD 2004, pp. 8–19 (2004)
2. Krovetz, R.: Viewing morphology as an inference process. In: Proceedings of SIGIR 1993, pp. 191–202 (1993)
3. Mei, Q., Xin, D., Cheng, H., Han, J., Zhai, C.: Generating semantic annotations for frequent patterns with context analysis. In: KDD 2006, pp. 337–346 (2006)
4. Mei, Q., Xin, D., Cheng, H., Han, J., Zhai, C.: Discovering Evolutionary Theme semantic annotations for frequent patterns with context analysis. In: KDD 2005 (2005)
5. Tzvetkov, P., Yan, X., Han, J.: TSP: Mining top-k closed sequential patterns. Knowledge and Information Systems 7, 438–457 (2005)
6. Varian, H.: Intermediate Microeconomics: A Modern Approach, 6th edn. W.W. Norton & Company Inc. (2003)
7. Xin, D., Han, J., Yan, X., Cheng, H.: On compressing frequent patterns. In: KIS 2007 (2007)
8. Xin, D., Han, J., Yan, X., Cheng, H.: Discovering Redundancy-Aware Top-K Patterns. In: KDD 2006, pp. 314–323 (2006)
9. Yan, X., Cheng, H., Han, J., Xin, D.: Summarizing Itemset Patterns: A Profile Based Approach. In: KDD 2005, pp. 314–323 (2005)
10. Yan, X., Han, J., Afshar, R.: Clospan: Mining closed sequential patterns in large datasets. In: SDM 2003, pp. 166–177 (2003)

Juicer: Scalable Extraction for Thread Meta-information of Web Forum*

Yan Guo[1], Yu Wang[1], Guodong Ding[1],
Donglin Cao[1], Gang Zhang[1], and Yi Lv[2]

[1] Institute of Computing Technology, Chinese Academy of Sciences
[2] State Key Laboratory of Computer Science, Institute of Software,
Chinese Academy of Sciences
guoy@ict.ac.cn

Abstract. In Web forum, thread meta-information contained in list-of-thread of board page provide fundamental data for the further forum mining. This paper describes a complete system named Juicer which was developed as a subsystem for an industrial application that involves forum mining. The task of Juicer is to extract thread meta-information from board pages of a great many of large scale online Web forums, which implies that scalable extraction is required with high accuracy and speed, and minimal user effort for maintenance. Among so many existed approaches about information extraction, we can not find any approach to fully satisfy the requirements, so we present simple scalable extraction approach behind Juicer to achieve the goal. Juicer is constituted by four modules: Template generation, Specifying labeling setting, Automatic extraction, Label assignment. Both experiments and practice show that Juicer successfully satisfied the requirements.

1 Introduction

The work in this paper focuses on Web forum extraction. Forum extraction produces structured data ready for postprocessing, and it is crucial to many applications of forum mining or forum retrieval.

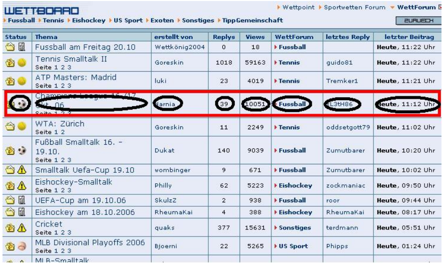
A system named Juicer will be described in the paper. Juicer was developed as a subsystem for an industrial application that involved forum mining. The task of Juicer is to extract thread meta-information from list-of-thread of board pages from a great many of large scale online forums in Chinese. Figure 1 shows an example board page [1]. Thread meta-information provides rich and fundamental data for the further forum mining. The task can be divided into two subtasks:(1)Data extraction: Extracting records and recognizing each items from list-of-thread of board page; (2)Label assignment: Assigning meaningful labels to the items in which user is interested; for example, thread-title, thread-author, and view count.

* This work is partially supported by the National Grand Fundamental Research 973 Program of China under Grant 2004CB318109, and the National High Technology Development 863 Program of China under Grant 2007AA01Z147.

[1] From the forum http://www.forum.wettpoint.com/

H. Chen et al. (Eds.): PAISI 2009, LNCS 5477, pp. 143–148, 2009.

© Springer-Verlag Berlin Heidelberg 2009



WETTBOARD

Wettpoint Sportwetten Forum WettForum

Fussball Tennis Eishockey US Sport Exoten Sonstiges TippGemeinschaft

ZURUECK

Status	Thema	erstellt von	Replys	Views	WettForum	letztes Reply	letzter Beitrag
	Fussball am Freitag 20.10	Wettkönig2004	0	18	Fussball		**Heute**, 11:22 Uhr
	Tennis Smalltalk II Seite 1 2 3	Goreskin	1018	59163	Tennis	guido81	**Heute**, 11:22 Uhr
	ATP Masters: Madrid Seite 1 2 3	luki	23	4019	Tennis	Tremker1	**Heute**, 11:21 Uhr
	Champions League 16./17. Okt. 06 Seite 1 2 3	Narnia	39	10051	Fussball	L3tH86	**Heute**, 11:12 Uhr
	WTA: Zürich Seite 1 2	Goreskin	11	2249	Tennis	oddsetgott79	**Heute**, 11:02 Uhr
	Fußball Smalltalk 16. - 19.10. Seite 1 2 3	Dukat	140	9039	Fussball	Zumutbarer	**Heute**, 10:20 Uhr
	Smalltalk Uefa-Cup 19.10	wombinger	9	671	Fussball	Zumutbarer	**Heute**, 10:02 Uhr
	Eishockey-Smalltalk Seite 1 2 3	Philly	62	5223	Eishockey	zockmaniac	**Heute**, 09:50 Uhr
	UEFA-Cup am 19.10.06	SkulzZ	2	938	Fussball	roor	**Heute**, 09:44 Uhr
	Eishockey am 18.10.2006	RheumaKai	4	388	Eishockey	RheumaKai	**Heute**, 08:17 Uhr
	Cricket Seite 1 2 3	quaks	377	15631	Sonstiges	terdmann	**Heute**, 05:51 Uhr
	MLB Divisional Playoffs 2006 Seite 1 2 3	Bjoerni	22	5265	US Sport	Phipps	**Heute**, 01:24 Uhr
	MLB-Smalltalk						

Fig. 1. An example board page

Most of the Web forum sites are designed as dynamic sites. Thread meta-information contained in list-of-thread of board page is usually stored in a database, and board page is generated dynamically with some pre-defined templates. So forum board pages generally have the following common characteristics: (1)The layout structures of board pages from the same Web forum are often similar, or to the worst, can be classified into several similar groups. (2)The layout structures of the post records in list-of-thread are very similar. Since Juicer was to be integrated in an industrial application, we require scalable extraction with high accuracy and speed, and minimal user effort for maintenance.

Web information extraction has received much attention over the last few years. Chia-Hui Chang et al. [1] presented a good survey on the major Web data extraction approaches. According to the task and the requirements for scalable extraction, MDR[2](further work are DEPTA [3] and NET [4]) was selected for the subtask of data extraction. However the experimental results showed that MDR can not fully satisfy the requirements. Inspired by MDR, we provide an unsupervised method to generate template offline, and then perform automatic extraction online based on the template.

For the subtask of label assignment, in order to get high precise we did not use a fully automatic method. We provide a friendly interface for user to specify labeling setting offline, and perform automatic label online for extracted data based on regular expression matching. Both experiments and practice showed that the manual labor cost for specifying labeling setting can be tolerant. Since the work for this subtask is not much of a novelty, the details about it will not present in this paper.

The underlying algorithms of Juicer are simple to implement. Juicer can fully satisfy the requirements for scalable extraction.

2 System Architecture

Juicer consists of four modules as shown in Figure 2. *Template generation* works offline. A sample board page is input, and template including extracting parameters is output. Using HTML tag tree, the module learns extracting parameters by an unsupervised method. *Specifying labeling setting* works offline. A friendly interface is provided for user to specify labeling setting. Labeling setting are

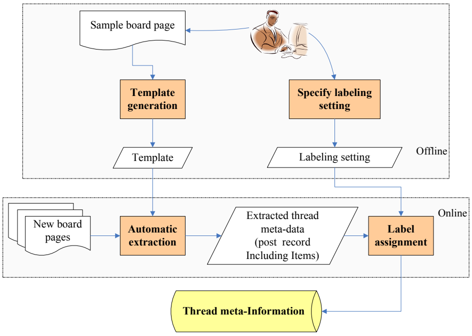


Fig. 2. System architecture of Juicer

output of this module. *Automatic extraction* works online. New board pages and template are input, and thread meta-data (post records including items) are output. For each new board page, thread meta-data are extracted based on the template automatically. *Label assignment* works online. Extracted thread meta-data and labeling setting are input, and thread meta-information is output. For each piece of thread meta-data, the meaningful label is assigned based on the labeling setting automatically.

3 Data Extraction

The nested structure of HTML tags in a Web page naturally forms an HTML tag tree. We have two observations for tag tree of forum board page:

1. Each post record in list-of-thread in a board page is formed by one and only one node with its sub-trees. Figure 3 shows part of HTML tag tree for board page of Figure 1. In Figure 3, a group of tags which contains a post record is labeled in an ellipse. Each post record is wrapped in a <tr> node with its sub-tree.
2. The post records in list-of-thread in a board page are a set of similar objects. Each object is formatted using similar HTML tags. For example, in Figure 3, the two <tr> nodes are similar, for each of them has 8 <td> child nodes, and each <td> node with its sub-tree contains an item of the corresponding post record. The node <tr> is called as *post-node*. A post-node and its sub-tree together contain a post record.

According to the observations, post-nodes share some common features. If the common features can be precisely summarized, all nodes of tag tree can be divide into several groups based on them. One and only one group including all post-nodes may be found. After experiments and analysis we classify board pages into three classes, named as Table-type, Div-type, and Other-type. These three classes of board pages account for about 60%, 30%, and 10% of the experimental collection respectively. Each class has its own features. For example, for a Table-type board page, the layout of list-of-thread relies on tags of <table>, <tr>,

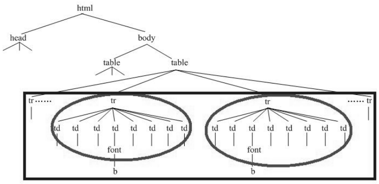


Fig. 3. Part of HTML tag tree for the board page of Figure 1

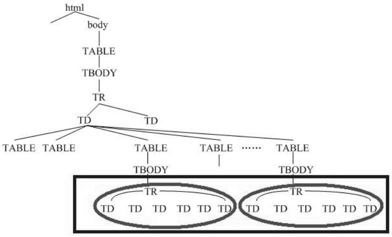


Fig. 4. Part of HTML tag tree for a board page

<td> etc. And post-nodes in the tag tree share following features: (a)The tag name of post-nodes are the same but not <div>; (b)The depth of post-nodes in the tag tree is the same; (c)Each post-node has the same number of children, and has more than 3 children. For the Table-type and the Div-type board page, we propose a fully automatic data extraction method. For the Other-type board page, we provide a friendly interface for user to write wrapper manually.

A post-node with its sub-tree contains a post record, and each child node with its sub-tree contains an item of the corresponding post record. The corresponding relation between the child node of post-node and the item of post record makes it easy to recognize items of post record.

Here the post-nodes are only restricted to be at the same depth. While in MDR the generalized nodes which forming a data region are restricted to having the same parent and being all adjacent. Compared with MDR, our hypothesis is more suitable for the task. For example, Figure 4 shows a tag tree for a kind of board page of Web forum (some details are omitted). The post-nodes are <TR> nodes. Each post record is also contained in a <TABLE> node (which is the father of the father of the <TR> node) with its sub-tree. Based on the restriction of MDR, the <TABLE> nodes can be mined, however each child node of <TABLE> node can not be corresponded with each item of the post record, which makes it difficult to recognize the items from the sub-tree of <TABLE> node.

4 Experiments and Evaluations

Experimental results are presented mainly in accuracy, speed and manual labor cost. All experiments were performed on a Windows XP with 3.4 GHz Pentium IV processors and 512MB memory. The testing data includes board pages from

45 real Web forum sites which are popular in China. Some of the forums are powered by popular forum tool in China, such as Discuz!, DVBBS, and PHPWIN, and others are self-developed forums. The open source software tidy [2] was used to make up the HTML sources . The target of MDR is at general web sites rather than web forum sites, so it seems unreasonable to compare with MDR. However in order to evaluate the ability of Juicer to finish the task, here just for our task, a comparisons with MDR is still made. We use the execute system of MDR [3] to experiment on the same testing data.

4.1 Accuracy

The experiments showed that once a post record is identified, all of its items can be recognized absolutely correctly. Based on labeling setting, the average correctness of label assignment is 100%. So we only provide the evaluation of the accuracy of extracting post records. We use the standard measures of Precision and Recall to evaluate the accuracy. The experimental results showed that: The average of Recall can reach to 100%. The average of Precision is 98.11% that is also very high but less than 100%, which is because that for some Web forums there is one and only one incorrectly extracted record. The incorrect one is the descriptive record containing the names of items of post record, such as “Title”, “Author”, “View count”. In some cases, the layout structure of the descriptive record is similar to the structure of the other post records, so it is often wrongly considered as a post record. We are satisfied with the accuracy of Juicer.

Data records extracted by MDR in experiment not only include post records but also include some other data records we are not interested in, such as announcements. It is due to the assumption of MDR that nontag tokens are data items to be extracted [1].

4.2 Speed

From the experiments, the average of time cost for generating template is 0.148 second. The average of time cost for automatic extraction and labeling is also 0.148 second, which not only includes the process of extracting post records and recognizing all items, but also includes the process of label assignment. They are really not time consuming. This speed is high enough for the scalable extraction.

By cursory counting, it seemed that the average time cost for extracting one board page by MDR is about 1 second, which includes the process of extracting similar data records without recognizing items of data records (MDR does not perform this process). So for our task, our method is faster than MDR. This result is mainly due to that: firstly, the process of finding similar data records in our method is simpler than MDR: MDR uses string matching whereas our method only uses some features of node, such as the tag name of node; secondly, during the process of extraction, our method deals with board page based on

[2] http://tidy.sourceforge.net/

[3] http://www.cs.uic.edu/∼liub/WebDataExtraction/

template, while MDR does not generate wrapper and performs the complex and time-consuming extraction for each board page.

4.3 Manual Labor Cost

Manual labor cost mainly includes two parts: (1)Find sample board page for template generation; (2)Manually specify labeling setting for label assignment.

For (1), from the two important characteristics of Web forum board page (discussed in section 1), it can be seen that for one forum, there is often only one template, or to the worst, there are only several templates. To generate one template, only one sample board page needs to be provided. So user effort for (1) is not high.

For (2), we try our best to make the interface much more friendly, and we have invited some persons who are not familiar to the task to specify label setting using the interface for test, and the average time cost for performing one board page is only about 2 minutes.

So the manual labor cost for maintainable can be endured.

5 Conclusions

We present a complete system named Juicer whose task is to extract thread meta-information from board pages of a great many of large scale online Web forums. Juicer has been used in the industrial application successfully. From the experimental and practical results, we can draw the conclusion that Juicer can fully satisfy the requirements both on high performance and low cost for maintenance. Besides the application for which Juicer has been developed, we believe Juicer is also very useful for other industrial applications involving forum mining or forum retrieval. In the further work, we will make Juicer more robust for practise.

References

1. Chang, C.-H., Kayed, M., Girgis, M.R., Shaalan, K.F.: A survey of web information extraction systems. IEEE transactions on knowledge and data engineering 18(10), 1411–1428 (2006)
2. Liu, B., Zhai, Y.: Mining data records in web pages. In: Proc. Intl. Conf. Knowledge Discovery in Databases and Data Mining (KDD), pp. 601–606 (2003)
3. Zhai, Y., Liu, B.: Web data extraction based on partial tree alignment. In: Proc. 14th Intl. Conf. World Wide Web (WWW), pp. 76–85 (2005)
4. Liu, B., Zhai, Y.: Net: a system for extracting web data from flat and nested data records. In: Proc. Sixth Intl. Conf. Web Information Systems Eng., pp. 487–495 (2005)

A Feature-Based Approach for Relation Extraction from Thai News Documents

Nattapong Tongtep and Thanaruk Theeramunkong

Sirindhorn International Institute of Technology
Thammasat University, Thailand
nattapong,thanaruk@siit.tu.ac.th
http://www.siit.tu.ac.th

Abstract. Relation extraction among named entities is one of the most important tasks in information extraction. This paper presents a feature-based approach for extracting relations among named entities from Thai news documents. In this approach, shallow linguistic processing, including pattern-based named entity extraction, is performed to construct several sets of features. Four supervised learning schemes are applied alternatively to investigate the performance of relation extraction using different feature sets. Focusing on four different types of relations in crime-related news documents, the experimental result shows that the proposed method achieves up to an accuracy of 95% using a data set of 1736 entity pairs. Effect of each set of features on relation extraction is explored for further discussion.

Keywords: Relation Extraction, Named Entity Extraction, Thai Language Processing, Supervised Learning, Local Features.

1 Introduction

There are several research works on transforming an unstructured text into knowledge base recently proposed [1,2,3]. The extracted knowledge is often used to respond to the 4W (who, what, when and where) questions. As one of main processes in Information Extraction (IE), named entity (NE) extraction finds a part of words such as a person name, a location and time expression. A number of name entity extraction methods were proposed in several languages, such as English, Japanese and Chinese. To utilize the extracted name entities, it is necessary to detect relations among them.

Named entities extraction in Thai is quite different from those in other languages. Kawtrakul et al. [4] summarized a state of the art of Thai language resources (e.g. corpus, lexicon and language analysis tools), Thai language behavior analysis (word, new word formation and sentence) and computational models (unknown word extraction, named entity identification, new word generation and noun phrase recognition). To the best of our knowledge, a study of extracting relations among named entities in Thai using different feature sets has not been examined.

H. Chen et al. (Eds.): PAISI 2009, LNCS 5477, pp. 149–154, 2009.
© Springer-Verlag Berlin Heidelberg 2009

This paper presents a feature-based approach to extract relations among named entities in a Thai news document using a number of supervised learning schemes. A number of different sets of features are investigated using crime-related news documents. The rest is organized as follows. Sect. 2 describes named entities in Thai and other related entities. Our feature-based approach is presented in Sect. 3. In Sect. 4, the experiment and results are discussed. A conclusion is illustrated in Sect. 5.

2 Thai Named Entities and Other Related Entities

Thai words are multi-syllabic words which may combine together to form a new word. Since Thai has no inflection and no word delimiters, morphological processing in Thai is just to recognize word boundaries, without recognizing a lexical form from a surface form, like in English. Almost all Thai new words are formed by means of compounding or adding prefix or suffix. Moreover, even it seems trivial for Thai people to recognize a proper name or transliterated words from a running text (a text without space or any delimiters), it is extremely hard for a machine to detect such name if it is not in a dictionary. Proper names and transliterated words are highly significant in news documents.

In news documents, there are several types of named entities such as date, time, location, person name, product name, plant name, and disease name, depending on their domains or categories. In our work, we focus on named entities involving with an event, such as person name (PER), location (LOC) including action (ACT). In our approach, each type of named entities can be viewed as a chunk of information. Moreover, person's position and product's name can be defined as other entities (OTH) in our work. Figure 1 shows an example of a sentence in Thai and the result of entity identification.

Sentence
นายอาคมฆ่านายสมคิดที่รัชดาผับ หลังก่อเหตุ นายอาคมได้หลบหนีไปที่จังหวัดราชบุรี Mr. Akom killed Mr. Somkid at Ratchada Pub. Thereafter Mr. Akom escaped to Ratchaburi Province.
Entity Identification
[PER01:นายอาคม][ACT01:ฆ่า][PER02:นายสมคิด][OTH01:ที่][LOC01:รัชดาผับ] [PER01:Mr. Akom][ACT01:killed][PER02:Mr. Somkid][OTH01:at][LOC01:Ratchada Pub] [OTH02:หลังก่อเหตุ][PER03:นายอาคม] [ACT02:ได้หลบหนีไป][OTH03:ที่][LOC02:จังหวัดราชบุรี] [OTH02:Thereafter][PER03:Mr. Akom][ACT02:escaped][OTH03:to][LOC02:Ratchaburi Province]

Fig. 1. Named entity identification

3 Our Approach

To extract relations in a Thai news document, pattern-based named entity extraction [5] is applied. In this work, extracted entities are of three types: person name, location and action (verb). After applying patterns on a news document, we manually correct the resultant named entities. In the rest of this section, we describe our framework, a set of features used for relation extraction, and relation types.

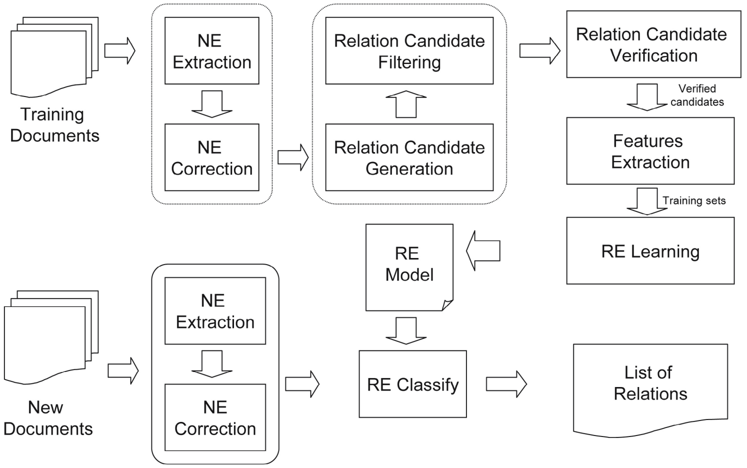


Fig. 2. The framework

The Framework. In our framework, two-stages are (1) training stage and (2) testing stage (see Fig. 2). In training stage, possible named entities are extracted (NE Extraction) from texts in a document and manually corrected their boundary (NE Correction), then possible pairs of relations are generated (Relation Candidate Generation) and only candidate relations are extracted (Relation Candidate Filtering). Relation's classes are verified by human using a correct answer constructing system (Relation Candidate Verification). Features are extracted (Features Extraction) and learnt by a classification model (RE Learning) such as Decision Tree (DT:C4.5), Random Forest (RF), Naïve Bayes (NB) and Rule Learner (RL). In testing stage, texts in a document are changed to be in the form of corrected named entity tags. Candidate relations are extracted and classified by a model from the training stage. For the purpose of evaluation, we apply five-fold cross validation with stratified sampling technique.

Features. Focusing on a pair of entities, a feature extraction process is applied to identify relation between two entities. In this work, these two entities are called preceding entity (PE) and succeeding entity (SE), respectively. To simplify the task with a reasonable constraint, we limit PE and SE to be located in the same paragraph in a document. To find relation, context in Thai language has been exploited. In this work, three sets of features are investigated, token space (SPC), number of named entities (nENT), and entity type (eTYP).

Relation Types. In this work, four types of relations are: (1) person and action [PER-ACT], (2) action and location [ACT-LOC], (3) location and action [LOC-ACT], and (4) action and person [ACT-PER]. From the example of named entities in Fig. 1, the relation among them can be investigated as in Table 1.

Table 1. Relation identification

Example	Relation Type	Class
[PER01:นายอาคม][ACT01:ฆ่า] [PER01:Mr. Akom][ACT01:killed]	PER-ACT	YES
[PER01:นายอาคม][ACT02:ได้หลบหนีไป] [PER01:Mr. Akom][ACT02:escaped]	PER-ACT	YES
[PER02:นายสมคิด][ACT02:ได้หลบหนีไป] [PER02:Mr. Somkid][ACT02:escaped]	PER-ACT	NO
[PER03:นายอาคม][ACT02:ได้หลบหนีไป] [PER03:Mr. Akom][ACT02:escaped]	PER-ACT	YES
[ACT01:ฆ่า][PER02:นายสมคิด] [ACT01:killed][PER02:Mr. Somkid]	ACT-PER	YES
[ACT01:ฆ่า][PER03:นายอาคม] [ACT01:killed][PER03:Mr. Akom]	ACT-PER	NO
[ACT01:ฆ่า][LOC01:รัชดาผับ] [ACT01:killed][LOC01:Ratchada Pub]	ACT-LOC	YES
[ACT01:ฆ่า][LOC02:จังหวัดราชบุรี] [ACT01:killed][LOC02:Ratchaburi Province]	ACT-LOC	NO
[LOC01:รัชดาผับ][ACT02:ได้หลบหนีไป] [LOC01:Ratchada Pub][ACT02:escaped]	LOC-ACT	NO

Table 2. The experimental results for the relation of ACT-LOC, LOC-ACT, ACT-PER and PER-ACT

Model	Feature	Accuracy (%)			
		ACT-LOC	LOC-ACT	ACT-PER	PER-ACT
DT	SPC	55.87	68.37	55.80	57.71
	eTYP	74.30	74.30	61.92	66.90
	nENT	69.99	68.14	64.16	65.50
	SPC+eTYP	77.41	86.69	69.87	66.39
	SPC+nENT	78.45	86.24	65.38	62.22
	eTYP+nENT	69.99	91.81	66.40	72.96
	SPC+eTYP+nENT	80.98	92.72	72.71	72.09
RF	SPC	56.45	66.22	55.80	54.59
	eTYP	75.33	78.95	63.55	70.54
	nENT	68.65	68.83	66.40	66.40
	SPC+eTYP	80.84	89.19	72.71	70.89
	SPC+nENT	82.62	89.99	75.35	64.31
	eTYP+nENT	68.65	93.86	78.00	77.48
	SPC+eTYP+nENT	**89.75**	**95.91**	**86.36**	**81.46**
NB	SPC	55.56	60.98	58.45	55.97
	eTYP	71.63	68.49	59.00	67.42
	nENT	65.54	66.89	63.96	59.10
	SPC+eTYP	70.74	70.42	63.96	65.34
	SPC+nENT	66.28	69.28	61.92	62.90
	eTYP+nENT	65.54	80.09	63.94	66.38
	SPC+eTYP+nENT	74.29	79.17	64.98	66.55
RL	SPC	56.61	67.92	55.58	57.02
	eTYP	71.77	73.03	60.90	68.12
	nENT	70.29	69.96	64.98	63.43
	SPC+eTYP	74.29	85.21	66.19	67.58
	SPC+nENT	74.15	80.54	67.42	65.33
	eTYP+nENT	70.29	85.21	67.62	72.62
	SPC+eTYP+nENT	76.23	90.67	69.45	74.36

4 Experiment and Results

Our named entity extraction aims at processing Thai news documents in the category of crime, collected from Thai three news publishers: Kom Chad Luek[1], Daily News Online[2], and Manager Online[3]. From a raw text, named entities are first detected and relations between pairs of named entity are identified. The experiment is done with 1736 pairs of relations in Thai news documents (crime). The experimental results from Table 2 shows that the performance of extracting relation of LOC-ACT is better than ACT-LOC, ACT-PER and PER-ACT. RF is the classifier which is preciser than DT, RL and NB, respectively.

RF with the combination of features SPC+nENT+eTYP is the best for the relation extraction of action-location, location-action, action-person and person-action with accuracy 89.75%, 95.91%, 86.36% and 81.46%, respectively. The classification model with the most combination of features is better than the model with the less combination of features.

5 Conclusion and Future Works

This paper presents a feature-based approach for extracting relations among entities from Thai news documents. Four supervised learning schemes, decision tree, rule learner, random forest and naïve Bayes are applied to investigate the performance of relation extraction using different feature sets. Focusing on four different types of relations in crime-related news documents, the experimental result shows that the random forest with the combination of three features achieves up to an accuracy of 95%. Generating more pairs of relation, and extracting relation automatically are our future works. Other news categories will be investigated.

Acknowledgments. This work has been supported by the Thailand Research Fund (TRF), under Grant No. BRG50800013 and a TRF Royal Golden Jubilee Ph.D. Program, under Grant No. PHD/0034/2550, and was supported by the National Electronics and Computer Technology Center, Thailand, under Grant No. NT-B-22-I4-38-49-05.

References

1. Zhu, J., Gonçalves, A., Uren, V., Motta, E., Pacheco, R.: Corder: Community relation discovery by named entity recognition. In: Proceedings of the 3rd int'l conference on Knowledge capture (K-CAP 2005), pp. 219–220. ACM, New York (2005)

[1] http://www.komchadluek.net

[2] http://www.dailynews.co.th

[3] http://www.manager.co.th

2. Hasegawa, T., Sekine, S., Grishman, R.: Discovering relations among named entities from large corpora. In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL 2004), Morristown, NJ, USA, ACL, p. 415 (2004)
3. Rosenfeld, B., Feldman, R.: Clustering for unsupervised relation identification. In: Proceedings of the sixteenth ACM conference on information and knowledge management (CIKM 2007), pp. 411–418. ACM, New York (2007)
4. Kawtrakul, A., Suktarachan, M., Varasai, P., Chanlekha, H.: A state of the art of thai language resources and thai language behavior analysis and modeling. In: Proceedings of the 3rd workshop on Asian language resources and int'l standardization (COLING 2002), Morristown, NJ, USA, ACL, pp. 1–8 (2002)
5. Tongtep, N., Theeramunkong, T.: Pattern-based named entity extraction for thai news documents. In: Proceedings of the 3rd Int'l Conference on Knowledge, Information and Creativity Support Systems (KICSS 2008), December 22-23, 2008, pp. 82–89 (2008)

An Incremental-Learning Method for Supervised Anomaly Detection by Cascading Service Classifier and ITI Decision Tree Methods

Wei-Yi Yu and Hahn-Ming Lee

Department of Computer Science and Information Engineering National Taiwan
University of Science and Technology Taipei, 106, Taiwan, R.O.C
{M9315912,hmlee}@mail.ntust.edu.tw

Abstract. In this paper, the incremental learning method to cascade Service Classifier and ITI (incremental tree inducer) methods for supervised anomaly detection, called "SC+ITI", is proposed for classifying anomalous and normal instances in a computer network. Since the ITI method can not handle new instances with new service value, the SC+ITI cascading method is proposed to avoid this. Two steps are in SC+ITI cascading methods. First, the Service Classifier method partitions the training instances into n service clusters according to different service value. Second, in order to avoid handling instances with new service value, the ITI method is trained with instances with the same service value in the cluster. In 2007, Gaddam et al. showed KMeans+ID3 cascading method which mitigates two problems 1) the Forced Assignment problem and 2) the Class Dominance problem. His method with Nearest Neighbor (NN) combination rule outperforms the other three methods (i.e., K-Means, ID3 and KMeans+ID3 with Nearest Consensus rule) over the 1998 MIT-DARPA data set. Since the KDD'99 data set was also extracted from the 1998 MIT-DARPA data set, Nearest Neighbor combination rule within K-Means+ITI and SOM+ITI cascading methods is used in our experiments. We compare the performance of SC+ITI with the K-Means, SOM, ITI, K-Means+ITI and SOM+ITI methods in terms of the Detection Rate and False Positive Rate (FPR) over the KDD'99 data set. The results show that the ITI method have better performance than the K-Means, SOM, K-Means+ITI and SOM+ITI methods in terms of the overall Detection Rate. Our method, the Service Classifier and ITI cascading method outperforms the ITI method in terms of the Detection Rate and FPR and shows better Detection Rate as compared to other methods. Like the ITI method, our method also provides the additional options of handling missing values data and incremental learning.

Keywords: anomaly detection system (ADS), K-Means clustering, Kohonens' self-organizing maps (SOM), ITI (incremental tree inducer), KDD'99.

1 Introduction

The intrusion detection systems (IDS) can be commonly classified into two categories according to the modeling methods used. One is misuse detection or rule-based

H. Chen et al. (Eds.): PAISI 2009, LNCS 5477, pp. 155–160, 2009.
© Springer-Verlag Berlin Heidelberg 2009

method that uses stored signatures of known intrusion instances to detect a malicious attack with low false-positive error. However this technique is hard to detect novel attacks and variants of known attacks whose rules are not stored. The other one is anomaly detection method that analyzes large amount of data to model a normal profile and attempts to identify patterns of activity that deviate from the defined profile. Although it remedies the problem of detecting novel attacks, the drawback of this technique is that normal behavior deviating from the defined profile may be labeled as an intrusion, resulting in high false-positive error.

In this paper, the incremental learning Service Classifier and ITI (SC+ITI) cascading method is proposed for classifying anomalous and normal instances. Since the ITI method can not handle new instances with new service value, the SC+ITI cascading method guarantees the ITI method is trained with instances with same service value. The SC+ITI cascading method has three phases which are described in section 2.2.

In 2007, Gaddam et al. [3] presented the novel method cascading the clustering method (K-Means) [4] with the decision tree (ID3) learning method [8] called "KMeans+ID3" which alleviates two problems in the cluster: 1) the Forced Assignment problem and 2) the Class Dominance problem. The first problem, Forced Assignment arises when similar anomaly and normal instances are assigned to the cluster. The second problem, Class Dominance arises in the cluster when subset of training data in the cluster contains an amount of instances from one particular class and few instances from the remaining classes. Since Gaddam et al. presented KMeans+ID3 cascading method which mitigates two problems: 1) the Forced Assignment problem and 2) the Class Dominance problem, SC+ITI cascading method is evaluated with the performance of K-Means, SOM , ITI, K-Menas+ITI, SOM+ITI methods using two measures (Detection Rate and False Positive Rate) in this paper.

The rest of the paper is organized as follow: In Section 2, we briefly discuss the K-Means, SOM, ITI, K-Means+ITI, SOM+ITI and SC+ITI learning-based anomaly detection methods. In Section 3, we discuss experiments, data sets and measures. In Section 4, we discuss the results of above six methods. We conclude our work and propose future work in section 5.

2 Methodologies for Anomaly Detection

Since anomaly detection with the K-Means [4], SOM [5], K-Means+ITI, SOM+ITI methods were quite similarly discussed in [3], these methods for anomaly detection will not be described in this section. In the section, we only briefly discuss the ITI 10 and SC+ITI methods for supervised anomaly detection. Nearest Neighbor combination rule within the K-Means+ITI and SOM+ITI cascading methods is adopted here instead of Nearest-Consensus rule because of two reasons: 1) Gaddam's K-Means+ID3 cascading method with Nearest Neighbor (NN) combination rule [3] outperform the other proposed methods over the 1998 MIT-DARPA data set 2) Nearest Consensus of the K-Means and ID3 cascading method probably doesn't exist if user defined parameter f is too small.

2.1 Anomaly Detection with the ITI Decision Tree

After trained with instances, the ITI method will build the binary decision tree. For detecting anomalies, ITI method outputs binary classification of "0" to indicate normal and "1" to indicate anomaly class. This is quite similarly to Gaddam's anomaly detection with ID3 which is described in [3]. We choose the ITI method because of four reasons: 1) inconsistent training instances 2) missing values data 3) incremental learning 4) numeric variables. Inconsistent training instances, missing values and incremental learning had been discussed in [10]. Numeric variables and limitations are discussed as follows:

2.1.1 Numeric Variables

For handling numeric variables, there are differences between the Gaddam's ID3 and ITI decision tree methods. In Gaddam's ID3 method, the training space was discretized into n equal-width intervals where n is predefined. Fayyad's splitting point selection [2] for numeric variables is adopted here in the ITI algorithm to deal with this problem.

2.1.2 Limitations

Two limitations were discussed in [6]. First, the ITI method needs to have "sufficient" training instances that cover as much variation of the normal behavior as possible. Second, this method can not handle new instances with new (i.e., unseen service) class labels. During the incremental-learning phase, this method can not incorporate instances with new service value into the binary decision tree and update the rules incrementally. However, at the testing phase, this method can be tested with these instances. Instances with new service value for the test at the decision node of service attribute will be passed down the false branch. Details were described in [6].

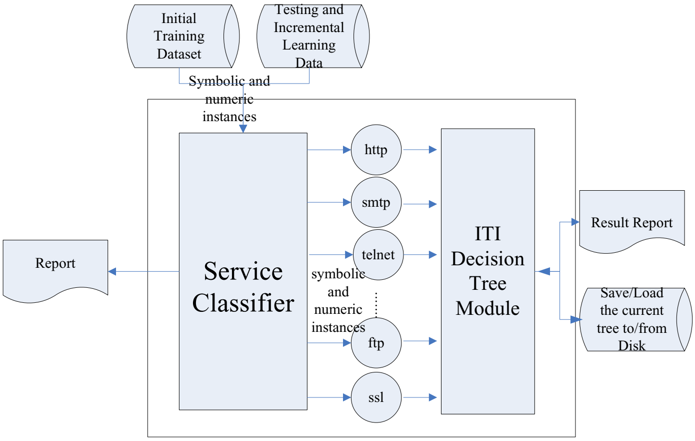

Initial Training Dataset
Testing and Incremental Learning Data
Symbolic and numeric instances
Service Classifier
http
smtp
telnet
symbolic and numeric instances
ftp
ssl
ITI Decision Tree Module
Report
Result Report
Save/Load the current tree to/from Disk

Fig. 1. The Service Classifier and ITI cascading method for Anomaly Detection

2.2 Anomaly Detection Using Service Classifier (SC) and ITI

Since the ITI method can not handle new instances with (i.e., unseen service) class labels, the service classifier and ITI cascading method is proposed to avoid this. Since Service Classifier ensures the ITI method is trained with instances with only one service value, the ITI method is trained with instances incrementally without the attribute service. The cascading method has three phases: 1) training, 2) testing and 3) incremental learning. During first training phase, the service classifier method is first applied to partition the training data set into m disjoint clusters according to different service. Then, the ITI method is trained with the instances in each cluster. The Service Classifier method ensures that each training instance is associated with only cluster. The testing phase, we will find the cluster to which the test instances are belong. Then, the ITI method is tested with the instances. At the last incremental learning phase, the Service Classifier and ITI cascading method will not be re-trained again and uses incremental learning data to train the existing ITI binary tree. The architecture of the Service Classifier and ITI cascading method is described in Fig. 1.

Two limitations are discussed. First, the cascading method needs to have "sufficient" training instances that cover as much variation of the normal behavior as possible. Second, the cascading method can not be tested with instances with new service value at the testing phase.

3 Experiments, Data Set and Measures

3.1 Experimental Setup and Data Set

SOM_PAK [11] and ITI [10] packages are adopted to evaluate their performance here. In our experiments, the k value of the K-Means method was set to 10, m*n value of the SOM method were set to 5*5 and 41 features of KDD’99 data set were all selected here. Neighborhood Parameters of SOM are Gaussian and Hexagonal.

Although a critique of the 1998 and 1999 DARPA IDS evaluations was discussed in [7], the KDD’99 data set [12] is commonly used for comparing the performance of IDSs. Four reasons to choose the KDD’99 data set were discussed in [9]. Other reason we choose this data set is noisy instances (ex: inconsistent training instances, error service value) occur in the KDD’99 data set. That represents real data in the reality. In our experiments, we simulate Sarasamma’s training data set which was described in [9]. 169,000 instances from the “10% KDD” data set and 311,029 instances from “Corrected KDD” (Test Set) data set were used for training and testing respectively. The training and test set consist of 22 and 39 attack types respectively which fall into four main categories: Denial of Service (DOS), Probe, Remote to User (R2L), and User to Root (U2R).

3.2 Performance Measures

Anomaly intrusion detection is a two-class classification problem. For each single prediction, there are four possible outcomes. The true-positives and true-negatives are correct classifications. A false-positive occurs when IDS/ADS classifies an instance as an anomaly when it is a normal instance. Measures such as False Positive Rate, Detection Rate, Accuracy, Precision and F-measure are defined in [1]. Other measures to compute the Area under an ROC (Receiving Operating Characteristic) curve, called AUC, mentioned in [1]. In our experiments, two measures (Detection Rate and FPR) are used.

4 Results

In this section, we present the results of K-Means, SOM, ITI, K-Means+ITI, SOM+ITI and SC+ITI methods over the KDD'99 data set. Table 1 summarizes the Detection Rate of these methods for five categories. The last two rows in Table 1 represent the Overall Detection Rate and FPR of these methods individually. The ITI and SC+ITI methods have better performance than K-Means, SOM, K-Means+ITI and SOM+ITI methods in terms of Detection Rate on U2R, R2L and PROBE attacks. The overall Detection Rate of the ITI and SC+ITI methods is better than other four methods, but the overall FPR of these methods is less than that of the ITI, and SC+ITI methods. In Table 1, we notice that:

- The overall Detection Rate of the K-Means, SOM and cascading K-Means+ITI and SOM+ITI methods is 89.95%, 85.97%, 91.31%, 91.07% respectively. The overall FPR of the K-Means, SOM and cascading K-Means+ITI and SOM+ITI methods is 1.29%, 1.49%, 0.81%, 0.73% individually. Since cascading methods mitigate the Class Dominance and Forced Assignment problems, cascading methods outperform the individual clustering methods in terms of the overall Detection Rate and FPR.

Table 1. Detection rate of attack catgory, overall detection rate and false positive rate of the methods

Method Category(Count)	KMeans	SOM	ITI	KMeans +ITI	SOM +ITI	SC+ITI
U2R(228)	55.70%	20.61%	75.44%	27.63%	42.98%	61.40%
R2L(16,189)	0.10%	0.03%	20.61%	3.96%	4.60%	21.22%
PROBE(4,166)	73.91%	71.80%	92.70%	85.14%	84.52%	95.15%
NORMAL(60,593)	98.71%	98.50%	98.14%	99.19%	99.26%	98.20%
DOS(229,853)	96.60%	92.35%	97.44%	97.64%	97.33%	97.65%
Overall Detection Rate	89.95%	85.97%	92.38%	91.31%	91.07%	92.63%
False Positive Rate	1.29%	1.49%	1.86%	0.81%	0.73%	1.80%

5 Conclusion and Future Work

In the Table 1, K-Means+ITI and SOM+ITI cascading methods outperform the individual K-Means and SOM clustering methods in terms of the overall Detection Rate and FPR respectively because they alleviate two problems: 1) the Forced Assignment problem and 2) the Class Dominance problem. The ITI method has better performance than these above four methods in terms of the overall Detection Rate. The SC+ITI cascading method shows better overall Detection Rate and FPR as compared to the ITI method. We conclude our work. For detecting anomalies, the incremental-learning SC+ITI cascading method shows better Detection Rate as compared to other methods and provides the additional options of handling missing values data and incremental learning.

Our future work includes 1) comparing performance of Gaddam's K-Means+ID3, 2) results of the methods (ex: ITI) tested with instances with new service value in terms of the Detection Rate and FPR, 3) statistical evaluation, 4) comparing performance of the different versions of ID3 or other decision trees, and 5) comparing performance of the incremental learning or multi-level classifier and decision tree (ex: ID4) cascading methods.

Acknowledgments. This work was supported in part by the National Digital Archive Program-Research & Development of Technology Division (NDAP-R&DTD), the National Science Council of Taiwan under grants NSC 95-2422-H-001-007, NSC 95-2422-H-001-024, and also by the Taiwan Information Security Center (TWISC), the National Science Council under grants NSC 95-2218-E-001-001, and NSC 95-2218-E-011-015.

References

1. Fawcett, T.: An introduction to ROC analysis. Pattern Recognition Letters 27(8), 861–874 (2006)
2. Fayyad, U.M., Irani, K.B.: On the handling of continuous-valued attributes in decision tree generation. Machine Learning 8(1), 87–102 (1992)
3. Gaddam, S.R., Phoha, V.V., Balagani, K.S.: K-Means+ID3: A novel method for supervised anomaly detection by cascading k-Means clustering and ID3 decision tree learning methods. IEEE Transactions on Knowledge and Data Engineering 19(3), 345–354
4. Hartigan, J.A., Wong, M.A.: A K-Means clustering algorithm. Applied Statistics 28(1), 100–108 (1979)
5. Kohonen, T.: The self-organizing map. Neurocomputing 21(1-3), 1–6 (1998)
6. Lee, W., Stolfo, S.J., Mok, K.W.: Adaptive Intrusion Detection: A Data Mining Approach. Artificial Intelligence Review 14(6), 533–567 (2000)
7. McHugh, J.: Testing intrusion detection systems: A critique of the 1998 and 1999 DARPA intrusion detection system evaluations as performed by Lincoln laboratory. ACM Transactions on Information and System Security 3(4), 262–294 (2000)
8. Quinlan, J.R.: Induction of decision trees. Machine Learning 1(1), 81–106 (1986)
9. Sarasamma, S.T., Zhu, Q.A.: Min-Max Hyperellipsoidal Clustering for Anomaly Detection in Network Security. IEEE Transactions on systems, man, and cybernetics-part B: Cybernetics 36(4), 887–901 (2006)
10. Utgoff, P.E., Berkman, N.C., Clouse, J.A.: Decision Tree Induction Based on Efficient Tree Restructuring. Machine Learning 29, 5–44 (1997)
11. Kohonen, T., Hynninen, J., Kangas, J., Laaksonen, J.: SOM_PAK: The Self-Organizing Map Porgram Package, http://www.cis.hut.fi/research/som_lvq_pak.shtml
12. Stolfo, S., et al.: The Third International Knowledge Discovery and Data Mining Tools Competition (2002), http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

Quantifying News Reports to Proxy "Other Information" in ERC Models

Kuo-Tay Chen[*], Jian-Shuen Lian, and Yu-Ting Hsieh

Department of accounting
College of Management
National Taiwan University
1 Roosevelt Road, Sec. 4, Taipei, Taiwan, 106
ktchen@management.ntu.edu.tw

Abstract. Many previous studies have investigated how earning announcement affects stock price. They measure the effect by employing earning response coefficient (ERC) models. However, the traditional models did not explicitly consider textual information received by investors. Rather they simply referred to it as "other information". However, investor's exposure to textual information (e.g. news report) might have significant influence on how stock prices will respond to earning announcements. This study attempts to investigate whether earning surprises cause stock fluctuations and how the effect is influenced by news coverage prior to earning announcements. We find that: (1) earning surprise significantly affects stock price; (2) more news coverage tends to decrease the ERC; (3) positive earning surprises have higher influence on stock price; and (4) different combinations of news sentiment and earning surprise result in different ERC.

Keywords: earning response coefficient (ERC) model, textual information, news coverage, news sentiment.

1 Introduction

Following Ball and Brown [1] and Fama et al. [2] many studies have investigated the effect of earning announcements on stock prices reaction. Most commonly, they measure the effect by relating a security's abnormal market return to the unexpected component of reported earning of the firm issuing that security [3]. This measure usually is referred to as "earning response coefficient" or simply ERC.

Over the past few decades, ERC-related research has progressed tremendously. These studies not only help to identify and explain the differential market response to earnings information, but also build up lots of useful approaches and models to investigate the relationship between earnings announcements and stock returns. However, these traditional ERC models have had a noticeable drawback – their explanation powers (namely, the R^2) are very low.

[*] Corresponding author.

H. Chen et al. (Eds.): PAISI 2009, LNCS 5477, pp. 161–168, 2009.
© Springer-Verlag Berlin Heidelberg 2009

The drawback might have resulted from the omission of textual information in the models. For example, financial statement footnotes or and concurrent news reports can certainly provide useful information for investors to predict a firm's future earning and affect their judgment about the firm's reasonable stock price. In addition, being exposed to new reports prior to earning announcement might have reduced the surprise shock to the investors, thus lowering the ERC. Traditional ERC models have failed to include these kinds of textual information. In this study, we attempt to partially rectify this critical omission.

> The main goal for this study attempts to investigate whether earning surprises cause stock fluctuations and how the effect is influenced by news coverage prior to earning announcements. We find that: (1) earning surprise significantly affects stock price; (2) negative earning surprises have lower influence on stock prices; (3) more news coverage tends to decrease ERC; and (4) different combinations of news sentiment and earning surprise result in different ERC.

This paper is organized as follows. Section 2 conducts literature review and develop hypothesis. Section 3 describes the research design. Section 4 explains the analysis results. Finally, Section 5 presents our conclusions and future research directions.

2 Literature Review and Hypothesis Development

Traditionally, earning response coefficient (ERC) is used to represent stock market response to earning announcement, which is measured by regressing cumulative abnormal return (CAR) on earning surprise. Previous studies [1, 5, 6, 7, 8, 9, 10, 11] have concluded that stock price changes move in the same direction as earnings changes; however, stock market responses differ between different kinds of firms. These studies identified four economic determinants of ERC variations: persistence of earning changes, risk, growth opportunities, and interest rate [7, 8, 9, 10, 11]. They found that ERC is higher when the unexpected current earnings changes are more persistent or the firm's growth opportunities are higher [8, 9, 10, 11]. Additionally, ERC is lower when firm's risk or risk-free interest rate is higher [9, 11]. Besides these four economic determinants, there are still others causes for differential market response, such as, capital structure [12], earnings quality [9,13,14], similarity of investor expectations [15], and informativeness of price [9,11], etc.

The main theme of this study is related to the "informativeness of price" argument. Before being officially announced, information related to a firm's earning might have been disclosed through various news reports, e.g. the firm has received a major order. Investors would have incorporated the information into their determination of reasonable stock prices; hence the stock price would reflect the prediction of the firm's earning. By the time earning is officially announced, market reaction would be less drastic as the stock price has already reflected some information of the announced earning. Based on this price-lead-earning logic, ERC will be lower if more information is "leaked" through news report prior to earning announcement. Following the aforementioned literature and the information leakage logic, we have developed the following hypotheses:

H1: A firm's announced earning surprise is positively related to the abnormal return of its stock.

H2: If a firm has higher news coverage before earning announcement, its ERC will be lower.

Other than the amount of leaked information, the direction of earning surprise might also affect ERC. When there is a positive earning surprise, people tend to jump on the bandwagon and rush to buy the stock. On the other hand, when this is a negative earning surprise, people are more reluctant to sell the stock due to the over-commitment bias. Therefore, we posit the following hypothesis:

H3: The ERC for positive earning surprises will be higher than that for negative earning surprises.

The ERC might also be affected by the directional combination of news sentiment and earning surprise. Arguably, if news sentiment and earning surprise are in the same direction, then the impact of earning surprise on stock price will be lower. For example, if the sentiment of prior news report is positive, but the announced actual earning is lower than the predicted earning (i.e. a negative earning surprise), then the ERC will be higher than that when both are positive. Therefore, we posit the following hypothesis:

H4: The ERC is lower when news sentiment and earning surprise are in the same direction; it is higher when news sentiment and earning surprise are in opposite direction.

3 Research Design

3.1 The Regression Model

Our analysis is based on the following simple regression model:

CAR=a + b x SUE

Where CAR: Cumulative abnormal return in (-2,2) window around earnings announcement dates
SUE: Earnings surprise based on IBES reported analyst forecasts and actual earnings.
b: the regression coefficient is interpreted as the ERC.

We define the earnings surprise (SUE) as actual earnings minus analyst forecasts earnings, scaled by stock price. We follow the method of Livnat and Mendenhall (2006) to calculate SUE:

$$SUE_{jt} = \frac{(X_{jt} - \bar{X}_{jt})}{P_{jt}}$$

Where X_{jt} is primary Earnings Per Share (EPS) before extraordinary items for firm j in quarter t, P_{jt} the price per share for firms j at the end of quarter t, and $\bar{X}_{jt}$ it the median of forecasts reported to I/B/E/S in the 90 days prior to the earnings announcement.

The daily abnormal returns are calculated as the raw daily return retrieved from the Center for Research in Security Prices (CRSP) minus the daily return on the portfolio of firms with approximately the same size and book-to-market (B/M) ratio. The daily returns for the size and B/M portfolios are from Professor Kenneth French's data library, which is based on classification of the population into six (two size and three B/M) portfolios.

3.2 The Data Collection

We follow the following procedure to collect data:

1. **Collect News Articles**: we collect news articles in the ***Wall Street Journal*** during the time period from Aug. 1999 to Feb. 2007. In total, there are 321993 articles.
2. **Collect Quarterly reports (10-Q reports) and report announcement dates**: We collect quarterly reports data of the company which has at least one news article in the field of 1999 to 2007 from **Compustat**, and use the rdq field (Report Date of Quarterly Earnings, the date on which quarterly earnings per share are first publicly reported) as the report announcement date.
3. **Calculate SUE**(Earnings surprise based on IBES reported analyst forecasts and actual): We collect data from IBES to calculate the **quarterly** SUE.
4. **Calculate CAR**(Cumulative abnormal return in (-2,2) window around earnings announcement dates): We use **CRSP** daily and monthly data to calculate the cumulative abnormal return.
5. **Identify and total positive and negative words in news articles:** Following Tetlock(2008), we identified words in news articles as positive or negative word by General Inquirer dictionary.
6. **Calculate news coverage:** We calculate the number of firm's news articles in the 30 days period prior to its quarterly report announcement date as the firm's news coverage. The mean of news coverage is 3.
7. **Determine the final sample:** We collect **26200** firm-quarter data sets and remove those with SUE in the upper and lower 0.05% from the sample. The final sample consists of **25940** firm-quarter data sets.
8. **Classify high and low news coverage groups:** We divided the sample into high and low news coverage groups based on the number of news coverage. If a data set has more than 3 (including 3) news articles, it is classified as high news coverage, otherwise, low news coverage.
9. **Classify positive and negative SUE groups:** We divide the sample into positive and negative SUE groups based on the sign of SUE's. There are 20047 data sets in the positive group and 5893 data sets in the negative group.
10. **Classify positive and negative sentiment groups**: We also divided the sample into positive and negative sentiment groups by the sentiment of their news articles. If the number of negative words in an article is greater than the number of positive words, we classify the news article as a negative news; otherwise, a positive news. Then, if a data set has more negative news articles than negative news article, it is classified as a negative sentiment data set; otherwise, a positive sentiment data set.

3.3 Hypotheses Testing

We use the following procedure to analyze our data and validate the hypotheses. Firstly, for H1, we use the entire data sets to run the regression model to see whether b is statistically significant.

For H2, we run two versions of the regression model; one for the high news coverage group, the other for the low new coverage group. We then compare the ERCs for these two models to determine whether they are significantly different.

For H3, we also run two versions of the regression model; one for the positive SUE group, the other for the negative SUE group. We then compare the ERCs for these two models to determine whether they are significantly different.

For H4, we run four versions of the regression model, one for each directional combination of sentiment and SUE. We then compare the ERCs for these four models to determine whether they are significantly different. There are four combinations: positive sentiment / positive SUE, positive sentiment / negative SUE, negative sentiment / positive SUE, and negative sentiment / negative SUE.

4 The results

Table1 shows the statistical result for hypothesis H1. Since b, namely the ERC, is significantly, we can conclude that earning surprises affect stock prices. That is, H1 is confirmed.

Table 2 shows the statistical results for hypothesis H2. The ERC for the high news coverage group is 2.56186, while the ERC for the low news coverage group is 3.22156. Since 2.56186 is apparently lower than 3.22156, H2 is also confirmed.

Table 1. Regression results for H1

CAR= a + b * SUE	ERC (b)	3.14290 (t=27.39 , P<0.0001)
	Adjusted R-square	0.0281
	Sample size	25942

Table 2. Regression results for H2

High news coverage	ERC	2.56186 (t=8.31 , P<0.0001)
	Adjusted R-square	0.0153
	Sample size	4390
Low news coverage	ERC	3.22156 (t=25.98 , P<0.0001)
	Adjusted R-square	0.0303
	Sample size	21552

Table 3 shows the statistical results for hypothesis H3. The ERC for the positive earning surprise group is 3.18871, while the ERC for the negative earning surprise group is 0.53500. Since 3.18871 is apparently higher than 0.53500, H3 is also confirmed.

Table 4 shows the statistical results for hypothesis H4. The ERC is 3.20904 for the positive news / positive earning surprise combination, 0.31425 for the positive news / negative earning surprise combination, 2.90845 for the negative news / positive earning surprise combination, and 2.21731 for the negative news / negative earning surprise combination. Therefore, for the negative news group, H4 is confirmed. However, for the positive news group, the result is opposite to H4.

Table 3. Regression results for H3

SUE>=0	ERC	3.18871 (t=18.03 , P<0.0001)
	Adjusted R-square	0.0159
	Sample size	20048
SUE<0	ERC	0.53500 (t=2.51 , P=0.0122)
	Adjusted R-square	0.0009
	Sample size	5894

Table 4. Regression results for H4

	SUE>=0		SUE<0	
Positive news	ERC	3.20904 (t=17.45 , P<0.0001)	ERC	0.31425 (t=1.39 , P=0.1634)
	Adjusted R^2	0.0159	Adjusted R^2	0.0002
	Sample size	18726	Sample size	5400
Negative news	ERC	2.90845 (t=4.52 , P<0.0001)	ERC	2.21731 (t=3.25 , P=0.0012)
	Adjusted R^2	0.0145	Adjusted R^2	0.019
	Sample size	1322	Sample size	494

5 Conclusions and Future Research Directions

This study attempts to investigate the influence of textual information on the ERC. We find that: (1) earning surprise significantly affects stock price; (2) more pre-announcement news coverage tends to decrease the ERC; (3) positive earning surprises have higher influence on stock price; and (4) different combinations of news sentiment and earning surprise result in different ERC. Since the last finding is partially opposite to our hypothesis, our future study will try to find out the reason. Also, we will try to test whether adding textual information to the ERC models can provide more explaining power.

References

1. Ball, R., Brown, P.: An Empirical Evaluation of Accounting Income Numbers. Journal of Accounting Research 6, 159–178 (1968)
2. Fama, E.F., Fisher, L., Jensen, M., Roll, R.: The adjustment of stock prices to new information. International Economic Review 10, 1–21 (1969)
3. Scott, W.R.: Financial Accounting Theory, 4th edn. Pearson Prentice Hall, Toronto (2006)
4. Tetlock, P.C., Saar-Tsechansky, M., Macskassy, S.: More Than Words: Quantifying Language to Measure Firms' Fundamentals. Journal of Finance 63, 1437–1467 (2008)
5. Beaver, W.: The Information Content of Annual Earnings Announcements. Journal of Accounting Research, 67–92 (1968)
6. Beaver, W., Clarke, R., Wright, W.: The Association Between Unsystematic Security Rerurns and the Magnitude of the Earnings Forecast Error. Journal of Accounting Research, 316—340 (1979)
7. Kothari, S.P.: Capital markets research in accounting. Journal of Accounting and Economics 31, 105–231 (2001)
8. Kormendi, R.C., Lipe, R.: Earnings Innovations, Earnings Persistence, and Stock Returns. Journal of Business, 323—346 (1987)
9. Easton, P.D., Zmijewski, M.E.: Cross-Sectional Variation in the Stock-Market Response to Accounting Earnings Announcements. Journal of Accounting and Economics, 117–141 (1989)
10. Ramakrishnan, R.T.S., Thomas, J.k.: Valuation of Permanent, Transitory and Price-Irrelevant Components of Reported earnings. Working paper, Columbia University Business School (1991)
11. Collins, D.W., Kothari, S.P.: An Analysis of the Intertemporal and Cross-Sectional Determinants of Earnings Response Coefficients. Journal of Accounting and Economics, 143—181 (1989)
12. Dhaliwal, D.S., Lee, K.J., Fargher, N.L.: The Association Between Unexpected Earnings and Abnormal Security Returns in the Presence of Financial Leverage. Contemporary Accounting Research, 20–41 (1991)
13. Dechow, P.M., Dichev, I.: The Quality of Accruals and Earnings: The Role of Accrual Estimation errors. The Accounting Review, 35–59 (2002)
14. Francis, J., LaFond, R., Olsson, P., Schipper, K.: Costs of Equity and Earnings Attributes. The Accounting Review, 967–1010 (2004)
15. Abarbanell, J.S., Lanen, W.N., Verrecchia, R.E.: Analysts' forecasts as Proxies for Investor Beliefs in empirical Research. Journal of Accounting and Economics, 31–60 (1995)
16. Beaver, W., Lambert, R., Morse, D.: The information content of Security Prices. Journal of Accounting and Economics 2, 3–28 (1980)
17. Foster, G., Olsen, C., Shevlin, T.: Earnings Releases, Anomalies, and the Behavior of Security Returns. The Accounting Review, 574–603 (1984)
18. Bernard, V.L., Thomas, J.: Post-Earnings Announcement Drift: Delayed Price Reaction or Risk Premium? Journal of Accounting Research, 1–36 (1989)

19. Ball, R., Bartov, E.: How Naïve Is the Stock Market's Use of Earnings Information? Journal of Accounting and Economics, 319–337 (1996)
20. Bartov, E., Radhakrishnan, S., Krinsky, S.: Investor Sophistication and Patterns in stock Returns after Earnings Announcements. The Accounting Review, 43–63 (2000)
21. Brown, L.D., Han, J.C.Y.: Do Stock Prices Fully Reflect the Implications of Current Earnings for Future Earnings for ARI firms? Journal of Accounting Research, 149–164 (2000)
22. Fama, E.F., French, K.R.: Common risk factors in the returns of stocks and bonds. Journal of Financial Economics 33, 3–56 (1993)
23. Livnat, J., Mendenhall, R.R.: Comparing the Post-Earnings Announcement Drift for Surprises Calculated from Analyst and Time Series Forecasts. Journal of Accounting Research 44, 177–205 (2006)
24. Fama, E.F., French, K.R.: The Cross-section of Expected Stock Returns. Journal of Finance 47, 427–465 (1992)

Author Index