Christopher C. Yang et al.

LNCS 5075

Intelligence and Security Informatics

IEEE ISI 2008 International Workshops:
PAISI, PACCF, and SOCO 2008
Taipei, Taiwan, June 2008, Proceedings

Lecture Notes in Computer Science 5075

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Christopher C. Yang Hsinchun Chen
Michael Chau Kuiyu Chang
Sheau-Dong Lang Patrick S. Chen
Raymond Hsieh Daniel Zeng Fei-Yue Wang
Kathleen Carley Wenji Mao Justin Zhan (Eds.)

Intelligence and Security Informatics

IEEE ISI 2008 International Workshops:
PAISI, PACCF, and SOCO 2008
Taipei, Taiwan, June 17, 2008
Proceedings

Springer

Volume Editors

Christopher C. Yang, E-mail: chris.yang@ischool.drexel.edu

Hsinchun Chen, E-mail: hchen@eller.arizona.edu

Michael Chau, E-mail: mchau@business.hku.hk

Kuiyu Chang, E-mail: kuiyu.chang@pmail.ntu.edu.sg

Sheau-Dong Lang, E-mail: lang@cs.ucf.edu

Patrick S. Chen, E-mail: chenps@ttu.edu.tw

Raymond Hsieh, E-mail: hsieh@cup.edu

Daniel Zeng, E-mail: zeng@email.arizona.edu

Fei-Yue Wang, E-mail: feiyue.wang@gmail.com

Kathleen Carley, E-mail: kathleen.carley@cs.cmu.edu

Wenji Mao, E-mail: wenji.mao@ia.ac.cn

Justin Zhan, E-mail: justinzh@andrew.cmu.edu

Library of Congress Control Number: 2008928272

CR Subject Classification (1998): H.4, H.3, C.2, H.2, D.4.6, K.4-5, K.6.5

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

ISSN 0302-9743
ISBN-10 3-540-69136-7 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-69136-5 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12278909 06/3180 5 4 3 2 1 0

Preface from the Pacific Asia Workshop on Intelligence and Security Informatics (PAISI 2008) Chairs

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 USA. 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. These past 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 2008 Pacific Asian Workshop on ISI (PAISI 2008) to especially provide a stimulating forum for ISI researchers in Pacific Asia and other regions of the world to exchange ideas and report research progress.

June 2008

Christopher C. Yang
Hsinchun Chen
Michael Chau
Kuiyu Chang
Daniel Zeng
Fei-Yue Wang

Preface from the Pacific Asia Workshop on Cybercrime and Computer Forensics (PACCF 2008) Chairs

As personal computers and access to the Internet become more prevalent, modern society is becoming increasingly dependent on the computer and networking technology for storing, processing, and sharing data, and for email and message communication. Cybercrime in the broadest sense refers to any criminal activity in which computers or networks play an essential role, where computers may be used as a tool to commit a crime, as the victim, or may contain evidence of a crime. Examples of cybercrime include: possession of illegal digital materials, spreading virus, worm, or malicious code, email spamming, hacking, ID theft, economic espionage, information warfare, etc. Law enforcement and government agencies, corporate IT officers, and software vendors have worked together to assemble forensic computing tools, incident response policies, and best practices to train and fight against the surge of this new crime wave.

The 2008 Pacific Asia Workshop on Cybercrime and Computer Forensics (PACCF 2008) provided a forum for professionals in the computer forensics community and IT security industry, forensic computing software vendors, corporate and academic researchers, to disseminate ideas and experiences related to forensic computing especially in the context of cybercrime investigation. The PACCF 2008 Workshop received high-quality papers dealing with topics in cybercrimes and computer (digital) forensics. The workshop organizers would like to thank contributing authors for their interest in the workshop and the Program Committee members for their effort and assistance in reviewing the papers and providing feedback to the authors.

June 2008

Sheau-Dong Lang
Raymond Hsieh
Patrick S. Chen

Preface from the Workshop on Social Computing (SOCO 2008) Chairs

Social computing can be broadly defined as computational facilitation of social studies and human social dynamics as well as the design and use of information and communication technologies that consider the social context. In recent years, social computing has become one of the central themes across a number of information and communications technology (ICT) fields and has attracted significant interest from not only researchers in computing and social sciences, but also software and online game vendors, Web entrepreneurs, political analysts, and digital government practitioners, among others.

The First Workshop on Social Computing (SOCO 2008) brought together social computing researchers to address a wide range of methodological and application-driven topics, employing research methods from both computational sciences and social sciences. The one-day workshop program included 12 long papers, 11 short papers, and 2 posters. The co-hosts of SOCO 2008 were the University of Arizona, the Chinese Academy of Sciences, and Carnegie Mellon University.

June 2008

Daniel Zeng
Fei-Yue Wang
Kathleen Carley
Weiji Mao
Justin Zhan

Organization

Pacific Asia Workshop on Intelligence and Security Informatics (PAISI 2008)

Organizing Committee

Honorary Co-chairs
Hsinchun Chen, The University of Arizona
Feiyue Wang, Chinese Academy of Science

Workshop Co-chairs
Chris Yang, Drexel University
Daniel Zeng, The University of Arizona

Program Co-chairs
Kuiyu Chang, Nanyang Technological University, Singapore
Michael Chau, The University of Hong Kong, Hong Kong

Program Committee

Clive Best, Joint Research Centre, European Commission, Italy
Robert W.P. Chang, Central Police University, Taiwan
Patrick S. Chen, Tatung University, Taiwan
Reynold Cheng, Hong Kong Polytechnic University, Hong Kong
Yiuming Cheung, Hong Kong Baptist University, Hong Kong
Vladimir Estivill-Castro, Griffith University, Australia
Eul Guy Im, Hanyang University, Korea
Kai Pui Lam, Chinese University of Hong Kong, Hong Kong
Wai Lam, Chinese University of Hong Kong, Hong Kong
Mark Last, Ben-Gurion University of the Negev, Israel
Ickjai Lee, James Cook University, Australia
You-lu Liao, National Central Police Univeresity, Taiwan
Ee-peng Lim, Nanyang Technological University, Singapore
Duen-Ren Liu, National Chiao-Tung University, Taiwan
Hongyan Liu, Tsinghua University, China
Xin (Robert) Luo, Virginia State University, USA
Anirban Majumdar, University of Auckland, New Zealand
Byron Marshall, Oregon State University, USA
Jialun Qin, The University of Massachusetts, USA
Dimitri Roussinov, Arizona State University, USA
Raj Sharman, State University of New York, Buffalo, USA
Andrew Silke, University of East London, UK
David Skillicorn, Queen's University, Canada
Aixin Sun, Nanyang Technological University, Singapore

Alan Wang, Virginia Tech University, USA
Fu Lee Wang, City University of Hong Kong, Hong Kong
Jau-Hwang Wang, National Central Police University, Taiwan
Shiuh-Jeng Wang, National Central Police University, Taiwan
Chih-Ping Wei, National Tsinghua University, Taiwan
Jennifer Xu, Bentley College, USA
Justin Zhan, Carnegie Mellon CyLab, Japan
Yilu Zhou, George Washington University, USA
William Zhu, The University of Auckland, New Zealand

Pacific Asia Workshop on Cybercrime and Computer Forensics (PACCF 2008)

Program Co-chairs

Sheau-Dong Lang, University of Central Florida, USA
Raymond Hsieh, California University of Pennsylvania, USA
Patrick S. Chen, Tatung University, Taiwan

Program Committee

Fahim Akhter, College of Information Technology, United Arab Emirates
Mohammed Arif, The British University in Dubai, United Arab Emirates
Feng-Cheng Chang, National Chiao Tung University, Taiwan
Wei Feng Chen, California University of Pennsylvania, USA
Hongmei Chi, Florida A&M University, USA
Hsiang-Cheh Huang, National University of Kaohsiung, Taiwan
San Kon Lee, Korean University of Technology and Education, South Korea
Dana J. Lesemann, Stroz Friedberg, LLC, USA
Yuh-Yih Lu, Minghsin University of Science and Technology, Taiwan
Muammer Ozer, City University of Hong Kong, Hong Kong
Lucille M. Ponte, Florida Coastal School of Law, USA
Daniel Purcell, Seminole County Sheriff's Office, USA
Chang-Lung Tsai, Chinese Culture University, Taiwan
Fredy Valenzuela, The University of New England, Australia
Jau-Hwang Wang, Central Police University, Taiwan
Mian Zhou, Bank of America, USA
Cliff Zou, University of Central Florida, USA

Workshop on Social Computing (SOCO 2008)

Organizing Committee

Workshop Co-chairs
Daniel Zeng, University of Arizona and Chinese Academy of Sciences
Fei-Yue Wang, Chinese Academy of Sciences
Kathleen Carley, Carnegie Mellon University

Program Co-chairs
Weiji Mao, Chinese Academy of Sciences
Justin Zhan, Carnegie Mellon University

Program Committee
Elisabeth Andre, University of Augsburg, Germany
Shizuo Asogawa, Cylab, Japan
Matt-Mouley Bouamrane, University of Manchester, UK
Longbin Cao, University of Technology Sydney, Australia
Guoqing Chen, Tsinghua University, China
Rui Chen, Institute of Policy and Management, Chinese Academy of Sciences, China
Xueqi Cheng, Institute of Computing Technology, Chinese Academy of Sciences, China
Chaochang Chiu, Yuanze University, Taiwan
Ruwei Dai, Institute of Automation, Chinese Academy of Sciences, China
Jonathan Gratch, University of Southern California, USA
Wendy Hall, University of Southampton, UK
Xiangpei Hu, Dalian University of Technology, China
Jenny Huang, AT&T and IFOSSF, USA
Zan Huang, Pennsylvania State University, USA
Changjun Jiang, Tongji University, China
Hady Wirawan Lauw, Nanyang Technological University, Singapore
Churn-Jung Liao, Academia Sinica, Taiwan
Huan Liu, Arizona State University, USA
Robert Lusch, University of Arizona, USA
Michael Shaw, University of Illinois at Urbana-Champaign, USA
Ron Sun, Rensselaer Polytechnic Institute, USA
Da-Wei Wang, Academia Sinica, Taiwan
Yingxu Wang, University of Calgary, Canada
Jennifer Xu, Bentley College, USA
Ning Zhong, Maebashi Institute of technology, Japan

International Conference on Intelligence and Security Informatics (ISI 2008)

Organizer

Central Police University, Taiwan

Co-organizers

Institute of Information Science, Academia Sinica, Taiwan
National Taiwan University, Taiwan
National Taiwan University of Science and Technology, Taiwan
University of Arizona, USA

Conference Advisors

Pei-Zen Chang, The Science and Technology Advisory Group of Executive Yuan, Taiwan
Tai-Lang Chien, Ministry of Interior, Taiwan
Wei-Hsien Wang, Science Research Office, National Security Bureau, Taiwan

Conference Chair

Ing-Dan Shieh, Central Police University, Taiwan

Conference Co-chairs

Hsinchun Chen, University of Arizona, USA
Shi-Shuenn Chen, National Taiwan University of Science and Technology, Taiwan
Der-Tsai Lee, Institute of Information Science, Academia Sinica, Taiwan
Si-Chen Lee, National Taiwan University, Taiwan

Program Committee Co-chairs

Wun-Hwa Chen, National Taiwan University, Taiwan
Wen-Lian Hsu, Academia Sinica, Taiwan
Tzong-Chen Wu, National Taiwan University of Science and Technology, Taiwan
Christopher C. Yang, The Chinese University of Hong Kong, Hong Kong
Daniel Zeng, University of Arizona, USA

Organizing Committee Co-chairs

Frank F.Y. Huang, Central Police University, Taiwan
Yuan-Cheng Lai, National Taiwan University of Science and Technology, Taiwan
I-An Low, Central Police University, Taiwan
Jau-Hwang Wang, Central Police University, Taiwan

Local Arrangement Committee Co-chairs

You-Lu Liao, Central Police University, Taiwan
Guan-Yuan Wu, Central Police University, Taiwan
Kou-Ching Wu, Central Police University, Taiwan

Publication Chair

Min-Yuh Day, Academia Sinica, Taiwan

Treasurers

Robert Weiping Chang, Central Police University, Taiwan
Chung-Yin Huang, Central Police University, Taiwan

Conference Secretaries

Lynn Chien, Central Police University, Taiwan
Jiunn-Cherng Chiou, Central Police University, Taiwan
Tai-Ping Hsing, Central Police University, Taiwan
Jia-Hung Huang, Central Police University, Taiwan
Yungchang Ku, Central Police University, Taiwan
Ya-Ren Teng, Central Police University, Taiwan
Chia-Chen Yang, Central Police University, Taiwan
Wen-Chao Yang, Central Police University, Taiwan
Chih-Ping Yen, Central Police University, Taiwan

Hosts and Sponsors

Central Police University, Taiwan
National Taiwan University, Taiwan
National Taiwan University of Science and Technology, Taiwan
Institute of Information Science, Academia Sinica, Taiwan
Drexel University, USA
The Chinese University of Hong Kong, Hong Kong
The University of Arizona, USA
Tatung University, Taiwan
California University of Pennsylvania, USA
University of Central Florida, USA
Chinese Academy of Sciences, China
Carnegie Mellon University, USA

Table of Contents

Pacific Asia Workshop on Intelligence and Security Informatics (PAISI 2008)

Information Retrieval and Event Detection

Internet Security and Cybercrime

Currency and Data Protection

Cryptography

Image and Video Analysis

Privacy Issues

Social Networks

Modeling and Visualization

Network Intrusion Detection

Pacific Asia Workshop on Cybercrime and Computer Forensics (PACCF 2008)

Forensic Information Management

Forensic Technologies

Forensic Principles and Tools

Workshop on Social Computing (SOCO 2008)

Social Web and Social Information Management

Social Networks and Agent-Based Modeling

Social Opinions, E-Commerce, Security and Privacy Considerations

Chinese Word Segmentation for Terrorism-Related Contents

Daniel Zeng[1,2], Donghua Wei[1], Michael Chau[3], and Feiyue Wang[1,2]

[1] Institute of Automation, Chinese Academy of Sciences, China
[2] The University of Arizona, Tucson, Arizona, USA
[3] The University of Hong Kong, Hong Kong
zeng@email.arizona.edu, donghuawei@gmail.com,
mchau@business.hku.hk, feiyue.wang@ia.ac.cn

Abstract. In order to analyze security and terrorism related content in Chinese, it is important to perform word segmentation on Chinese documents. There are many previous studies on Chinese word segmentation. The two major approaches are statistic-based and dictionary-based approaches. The pure statistic methods have lower precision, while the pure dictionary-based method cannot deal with new words and are restricted to the coverage of the dictionary. In this paper, we propose a hybrid method that avoids the limitations of both approaches. Through the use of suffix tree and mutual information (MI) with the dictionary, our segmenter, called IASeg, achieves a high accuracy in word segmentation when domain training is available. It can identify new words through MI-based token merging and dictionary update. In addition, with the Improved Bigram method it can also process N-grams. To evaluate the performance of our segmenter, we compare it with the Hylanda segmenter and the ICTCLAS segmenter using a terrorism-related corpus. The experiment results show that IASeg performs better than the two benchmarks in both precision and recall.

Keywords: Mutual information, N-gram, suffix tree, Ukkonen algorithm, Heuristic rules, Lidstone flatness.

1 Introduction

Extremists and terrorists have been using the Internet to spread their ideology and recruit new members. It is important for government and anti-terrorist organizations to analyze such online information in order to enhance national and international security. Previous research has reported on how to collect and analyze relevant documents from the Internet (e.g., Web pages, blogs, newsgroup postings) in English and it has been shown possible to extract extremist or terrorist information and their relationships from these documents (Chen & Xu, 2006; Chau & Xu, 2007). However, little work has been done in the analysis of extremist or terrorist related Web documents in Chinese. In this paper we propose a method that combines mutual information and suffix tree to address the word segmentation problem in Chinese document analysis. We apply the proposed method to perform word segmentation on

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 1–13, 2008.

a terrorism-related corpus. The rest of the paper is structured as follows. Section 2 reviews related work in Chinese word segmentation. In Section 3 we describe our proposed algorithm based on mutual information and suffix tree. Section 4 reports the results of our evaluation study, in which we tested our proposed algorithm using a terrorism-related data set. We discuss the findings in Section 5 and conclude our study in Section 6.

2 Related Work

Chinese word segmentation has been studied for many years, but two problems in word segmentation, namely unknown word identification and ambiguity parsing, are still not completely solved. Studies on Chinese word segmentation can be roughly divided into two categories: heuristic dictionary-based methods, and statistical machine learning methods. Readers are referred to (Wu et al., 1993) for a more detail survey. In the following, we review previous research in each category.

2.1 Dictionary-Based Methods

Dictionary-based methods mainly employ a predefined dictionary and some hand-generated rules for segmenting input sequence. These rules can be generally classified based on the scanning direction and the prior matching length. The Forward Matching Method (FMM), the input string will be scanned from the beginning to the end and matched against dictionary entries. In the Reverse Matching Method (RMM), the input string will be scanned from the end to the beginning. Bidirectional Matching Methods scan the input string from both directions. The matching length can be based on maximum matching or minimum matching. Most popular dictionary-based segmenters use a hybrid matching method. The main disadvantage of dictionary-based methods is that their performance depends on the coverage of the lexicon, which unfortunately may never be complete because new words appear constantly. Consequently, these methods cannot deal with the unknown words (sometimes called Out-Of-Vocabulary or OOV) identification and may result in wrong segmentation.

2.2 Statistical and Machine Learning Methods

Statistical methods rely on different measure to decide on the segmentation boundary in Chinese. Sun et al.(2004) use a liner function of mutual information (MI) and difference of t-test to perform text segmentation. Many researchers also concentrate on two-character words (bigrams), because two is the most common length in Chinese words. Dai et al. (1999) use contextual and positional information, and found that contextual information is the most important factor for bigram extraction. They found that positional frequency is not helpful in determining words. Yu et al. (2006) proposed a cascaded Hidden Markov Model (HMM) for location and organization identification. Other researchers, such as Jia et al. (2007), Xue et al. (2003), and Low et al. (2005) focus on the Maximum Entropy (ME) models. Li et al. (2002) use Expectation Maximization and Maximum Likelihood Prediction to deal with Chinese word segmentation. Zhou & Liu (2002) construct state chart using the information of

whether several characters can compose one word and propose an algorithm to generate candidate words. Sproat et al. (1996) propose a Stochastic Finite-State Word-Segmentation method by combining character states with dictionary-based heuristic rules. There are several others: Hockenmaier and Brew (1998) present an algorithm, based on Palmer's (1997) experiments, that applies a symbolic machine learning technique to the problem of Chinese word segmentation. Many other statistic-based machine learning methods have been used in Chinese word segmentation, such as SVM-based segmentation (Li et al., 2001), the CRF method segmentation (Peng, et al., 2004), unsupervised models (Creutz et al, 2007).

Ponte and Croft (1996) introduce two models for word segmentation: word-based and bigram-based models. Both utilize probabilistic automata. In the word-based method, a suffix tree of words in the lexicon is used to initialize the model. Each node is associated with a probability, which is estimated by segmenting training text using the longest match strategy. This makes it easy to apply the segmenter to new languages. The bigram model uses the lexicon to initialize probability estimates for each bigram, and the probability with which each bigram occurs, and uses the Baum-Welch algorithm (Rabiner 1989) to update the probabilities as the training text is processed.

Some researchers concentrate on the named entity identification, such as Sproat et al. (1996), who developed special-purpose recognizers for Chinese names (and translated foreign names). They implemented special recognizers not only for Chinese names and transliterated foreign names, but also for components of morphologically obtained words.

As we know, pure dictionary-based methods rely on the coverage of the dictionaries, and general statistical methods require segmented training corpus. Teahan et al. (2000) proposed using text compression model to do word segmentation. This method uses neither manual dictionary nor training corpus, but just uses finite-context models of characters to predict the upcoming one. Each prediction takes the form of a probability distribution that is provided to an encoder. The conditional probability distribution of characters, conditioned on the preceding few characters, is maintained and updated as each character of input is processed.

In sum, all methods have to rely on either character-level information indicated by the co-occurrence probability, conditional probability, position status, or word-level information provided by the dictionary or the language knowledge, such as the part-of-speech, morphological, syntactic and semantic knowledge (Cui et al., 2006). Many researchers combine the available information and achieved better performance in both unsupervised learning (Peng and Schuurmans, 2001) and supervised learning (Teahan et al., 2000).

3 Proposed Algorithm

In this paper, Mutual Information (MI) and Suffix Tree are combined to perform Chinese word segmentation. While Ponte and Croft (1996) just deal with bigrams, we focus more on segmentation of trigrams and longer words. We first use a training corpus to train the bigram model and use a lexicon to establish the improved bigram model. We then use MI and the improved bigram model combining with the Suffix Tree to parse the given text.

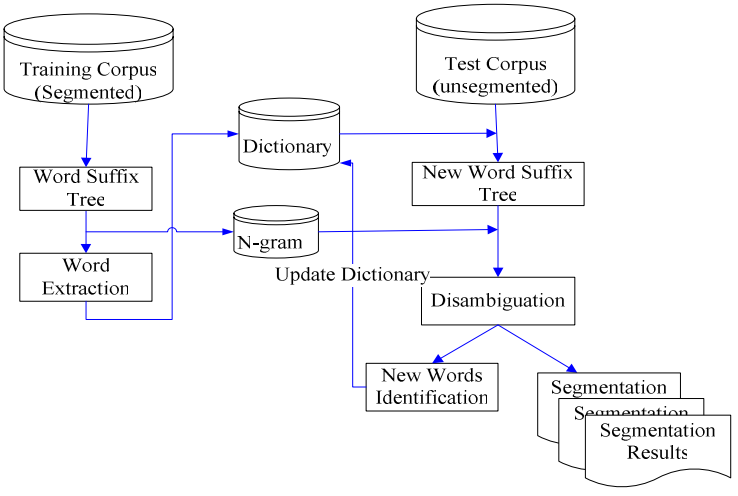


Fig. 1. Overall architecture of our system

In this Section, we describe our proposed algorithm, called the IASeg. We separate the segmentation process into two phases - the training phase and the test phase. The overall algorithm is shown in Figure 1. In the training phase, we construct a dictionary and the N-grams, which include the Unigram, Bigram and the Improved Bigram. In the test phase, we first split the input string into tokens using dictionary-based FMM heuristic. Then we calculate the strings' Mutual Information to predict unknown words and decide whether we should merge the two adjacent tokens as one new word. If the formation of the new word is supported, we can update the dictionary dynamically. Finally, we output the segmentation results.

3.1 N-Gram Construction

The n-gram word model is based on the Markovian assumption. If the n-th character is related only with its preceding (N-1) characters of context, and without any other correlations, we call it the N-gram word model or (N-1)-order Markov model.

Given a string $w_1w_2\ldots w_n$, with its length being n, we have the following equations.

In unigram, we have.:

$$p(w_1w_2\ldots w_n)= p(w_1)\, p(w_2)\ldots\, p(w_n)$$

Using bigram (1-order Markov model), we have:

$$p(w_1w_2\ldots w_n)= p(w_1)\, p(w_2 \mid w_1)\, p(w_3 \mid w_2)\ldots p(w_n \mid w_{n-1})$$

And using trigram (2-order Markov model), we have:

$$p(w_1w_2\ldots w_n)= p(w_1)\, p(w_2 \mid w_1)\, p(w_3 \mid w_1w_2)\, p(w_4 \mid w_2w_3)\ldots p(w_n \mid w_{n-2}\, w_{n-1})$$

For an N-gram model, we have:

$$p(w_1w_2\ldots w_n)= p(w_1)\, p(w_2 \mid w_1)\, p(w_3 \mid w_1w_2)\, p(w_4 \mid w_1w_2w_3)\ldots p(w_n \mid w_1\ldots w_{n-1})$$

The above equations can be expressed by one equation:

$$p(w_1w_2\ldots w_n)= \prod_{i=1}^{n} p(w_i \mid w_{i-(N-1)}\, w_{i-(N-1)+1}\, w_{i-(N-1)+2}\ldots w_{i-1})$$

Table 1. Expressions in equations

Expression	Meaning
w_i	a character
$length(str)$	Number of characters in *str*
w_i^j	simple expression of string of $w_i w_{i+1} \ldots w_j$, $i < j$
$p(w)$	probability of string w in a given corpus
$count(w_1 w_2 \ldots w_n)$	frequency of n-gram $w_1 w_2 \ldots w_n$ in a given corpus
$p(x, y)$	probability of co-occurrence of x, y
$f(x)$	frequency estimate of x
N	number of training instances

where $n = length(w_1 w_2 \ldots w_n)$, and N is the length of N-grams to be considered. Table 1 gives a summary of the expressions used.

To make calculation simpler we often use the following parameter evaluation equations, based on their relative frequency, using a Maximum Likelihood Estimation (MLE) method.

$$p_{MLE}(w_n / w_{n-1}) = \frac{count(w_{n-1} w_n)}{count(w_{n-1})}$$

$$p_{MLE}(w_n / w_1^{n-1}) = \frac{count(w_1 w_2 w_3 \ldots w_n)}{count(w_1 w_2 w_3 \ldots w_{n-1})}$$

$$p_{MLE}(w_n / w_{n-N+1}^{n-1}) = \frac{count(w_{n-N+1}^{n-1} w_n)}{count(w_{n-N+1}^{n-1})}$$

For an N-gram model a large number of parameters need to be estimated. Many previous studies have focused on bigrams only because of computational efficiency considerations. In this study, we use an Improved Bigram to deal with longer grams. More details will be discussed in the following subsection.

3.2 MI Measure

3.2.1 Basic Concepts and MI Calculation Equation

The concept of Mutual Information comes from <Information Theory>, which indicates two events' dependence (compactness) using:

$$MI(x, y) = log_2 \frac{p(x, y)}{p(x) p(y)}$$

In Natural Language Processing (NLP), $MI(x,y)$ is used to estimate the *compactness* of two characters: x, y. If $MI(x,y)$ is higher than a given threshold value (usually estimated through experiments, denoted μ), we can regard them as one word. To simplify calculation, we define $p_{MLE}(x) = f(x) = count(x)/N$, so we can rewrite the MI equation as follows:

$$MI(x, y) = log_2 \frac{count(x, y) \cdot N}{count(x) count(y)}$$

Its full meanings are as followings: (On condition that the characters *a*, *b* both have the normal distribution in text, their dependence equals to their correlation.)

1. If $MI(a, b) >> 0$, i.e. $count(a, b) \cdot N >> count(a) \cdot count(b)$, the characters have positive correlation. And if $MI(a, b) > \mu$, we can regard the string '*ab*' as a word.

2. If $MI(a, b) << 0$, i.e. $count(a, b) \cdot N << count(a) \cdot count(b)$, the characters have negative correlation, and we do not regard '*ab*' as a word.

3. If $MI(a, b) \approx 0$, i.e. $count(a, b) \cdot N \approx count(a) \cdot count(b)$, then we say the characters have no correlation and they can't be viewed as a word either.

Researchers(Fang et al., 2005) also have used the following equation to calculate the *MI* of two bigrams *c* and *d*, each of them being a bigram string, and achieve better performance in such cases where a 4-character word is composed of 2 bigrams, e.g., "高音喇叭", in which both "高音" and "喇叭" are words.

$$MI(c,d) = log_2 \frac{N_c^2 f(c,d)}{N_w \times f(c) \times f(d)}$$

where N_c is the total characters in the corpus, N_w is the total number of the tokens in the corpus.

In our approach, however, we do not adopt this method because we have different classes of n-grams but just use one threshold. In order to come up with one measure standard, we using MLE calculation equations together with the *Lidstone* flatness algorithm to avoid the sparseness of the co-occurrences.

3.2.2 Flatness Algorithm

Most of the probabilities involved in the MI calculation are very small and can result in *zero* probability. To avoid numerical problems associated with these zero probabilities, we use the Lidstone flatness function:

$$P_{Lid}(w_1 ... w_n) = \frac{count(w_1 ... w_n) + \lambda)}{N + \lambda * B}$$

where $\lambda = 0.5$, B is the number of bins that the training instances are divided into (usually based on the number of dictionary items), and N is the corpus size (number of tokens).

For the forecast, we use

$$P(b/a) = \frac{count(a,b) + \lambda}{(count(a) + \lambda B)}, 0 < \lambda < 1$$

Especially, to deal with the probability of single word *w*, using following:

$$P(w) = \frac{count(w) + \lambda}{N + \lambda B}, 0 < \lambda < 1$$

on condition that *a*, *b*, w are non-empty strings.

In our research, we store the tokens and their frequencies. If we need to calculate their MI, we first retrieve the tokens and their frequencies, then calculate the MI using the equations described earlier.

3.2.3 Improved Bigram

In this subsection we describe our improved bigram model which is used to deal with n-gram words with $n \geq 3$. We store patterns using a hash table for MI computation with the unigram and simple bigram. This approach allows us to process multi-gram words, such as 4-gram words and 5-gram words, and even more parameters prediction.

We show an example of our improved Bigram in Figure 2. Consider the words: "东/土耳其/斯坦/信息/中心 (East Turkistan Information Center)". The MI of "土耳其" and "斯坦" will be calculated. As this is greater than the threshold, the frequency of the term "土耳其斯坦" will be stored and the MI of "东" and "土耳其斯坦" will be further calculated to obtain the correct term "东土耳其斯坦(East Turkistan) ".

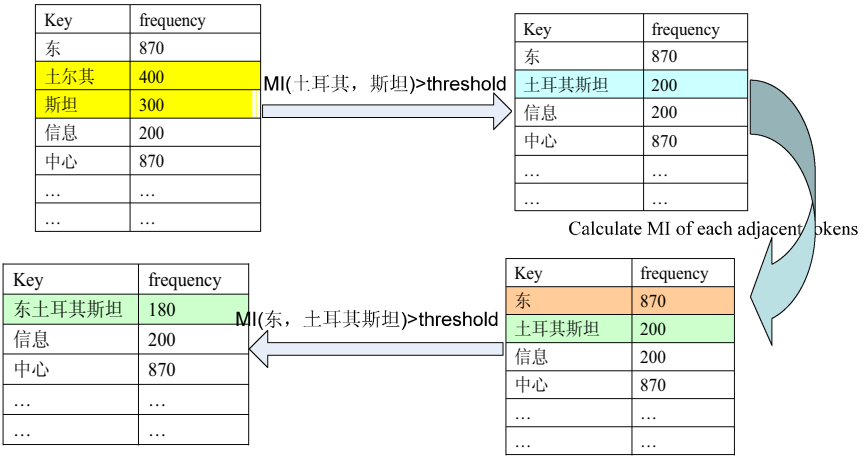


Fig. 2. Example of Improved Bigrams

This method is useful for segmenting terms that are named entities, such as combining "邓/小平" to "邓小平(Deng Xiaopeng)", and "中国/人民/银行" to "中国人民银行 (The People's Bank of China) ", etc. It can also deal with some ambiguous pairs, such as "新西兰花". For instance, if the training corpus are mostly about vegetables (i.e., having a lot of individual occurrences of "西兰花"), then it would split into "新/西兰花 (new broccoli)"; but if the corpus are about the country (i.e., having a lot of individual occurrences of "新西兰"), then it would be segmented as "新西兰/花 (New Zealand flowers)".

3.3 Suffix Tree

Given a string $S \in \Sigma^n$, the suffix tree T_S of S is the compacted Trie of all the suffixes of $S\$$, $\$ \notin \Sigma$. The suffix tree is the basic data structure in combinatorial pattern matching because of its many elegant uses. Furthermore, it has a compact $O(n)$ space representation that can be constructed in $O(n)$ optimal time for a constant-size alphabet (Weiner 1973). The original construction and its analysis are nontrivial.

Some efforts have been spent on producing simplified linear time algorithms (Chen and Seiferas 1985; McCreight 1976), though all such efforts have been variants of the original approach of Weiner. Due to limited space, readers are referred to other papers for the details of the details of the model and implementation of suffix tree (Chen and Seiferas 1985; Giegerich, et al., 1997; McCreight 1976; Weiner 1973; Zhang et al, 2004).

Suffix Tree has a low complexity (Chan et al., 2005). Let $C=\{T_1, T_2,..., T_k\}$ be a collection of texts with total length n. We can maintain a compressed suffix tree for C, which uses $O(n)$ space and supports the following queries about the suffix tree for C: finding the root in $O(1)$ time and finding the parent, left child, left sibling, right sibling, and suffix link of a node in $O(log\ n)$ time. The edge label and leaf label can be computed in $O(log\ n)$ time. Inserting or deleting of a text T in C can be done in $O(|T| log\ n)$ time.

How to construct a suffix tree efficiently is the critical problem of using suffix tree. Since *Weiner* proposed this data structure on 1973, a lot of work has been done in this area. The serial suffix tree construction is very mature and there are several well known algorithms which have linear complexity with respect to the length of the given string on both time and space. These algorithms include *Weiner*'s algorithm (1973), *McCreight*'s(1976) algorithm McC and *Ukkonen*'s algorithm (1992). Weiner's method is a monument on the character processing domain. McC uses suffix link technique to reduce the time complexity further and uses less space than the other two algorithms. The Ukkonen algorithm also uses the suffix link, and is an online arithmetic that builds suffix tree from left to right, i.e. adding t_{i+1} (a new character or label edge) to the current suffix tree $STree(T_n)$ and forms another new suffix tree $STree(T_{n+1})$, $0 < n \le |T|$, where $|T|$ is the length of the string. In this study we use the *Ukkonen* algorithm.

3.4 Lexicon Construction

Our algorithm uses the known words (dictionary) generated from the previous stage (training processing) to segment the test corpus through the FMM heuristic rule. In order to improve the efficiency of matching, we can either sort all words in the dictionary according to the frequency of words or first classify words according to their length and then sort by their frequencies. However, these methods cannot solve the problem fundamentally. Our system stores the dictionary using a **Trie** structure, and sorts the items according to the order of the Chinese characters based on their Pinyin Romanization.

The construction steps are as follows:

1. The entire dictionary is stored as a Forest;
2. Each tree contains all the words which have the same first character;
3. All the second characters of these words in the same tree are children of the root node;
4. Other characters follow the same token.

For example, the first character of all the following words have the same Pinyin "bao": "爆炸", "爆炸装置", "爆炸药", "爆炸性", "爆发性质", "爆发性", "爆裂", "爆裂声", "爆裂声响", "爆破", "爆破突击队", "爆破井", "暴戾".

3.5 Comparison with Existing Methods

Previous research has used mutual information for Chinese word segmentation. For example, both *Chien et al.* (1997) and *Ong et al.* (1999), utilize MI in their key phrase extraction. Our proposed algorithm is different from these existing studies. First, MI is used in different ways and different stages in our segmenter. Chien et al. (1997) first split a given string into tokens with different lengths and use MI to filter out the strings with an MI value lower than the threshold. Ong et al. (1999) extend Chien's work by suggesting an updateable PAT-tree that allows the update of string frequencies dynamically. Different from their methods, we first split a given string coarsely, then compute MI of the neighbor tokens, and compare the MI value with the threshold. If the MI value is higher, then we merge the tokens and add the new word into the dictionary. Otherwise, we keep them unmerged. Another major difference is that we use a hybrid approach. In the first stage, we perform coarse splitting using a dictionary-based method to split the given texts, while the other two methods directly compound the characters according to their compositions.

4 Experiments

In order to evaluate the performance of our IASeg system, we compare it with the Hylanda segmenter (www.hylanda.com) and the ICTCLAS segmenter (Zhang et al., 2003). The Hylanda segmenter is a dictionary-based segmenter that has been widely used in practice (e.g., the search engine Zhongsou). ICTCLAS is an HMM-based segmenter. Both segmenters were chosen because they have shown very good performance in previous studies.

We use precision, recall, and F-measure to evaluate the performance of the segmenters. The calculations are as follows:

$$precision = \frac{correctNum}{autoTotalNum} \qquad recall = \frac{correctNum}{manualTotalNum}$$

$$F-measure = \frac{2 \times recall \times precision}{recall + precision}$$

where *correctNum* is the number of words correctly identified by the automatic method, *autoTotalNum* is the total number of words identified by the automatic method, and *manualTotalNum* is the number of words identified in the manual segmentation. A perfect segmenter will have a recall and precision of 100%. All these measures can be calculated automatically from a machine-segmented text, along with the human-segmented gold standard.

We collected a set of terrorism-related documents from the Web using crawlers. Within our list of *seed Websites* for terrorism-related content, we crawled news content using our page filter, which discards irrelevant materials like the advertising anchor text and the fringe links, and just keep the news content. As a result, we obtained a set of 330 news articles.

With these data, we setup our own gold set and training set, and run the three segmenters on the corpus. In our algorithm, the dictionary trie is updated dynamically

after the first run of the training corpus. At last we obtained 57,339 words through the terrorism corpus, which also includes some high frequency general words.

The segmentation results of the three segmenters are shown in Table 2:

Table 2. Results of the three segmenters on terrorism-related content

	Hylanda	ICT	IASeg
Precision	0.8603	0.7759	0.9477
Recall	0.9160	0.8658	0.9480
F-measure	0.8874	0.8223	0.9477

Overall we can see that our segmenter achieves the best performance among the three segmenters in terms of precision, recall, and F-measure. Note that some of the ICT scores are *zero* because those tests could not be executed successfully due to, for example, problems with common Web page patterns like "……" or email addresses.

5 Discussion

Based on our testing of the segmenter on the terrorism-related corpus and other corpuses (not reported here), we found that two aspects of the training data have a profound influence on the model's accuracy. First, some errors are obviously caused by deficiencies in the training data, such as improperly segmented common words and name entities. Second, some errors stem from the topics covered by the corpus. It is not surprising that the error rate increases when the training and testing text represent different topic areas--such as training on military affairs news text and testing on medical text.

We observe that our algorithm has the following characteristics:

1. Using Mutual Information value as the new words identification threshold is greatly different from simple term frequency confidence.
2. Different thresholds will achieve different results. For example, a threshold of *20* may just keep all the tokens in their original form, while a threshold of *9* will result in merging some high co-occurrence adjacent tokens as one word. In general, we found that a lower threshold will make the segmenter to prefer longer words, thus resembling more closely with named entity extraction tools.
3. By using suffix tree, we can do searching and matching more easily and efficiently. Using the *Ukkonen* algorithm, we can construct the suffix tree in *O(n)* time complexity and *O(n)* space complexity.
4. Through our improved bigram structure, we can filter the low MI token-pairs, which greatly improves the boundary forecast accuracy.

6 Conclusion

In this paper, we propose a method on Chinese word segmentation based on suffix tree and mutual information. We integrate character-level information and word-level

information and achieve encouraging results in segmenting a terrorism-related corpus. Our algorithm uses a two-stage statistical word segmentation. In the first stage, word suffix tree are used to generate a dynamic dictionary and N-gram model on input text, and then a hybrid approach is employed in the second stage to incorporate word N-gram probabilities, and mutual information with word-formation patterns to detect Out-Of-Vocabulary words.

Our future work includes the following:

- Improve our strategies by adding more words' position information and part-of-speech to develop an integrated segmenter which can perform known word segmentation and unknown word identification at the same time.
- Address the OAS (overlap ambiguity string) problem using syntax rules and address the "CAS" (combination ambiguity string) problem using SVM classifier.
- Study the possibility of performing Chinese named entity recognition using the HMM-based tagger and its integration with this Chinese analyzer.
- Investigate the problem of event information extraction based on syntax structure.

Acknowledgments

This work is supported in part by NNSFC #60621001 and #60573078, MOST #2006CB705500, #2004CB318103, and #2006AA010106, CAS #2F05N01 and #2F07C01, and NSF #IIS-0527563 and #IIS-0428241. We thank our team member Qingyang Xu for his help with the experiments. We also thank Fenglin Li and Shufang Tang for their help with data preparation and processing.

References

1. Chan, H.L., Hon, W.K., Lam, T.W., Sadakane, K.: Dynamic dictionary matching and compressed suffix trees. Society for Industrial and Applied Mathematics, 13–22 (2005) ISBN:0-89871-585-7
2. Chau, M., Xu, J.: Mining Communities and Their Relationships in Blogs: A Study of Online Hate Groups. International Journal of Human-Computer Studies 65(1), 57–70 (2007)
3. Chen, H., Xu, J.: Intelligence and Security Informatics. Annual Review of Information Science and Technology 40, 229–289 (2006)
4. Chen, M.T., Seiferas, J.: Efficient and elegant subword-tree construction. In: Combinatorial Algorithm on Words, NATO Advanced Science Institutes. Series F, vol. 12, pp. 97–107. Springer, Berlin (1985)
5. Chien, L.F.: PAT-Tree Based Keyword Extraction for Chinese Information Retrieval. In: ACM SIGIR (1997)
6. Creutz, M., Lagus, K.: Unsupervised Models for Morpheme Segmentation and Morphology Learning. ACM Transactions on Speech and Language Processing 4(1) (January 2007)
7. Cui, S.Q., Liu, Q., Meng, Y., Yu, H.: Nishino Fumihito. New Word Detection Based on Large-Scale Corpus 43(05), 927–932 (2006)

8. Dai, Y.B., Khoo, S.G.T., Loh, T.E.: A new statistical formula for Chinese word segmentation incorporating contextual information. In: Proc. of the 22nd ACM SIGIR Conf. on Research and Development in Information Retrieval, pp. 82–89 (1999)
9. Fang, Y., Yang, H.E.H.: The Algorithm Design and Realization to Calculate The Mutual Information of Four- Word- String in Large Scale Corpus. Computer Development & Applications 1 (2005)
10. Giegerich, R., Kurtz, S.: From Ukkonen to McCreight and Weiner: A unifying view to linear-time suffix tree construction. Algorithmica 19, 331–353 (1997)
11. Hockenmaier, J., Brew, C.: Error-driven segmentation of Chinese. Communications of COLIPS 1(1), 69–84 (1998)
12. Jia, N., Zhang, Q.: Identification of Chinese Names Based on Maximum Entropy Model. Computer Engineering 33(9), 31–33 (2007)
13. Li, J.F., Zhang, Y.F.: Segmenting Chinese by EM Algorithm. Journal of the China Society for Scientific and Technical Information 03, 13–16 (2002)
14. Li, R., Liu, S.H., Ye, S.W., et al.: A method of crossing ambiguities in Chinese word segmentation based on SVM and k-NN. Journal of Chinese Information Processing 15(6), 13–18 (2001) (in Chinese)
15. Maaß, M.: Suffix Trees and their Applications. Ferienakademie 1999 Kurs 2: Bäume-Algorithmik and Kombinatorik (1999)
16. Low, J.K., Ng, H.T., Guo, W.: A Maximum Entropy Approach to Chinese Word Segmentation. In: Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing, Jeju Island, Korea, pp. 161–164 (2005)
17. McCreight, E.M.: A space-economical suffix tree construction algorithm. Journal of ACM 23(2), 262–272 (1976)
18. Ong, T.H., Chen, H.: Updateable PAT-Tree Approach to Chinese Key Phrase Extraction using Mutual Information: A Linguistic Foundation for Knowledge Management. In: Proceedings Asian Digital Library Conference, Taipei, Taiwan, pp. 63–84 (1999)
19. Palmer, D.: A trainable rule-based algorithm to word segmentation. In: Proceedings of the 35th Annual Meeting of the Association of Computational Linguistics, Madrid, Spain (1997)
20. Peng, F.C., Feng, F.F., McCallum, A.: Chinese segmentation and new word detection using conditional random fields. In: COLING 2004, Geneva, Switzerland (2004)
21. Peng, F.C., Dale, S.: Self-supervised Chinese Word Segmentation. In: Proceedings of the 4th International Symposium of Intelligent Data Analysis, pp. 238–247 (2001)
22. Ponte, J.M., Croft, W.B.: Useg: A retargetable word segmentation procedure for information retrieval. In: Proceedings of SDAIR 1996, Las Vegas, Nevada (1996)
23. Rabiner, L.: A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE 77(2), 257–286 (1989)
24. Sproat, R., Shih, C., Gale, W., Chang, N.: A stochastic finite-state word-segmentation algorithm for Chinese. Computational Linguistics 22(3), 377–404 (1996)
25. Sun, M.S., Xiao, M., Zou, J.Y.: Chinese Word Segmentation without Using Dictionary Based on Unsupervised Learning Strategy. Chinese Journal of Computers 27(6), 736–742 (2004)
26. Teahan, W.J., Wen, Y., McNab, R.J., Witten, I.H.: A compression-based algorithm for Chinese word segmentation. Computational Linguistics 26, 375–393 (2000)
27. Ukkonen, E.: Constructing Suffix Trees On-Line in Linear Time. In: Leeuwen, J.v. (ed.) Algorithms, Software, Architecture, Proc. IFIP 12th World Computer Congress, Information Processing 1992, Madrid, Spain, vol. 1, pp. 484–492 (1992)
28. Ukkonen, E.: On-line Construction of Suffix-Trees. Algorithmica 14(3) (1995)

29. Weiner, P.: Linear Pattern Matching Algorithms. In: Proc. 14th IEEE Annual Symp. on Switching and Automata Theory, pp. 1–11 (1973)
30. Wu, Z., Tseng, G.: Chinese text segmentation for text retrieval achievements and problems. JASIS 44(9), 532–542 (1993)
31. Xue, N.W., Chiou, F.-D., Palmer, M.: Building a large annotated Chinese corpus. In: The Proceedings of the 19th International Conference on Computational Linguistics, Taipei, Taiwan (2002)
32. Xue, N.W.: Chinese word segmentation as character tagging. International Journal of Computational Linguistics and Chinese Language Processing 8(1), 29–48 (2003)
33. Yu, H.K., Zhang, H.P., Liu, Q., Lv, X.Q., Shi, S.C.: Chinese named entity identification using cascaded hidden Markov model. Journal on Communications 27(2), 87–94 (2006)
34. Zhang, H.P., Yu, H.K., Xiong, D.Y., Liu, Q.: HMM-Based Chinese lexical analyzer ICTCLAS. In: Proc. of the 2nd SIGHAN Workshop, pp. 184–187 (2003)
35. Zhang, Ch.L., Hao, F.L., Wan, W.L.: An automatic and dictionary-free Chinese word segmentation method based on suffix array. Journal of Jilin University (Science Edition) 4 (2004)
36. Zhou, L.X., Liu, Q.: A Character-net Based Chinese Text Segmentation Method. In: SEMANET: Building and Using Semantic Networks Workshop, attached with the 19th COLING, pp. 101–106 (2002)

A Term Association Inference Model for Single Documents: A Stepping Stone for Investigation through Information Extraction

Sukanya Manna and Tom Gedeon

Department of Computer Science,
The Australian National University, Canberra, ACT, Australia
{sukanya.manna,tom.gedeon}@anu.edu.au

Abstract. In this paper, we propose a term association model which extracts significant terms as well as the important regions from a single document. This model is a basis for a systematic form of subjective data analysis which captures the notion of relatedness of different discourse structures considered in the document, without having a predefined knowledge-base. This is a paving stone for investigation or security purposes, where possible patterns need to be figured out from a witness statement or a few witness statements. This is unlikely to be possible in predictive data mining where the system can not work efficiently in the absence of existing patterns or large amount of data. This model overcomes the basic drawback of existing language models for choosing significant terms in single documents. We used a text summarization method to validate a part of this work and compare our term significance with a modified version of Salton's [1].

Keywords: Information retrieval, investigation, Gain of Words, Gain of Sentences, term significance, summarization.

1 Introduction

Information retrieval (IR) deals with text analysis, text storage, and the retrieval of stored records having similarity between them [2]. Among various IR models, vector based model is the significant one assigning weights based on the discriminative powers [3]. Inverse Document Frequency is the most common language model. But there are also modifications of the above concept into inverse sentence frequency and inverse term frequency, which all work over a large corpus to find a solution to the problem where document space language models do not work [4]. There are situations when the user query is not the only desired need but the relations between different contexts within a single text, which provide an insight into the semantic relations, might be of interest in some specific applications like official investigations, or counter terrorism, text summarization [5], question answering systems [6] and so on.

There are different computational models for natural language discourse structures, which are mainly used for summarization and question answering systems [7],[8], [9], [10]. In [11], the authors generate intra-document semantic hyperlinks and characterize the structure of a text based on the intra document linkage pattern. Again the concept of

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 14–20, 2008.

Latent Semantic Analysis [12] exploits knowledge induction and representation. A related concept to our work was analyzed by Rocha [13], where he presented keyword semantic proximity and its semi-metric behaviour in a recommendation system TalkMine to advance adaptive web and digital library technology.

IR following conventional predictive data mining techniques has proved to be ineffective in handling cases where there are no previous patterns of data available [14]. If we consider the act of terrorism, we do not find any similar indicia. With a relatively small number of attempts every year and only one or two major terrorist incidents every few years- each one distinct in terms of planning and execution- there are no meaningful patterns that show what behaviour indicates planning or preparation for terrorism. So, it is preferable to handle these types of scenarios with subjective data analysis or computational linguistic technologies to exploit the semantic and syntactic structure of texts.

Almost every document has some hierarchical structure concerning the importance of the words or concepts occurring in it [15]. The basic idea of linking the terms (entities + significant keywords) in a document is based on their frequency of occurring together in different paragraphs or sentences, presuming them to have some relationship. This approach does not require any previous knowledge about the data pattern. It is based on the degree of linkages found between different terms and brings out the relevant ones.

2 Motivation

In the previous section we have already mentioned that predictive data mining is not that useful to analyze cases like terrorism [14], or social crimes. Trained officials need to analyze every witness statements to find some clues to assume a possible solution to solve a legal problem. The basic objective of our work is to enhance the performance of these people and make their work easier in getting a solution.

There are several works related to information extraction, but the established models [3], [12], and [13] mainly deal with huge corpora for their analysis. Hence, it is challenging to work with a single document or very few documents to extract the most important facts and create a possible network to find patterns between different discourse segments within the text.

3 Term Significance Models

3.1 Modification of Salton's Indexing Method for Choosing Significant Terms in Single Documents

In this section we have modified Salton's [1] term discrimination model in such a way so that the documents in his model refers to the sentences in our version. Instead of calculating the similarities between the document pairs, we calculated here the similarity between the sentence pairs respectively. Our main aim of calculating the discrimination value was to identify the significant terms.

Let sentences be the discourse structure in this case. So, similarity between sentences is calculated by,

$$sim(s_i, s_j) = \sum_{k=1}^{t} w_{ik} w_{jk} \tag{1}$$

where, $t = no.\ of\ \ attributes\ (\ or\ terms)$, w refers to the binary weights, i.e.,

$$w = \begin{cases} 1, if\ \ the\ term\ is\ \ present\ in\ the\ sentence \\ 0, otherwise \end{cases} \tag{2}$$

and $i, j \in sentences$.

The average similarity between the sentence pairs is calculated by,

$$sim_{avg} = K \sum_{\substack{i=1, j=1 \\ i \neq j}}^{n} sim(s_i, s_j) \tag{3}$$

Now, consider the original sentence collection with the term k removed from all the sentence descriptions and let sim^{k}_{avg} be the average sentence pair similarity in that case.

So, *discrimination value (DV)* can be computed as,

$$DV = (sim^{k}_{avg} - sim_{avg}) \tag{4}$$

According to Salton [1], if $DV > 0$, it refers to good discriminators and if $DV < 0$, it refers to bad discriminators.

3.2 Our Approach: Gain of Words (GOW)

We present here a method, whose major purpose is to discriminate between the significant and non significant terms (or words). As a preprocessing step, we have initially considered all words from the document including the stop words.

Now, let n be the no. of words/ terms considered. Let S be the vector of sentences present in the document. So, we calculated the gain of words by,

$$GOW = \frac{\sum f_{ij}}{\sum i} \times \sum w_{ij}\ , \tag{5}$$

Where, f_{ij} is the frequency of the term (no. of occurrences) j in the sentence i and w is the weight as mentioned in the previous model.

Words having very high *GOW* values are discarded, maintaining a threshold of *0<GOW<10*.

4 Sentence Extraction: *Gain of Sentences (GOS)*

Gain of Sentences, refer to the value which signifies the importance of sentences in a document. The greater the value, higher is the importance. Before computing this, in the preprocessing stage, we discarded all the stop words. As mentioned above, let n be the no. of words / terms considered. Let *S* be the vector of sentences present in the document.

So, we compute the gain of sentences by,

$$GOS = \frac{\sum f_{ij}}{\sum j} \times \sum w_{ji}\ , \tag{6}$$

Where, f_{ij} is the frequency of the term (which means no. of occurrence of the term) j in the sentence i and w is the weight as mentioned in the previous model.

This concept is also used for summarizing a document as it ranks the sentences as per their importance.

5 Experimental Results

In this section we illustrate the experimental results related to the methods discussed in the previous section.

We used the CST data set [16], related to a Milan plane crash. There are multiple single texts in the data set. Since, we focus on single documents, we used each of those for analysis.

Gain of Words: We explained the significance of using G*ain of Words* (GOW) in the previous section. Using GOW, we can eliminate the unwanted words, at the same time keep the possible important words including the entities (the ones generally obtained using named entity extractors). Here we have taken ten words, randomly chosen from the files separately. The tables below show the nature of results obtained using Salton's term significance measure on a single document as well ours. It clearly shows that, for certain words, it gives some meaningful results showing that negative discriminative value, signifying that those words are poor terms. But on the other hand, it cannot differentiate between the good words also. The zeros in the tables show that it cannot identify the terms. Our result overcomes this drawback. The value of the gain computed easily helps us to identify the words between useless, useful and less useful. When the gain values are very large, it shows that the words are useless.

Table 1 and Table 2 illustrate the term significance based on two different methods. It is clearly seen in fig.1 that the highest value for the DV is 0. It is just capable of discarding the most useless terms. The words like "the", "in" (shown in table 1) are the stop words which can be discarded using both the methods. But words like "crash", "plane", "Milan" bear meaningful content, but can be identified by GOW method, not with DV.

Table 1. Comparison between two term significance methods

Document 1		
Words	DVof Salton's method	GOW
the	-0.462	13.847
in	-0.462	13.154
plane	-0.038	0.692
building	-0.077	1.231
crash	-0.038	0.692
april	0	0.077
skyscraper	-0.013	0.308
milan	-0.013	0.308
cnn	-0.0123	0.462
bombing	0	0.077

Table 2. Comparison between two term significance methods

Document 2		
Words	DV of Salton's method	GOW
Are	-0.035	0.842
smoke	-0.006	0.211
police	-0.006	0.211
people	-0.006	0.211
milan	-0.006	0.211
scene	-0.018	0.474
pirelli	-0.018	0.474
italian	-0.018	0.474
from	-0.018	0.474
work	-0.018	0.474

The DV identifies all words as useless, except for “april” and “bombing” and identifies these as almost useless. Our technique also identifies these two as almost useless, but also identifies stop words and useful words and clearly differentiates them.

We basically maintained a threshold of $0<GOW<10$ approximately to choose the words. But there is some noise in out data also.

In table 2, the words like “are”,” from” fall within the threshold limit we have chosen. So these words could not be identified

Gain of Sentences: The Gain of sentences (GOS) is another useful method we present here. The basic target of this part is to analyze and extract the important regions of the text so it partly behaves as a summarization. We had to do some preprocessing before obtaining GOS. Using the threshold mentioned above, we selected the possible significant words from the text. But in order to reduce the noise, we removed the stop words from this new set of words. After this we ran our simulation to obtain GOS. We used the MEAD [17] summarization tool to compare our method. We present here the nature of summaries extracted using this tool as well as with our method. Since GOS creates sentence importance in the document, we here present the five most important sentences to see how far it holds with the MEAD’s process.

For Document 1:

MEAD summarization:

CNN.com - Plane hits skyscraper in Milan - April 18 2002 CNNenEspanol.com A small plane has hit a skyscraper in central Milan setting the top floors of the 30-story building on fire an Italian journalist told CNN.

The crash by the Piper tourist plane into the 26th floor occurred at 5:50 p.m. 1450 GMT on Thursday said journalist Desideria Cavina.

U.N. envoy horror at Jenin camp U.S. bombing kills Canadians Chinese missiles concern U.S. 2002 Cable News Network LP LLLP.

Our Approach: GOS

CNNenEspanol.com A small plane has hit a skyscraper in central Milan, setting the top floors of the 30-story building on fire, an Italian journalist told CNN.

U.N. envoy horror at Jenin camp U.S. bombing kills Canadians Chinese missiles concern U.S. 2002 Cable News Network LP, LLLP.

The crash by the Piper tourist plane into the 26th floor occurred at 5:50 p.m. (1450 GMT) on Thursday, said journalist Desideria Cavina.

Italian TV says the crash put a hole in the 25th floor of the Pirelli building, and that smoke is pouring from the opening.

Many people were on the streets as they left work for the evening at the time of the crash.

For Document 2:

MEAD summarization:

CNN.com - Plane hits skyscraper in Milan - April 18 2002 CNNenEspanol.com A small plane has hit a skyscraper in central Milan setting the top floors of the 30-story building on fire an Italian journalist told CNN.

The crash by the Piper tourist plane into the 26th floor occurred at 5:50 p.m. 1450 GMT on Thursday said journalist Desideria Cavina.

I heard a strange bang so I went to the window and outside I saw the windows of the Pirelli building blown out and then I saw smoke coming from them said Gianluca Liberto an engineer who was working in the area told Reuters.

U.N. envoy horror at Jenin camp U.S. bombing kills Canadians Chinese missiles concern U.S. 2002 Cable News Network LP LLLP.

Our Approach: GOS

"I heard a strange bang so I went to the window and outside I saw the windows of the Pirelli building blown out and then I saw smoke coming from them," said Gianluca Liberto, an engineer who was working in the area told Reuters.

TV pictures from the scene evoked horrific memories of the September 11 attacks on the World Trade Center in New York and the collapse of the building's twin towers.

CNNenEspanol.com A small plane has hit a skyscraper in central Milan, setting the top floors of the 30-story building on fire, an Italian journalist told CNN.

U.N. envoy horror at Jenin camp U.S. bombing kills Canadians Chinese missiles concern U.S. 2002 Cable News Network LP, LLLP.

The crash by the Piper tourist plane into the 26th floor occurred at 5:50 p.m. (1450 GMT) on Thursday, said journalist Desideria Cavina.

The alignment of the sentences we presented might vary from the MEAD summarization. Many of the documents include the same sentences as news sources import the same sentences into different documents. We have presented here the sentences based on their importance in the document. Clearly, our summaries are qualitatively equivalent to the MEAD summarizations.

6 Conclusion

This work is a two way approach of term association where we find the significant words as well as extract the important sentences from a text. It is a simple method based on the syntactic appearances of the terms/ words in a single document. It is very useful to analyze the cases where no predefined data pattern is available. We have also shown that a classic method which has been used successfully for term extraction fails to work when there is a single document or very few documents. Though we have seen that the performance of this model is better, but still we need to improve this in order to get rid of the noise.

References

1. Salton, G.: A Theory of Indexing. In: Regional Conf. Series in Applied Mathematics, Philadelphia, Pennsylvania (1975)
2. Salton, G., Fox, E.A., Wu, H.: Extended Boolean Information Retrieval. Comm. of ACM 26(12), 1022–1036 (1983)
3. Salton, G., Buckley, C.: Term Weighting Approaches in Automatic Text Retrieval. Inf. Processing and Management 24(5), 513–523 (1988)
4. Blake, C.: A Comparison of Document, Sentences, and Term Event Spaces. In: Proc. of 21st intl. Conf. on Comp. Linguists and 44th Annual Meeting of the ACL, pp. 601–608 (2006)
5. Zhang, Z., Goldensohn, S.B., Radev, D.R.: Towards CST-Enhanced summarization. In: Eighteenth national conf. on Artificial intelligence, pp. 439–445 (2002)
6. Katz, B., et al.: START, Natural Language question answering system (1993)
7. Zhang, Z., Otterbacher, J., Radev, D.: Learning Cross-document structural Relationships using Boosting. In: CIKM (2003)
8. Grosz, B.J., Sidner, C.L.: Attention, Intentions, and the Structure of Discourse. Comp. Linguistics 12(3), 175–204 (1986)

9. Radev, D.R.: A Common Theory of Information Fusion from Multiple Text Sources Step One: Cross-Document Structure. In: Proc. of the First SIGdial Workshop on Discourse and Dialogue (2000)
10. Mann, W.C., Thompson, S.A.: Rhetorical Structure Theory: towards a functional theory of text organization. Text 8(3), 243–281 (1988)
11. Salton, G., Singhal, A., Mitra, M., Buckley, C.: Automatic text structuring and summarization. Inf. Processing & Management 33, 193–207 (1997)
12. Landauer, T.K., Foltz, P.W., Laham, D.: An introducition to Latent Semantic Analysis. Discourse Process 25, 259–284 (1998)
13. Rocha, L.M.: TalkMine. A Soft Computing Approach to Adaptive Knowledge Recommendation. In: Loia, V., Sessa, S. (eds.) Soft Computing Agents: New Trends for Designing Autonomous Systems. Series on Studies in Fuzziness and Soft Computing. Physica-Verlag, Springer (2001)
14. Jonas, J., Harper, J.: Effective Counterterrorism and the Limited Role of Predictive Data Mining. Policy Analysis 584, 1–12 (2006)
15. Gedeon, T.D., Koczy, L.T.: Hierarchical co-occurrence Relations. In: Proc. Systems, Man, and Cybernetics, vol. 3, pp. 2750–2755 (1998)
16. Radev, D., Otterbacher, J., Zhang, Z.: CST Bank: A Corpus for the Study of Cross-document Structural Relationships. In: Proc. of LREC 2004 (2004)
17. Radev, D., et al.: MEAD - a platform for multidocument multilingual text summarization. In: LREC, Lisbon, Portugal (2004)

Method for Evaluating the Security Risk of a Website Against Phishing Attacks

Young-Gab Kim[1], Sanghyun Cho[2], Jun-Sub Lee[3], Min-Soo Lee[3], In Ho Kim[4], and Sung Hoon Kim[4]

[1] Graduate School of Information Management and Security,
Center for Information Security Technologies (CIST), Korea University,
1, 5-ga, Anam-dong, SungBuk-gu, 136-701, Seoul, Korea
always@korea.ac.kr
[2] NHN Corporation IT Security Analysis Team,
Venture Town Bldg., Jeongja-dong, Bundang-gu, Seongnam-si, 25-1, Gyeonggi-do, Korea
bungae@nhncorp.com
[3] Div. of Computer Science Dept. of EECS,
Korea Advanced Institute of Science and Technology (KAIST),
335 Gwahangno (373-1 Guseong-dong), Yuseong-gu, 305-701, Daejeon, Korea
{jslee,mslee}@dependable.kaist.ac.kr
[4] Korea Information Security Agency (KISA)
IT Venture Tower, Jungdaero 135 (Garak-dong 78), Songpa-gu, 138-950, Seoul, Koera
{kih,kimsh}@kisa.or.kr

Abstract. As Internet technologies evolve, phishing and pharming attacks frequently occur and diversify. In order to protect the economic loss and privacy of Internet users against the phishing attacks, several researches such as website authentication and email authentication have been studied. Although, most of them use website black-list (WBL) or website white-list (WWL), there are several weak points, such as validity of WBL DB (database) and the short life-cycle of phishing websites. That is, it is impossible to discriminate between legitimate and forged websites until the phishing attacks are detected and recorded into WBL DB. Furthermore, the existing WBL and WWL approaches hardly counter the new generation of sophisticated malware pharming attacks. In this paper, in order to overcome the limitation of WBL and WWL approaches, new approach based on the WWL approach, which can quantitatively estimate the security risk of websites that is security risk degree representing the phishing websites, is proposed.

1 Introduction

As Internet technologies (e.g. E-Commerce and Internet Banking) evolve, phishing and pharming attacks frequently occur and diversify. The phishing attacks steal user identity data and financial account credentials using social engineering and technical subterfuge [1]. The social engineering schemes use 'spoofed' e-mails to lead users to forged websites designed to trick attacker into divulging financial data such as credit card numbers, account usernames, passwords and social security numbers. The pharming attacks, e.g. hijacking brand names of banks, are more advanced phishing

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 21–31, 2008.

attacks that it misdirects users to fraudulent sites or proxy servers typically through DNS hijacking, often using crimeware such as Trojan keylogger spyware [1]. Phishing attacks not only exploit software vulnerabilities but also human vulnerabilities since Internet users of average skill often do not understand security indicators and cannot distinguish between legitimate and forged websites [2]. In order to protect the economic loss and privacy of Internet users against the phishing attacks, several researches such as website authentication and email authentication have been studied. Most of them use website black-list (WBL), which is a list of e-mail addresses or IP addresses that are originating with known safe website, or website white-list (WWL), which is a list of e-mail addresses or IP addresses that are considered safe website. However, there are several weak points with the WBL and WWL approach. First, validity of WBL DB (database) is low because the life-cycle of phishing websites is short. Second, it is impossible to discriminate between legitimate and forged websites until the phishing attacks are detected and recorded into WBL DB. Furthermore, the WBL and WWL approaches hardly counter the new generation of sophisticated malware phishing attacks, pharming attacks, designed to target certain services. In addition, many of the existing WWL approach use only website URL to distinguish between legitimate and forged websites. In this paper, in order to overcome the limitation of WBL and WWL approaches, new approach, which can quantitatively estimate the security risk of websites that is security risk degree representing the phishing websites, is proposed. The proposed approach includes the definition of security risk elements, which are used to quantitatively calculate the security risk of websites, and procedure for quantitative analysis of the security risk of the website.

The sebsequent sections of this paper are organized as follows: In Section 2, the background and related works about the phishing and pharming researches are presented. Section 3 shows the approach method to estimate the security risk of a website, including the definition of security risk elements and the steps for evaluating the security risk of the website. In addition, a case study to apply it to real websites is presented. Section 4 presents a short discussion about the proposed approach and its implementation. Section 5 concludes this paper.

2 Background and Related Works

Anti-Phishing Working Group (APWG) [1] is the global industry working group to eliminate the fraud and identity theft that result from phishing, pharming and email spoofing. The APWG, which is composed of many organizations and security companies, provides diverse information such as phishing reports, research data, and resources related with the phishing, pharming, and crimeware.

Several research efforts have been made to prevent phishing attacks. In this section, we briefly review some typical approaches as divided into two parts: Server-side approach, and Client-side approach.

Server-Side Approach. In server-side approach, server authentication is required to defend against phishing attacks. One of main reasons why phishing attacks are possible is because e-mails can be spoofed easily. Although spam filters researches quite well today, they cannot guarantee that all phishing e-mails are intercepted. As one of

these solutions, which authenticate the sender's e-mail, and prevent phisher from using hijacked mail address, Miscrosoft presents the Sender ID Framework [3], and Yahoo uses its own technique called DomainKey [4]. Currently, Yahoo and other industry leaders are in the process of standardizing a technique called DKIM (DomainKeys Identified Mail) [5]. Another authentication approach is to share a secret such as a password and an image, between server and client. Dhamija et al [6, 7] proposes Dynamic Security Skins, which allows that users visually verify whether the image from the server matches its corresponding local images. Finally, Fu et al. [8, 9] proposes a visual similarity assessment-based antiphising strategy, which uses visual characteristics to identify potential phishing websites and measure suspicious webpages's similarity to actual sites registered with the system.

Client-Side Approach. In client-side approach, most solutions are supported as a toolbar, which show different types of security messages to help users to detect phishing websites, built-in the web browser [10]. Chou et al. [11] proposes a framework for client-side defense using a browser plug-in called SpoofGuard that examines webpages and warns the user when request for data may be part of a spoof attack. It uses domain names, invalid links, URL obfuscation and images to measure the similarity between a given page and the pages in the caches. Beside this solution, there are many toolbars, such as TrushWatch [12], Netcraft [13], EarthLink [14], MS Phishing Filter [15], which are designed to detect and prevent phishing attacks. Most of them use WBL and WWL, which depends on phishing reports. As mentioned earlier, as long as a phishing website has not been reported, phishers may steal personal data from visitors to the website. Wu et al. [16] presents the Web Wallet, which prevents phishing attacks by forcing users to compare, then confirm before going to a website instead of just confirming. More comprehensive survey of antiphishing solutions can be found in [17].

3 Method for Evaluating the Security Risk of a Website

In this section, in order to evaluate a security risk of a website, the security elements, which are used to calculate a degree of the website's security risk, are proposed. Furthermore, steps for quantitative analysis of the website's security risk using the security elements are explained in detail.

The security risk of a website is calculated through 6 steps as depicted in Fig. 1: Definition of the risk elements, Weight between risk elements, Measurement of the risk grade, Calculation of the total security index (TSI), Calculation of the max security index (MSI), and finally Calculation of the website security risk index (WSRI).

In order to quantitatively calculate the security risk of a website, the security risk evaluation matrix (SREM) based on the scoring method is proposed as shown in Fig. 2. The SREM can be used and implemented as WWL DB elements or security features related with web pages.

The security risk element depicted in the SREM means a condition or a situation which can cause security attacks. A more detailed description about the security risk elements and steps will be presented in the following subsections.

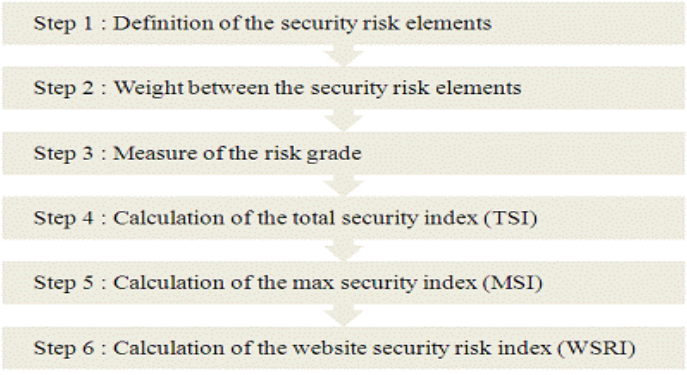


Fig. 1. Steps for evaluating the security risk of a website

Security Risk Element	Weight	Risk Grade					SRI
		4	3	2	1	0	
…..							
TSI	-	-	-	-	-	-	

Fig. 2. Security Risk Evaluation Matrix (SREM)

3.1 Steps for Quantitative Analysis of the Website

Step 1: Definition of the security risk elements
There are many kinds of elements related with a website or a webpage. In this paper, we propose eight security risk elements, which are especially applicable elements related with an occurrence of the phishing attacks, are proposed as follows. These elements can be implemented as WWL DB.

Server-Name. This element is used to judge whether the domain name of a website coincides with the registered IP address. As mentioned previously, most of pharmming attacks misdirect users to the phishing website through DNS hijacking using crimeware such as Trojan. Therefore, pharming attacks can be detected using check of this element.

Domain-Country. In order to detect suspicious websites as considered phishing routes or phishing websites, this element checks whether the domain country information requested by a user coincides with the IP address assigned by an organization of real country.

Domain-Life. The Domain-Life means a period of time from the registration date to the expiration date of a domain. This element considers the characteristic of a phishing website, which has a short life-span of time. That is, it is considered that the

websites, which have a longer Domain-Life span, are safer. The Domain-Life can be calculated using the formula (1).

$$Domain\text{-}Life = Expiration\ Date - Registration\ Date \quad (1)$$

Domain-Age. Like the Domain-Life, the Domain-Age means a period of time from the registration date of a website to the current time. This element complements the false positive feature, which is the characteristic of the Domain-Life. The Domain-Age can be calculated using the formula (2)

$$Domain\text{-}Age = Current\ Date - Registration\ Date \quad (2)$$

Domain-Famous. Domain-Famous means an eminence of the website that is the number of search result related with the website, evaluated by search engine such as Google, Yahoo, and so on. If the search result of website is high, it means that many people use the website information, and secures. On the contrary, if the search result is low, it is possible that the website is create newly or a phishing website.

DNS-Ranking. This element means the DNS query ranking information supported by the specific companies such as Rankey [18] and Alexa [19]. This element also considers the characteristic of the phishing website that has shorter lifecycle with that a more little user contact with this website.

Website-Type. The Website-Type is decided by the subjects of supporting services such as a government, a public institution, a company, and a personal website. If the web services are supported by a trusted website as one of the government and the public institution website, the website will have a lower security risk than that of the personal website.

Security-Manager. The security manager plays critical role in fraud management such as fraud prevention efforts and deliver actionable results that allow organizations to take proactive action in preventing and remediating fraud. This element is used to decide whether the website can be used as a phishing website by a hacker. Generally, a website managed by a great security manager has a low possibility that the website is used as a phishing website or a routing path of the phishing website.

Step 2: Weight between security risk elements

In this Step, weight is given to the security risk elements defined in previous step according to their impact to security risk of the website. In a relative security risk evaluation (or an absolute security risk evaluation), for example, the weight value ranges from 1 to 5. Higher weight value is more important element, which can increase security risk of a website. In this paper, in order to show case study in subsection 3.2, the security risk elements defined in previous step are assigned the weight value using the following criteria.

5: Server-Name
4: Website-Type
3: Domain-Life, Domain-Age, Domain-Famous
2: Domain-Country, DNS-Ranking
1: Security-Manager

Step 3: Measure of the risk grade
In Step 3, the risk grade for the each security risk elements is measured. In this point, the risk grade is a possibility of causing the threat occurrence. In this paper, the risk grade can be from 0 to 4 that the grade '4' is the highest value to cause the risk occurrence. Each the risk grade is measured by the statistic information related with the security risk elements as presented in subsection 3.2 Case Study.

Step 4: Calculation of the total security index (TSI)
Step 4 is a process to calculate the total security index (TSI), which is a score that represents a security risk degree embedded in a website. As depicted in formulas (3) and (4), the TSI is calculated as a sum of security risk index (SRI) that is a multiplication of the risk grade and the weight for the each security risk elements.

$$TSI = \sum_{i=1}^{n} SRI_i \tag{3}$$

where n is a number of the security risk elements and i is a specific security risk element.

$$SRI = Risk\,Grade \times Weight \tag{4}$$

Step 5: Calculation of the max security index (MSI)
In Step 5, the max security index (MSI), which means the maximum security risk degree, is calculated. The MSI is a theoretically max risk value of the website as formula (5). That is, in this paper, the MSI is a total value of the SRI when the all risk grades for the each security risk elements is a max risk grade value 4.

$$MSI = \sum_{i=1}^{n} SRI_i = \sum_{i=1}^{n} (Risk\,Grade_{\max} \times Weight)_i \tag{5}$$

where n is a number of the security risk elements, and i is a specific security risk element.

Step 6: Calculation of the website security risk index (WSRI)
Finally, the website security risk index (WSRI), which is the security risk degree representing phishing websites, is calculated using the SRI and the MSI calculated in steps 4 and 5 as formula (6).

$$WSRI = \frac{TSI}{MSI} \times 100 \tag{6}$$

The value of WSRI ranges from 0 to 100 that the website with higher value means a more suspicious website as phishing websites. In order to illustrate the motivation of our research, a case study is presented in the following section.

3.2 Case Study

In order to show a case study, let us suppose several assumptions related with the security risk elements and the website, which users want to contact. First, the eight security risk elements proposed in Step 1 in subsection 3.1 are used to quantitatively evaluate the security risk of websites. Second, weight for each security risk elements is assigned by the criteria proposed in Step 2 in subsection 3.1. Finally, a website that user will contact, is defined as follows: A website has a URL, which coincides with the registered IP address. Country information for a specific domain coincides with

Security Risk Element	Weight	Risk Grade					SRI
		4	3	2	1	0	
Server-Name	5						
Domain-Country	2						
Domain-Life	3						
Domain-Age	3						
Domain-Famous	3						
DNS-Ranking	2						
Website-Type	4						
Security-Manager	1						
TSI	-	-	-	-	-	-	

Fig. 3. SREM through Step 1 to Step 2

Security Risk Element	Explanation	Risk Grade				
		4	3	2	1	0
Server-Name	Checks the registered URL and IP address	No Match	-	-	-	Match
Domain-Country	Checks countries of domain and IP address	No Match	-	-	-	Match
Domain-Life	Expired Date – Created Date	Under 2 years	Under 4 years	Under 6 years	Under 8 years	Over 8 years
Domain-Age	Current Date – Created Date	Under 1 year	Under 2 years	Under 3 years	Under 4 years	Over 4 years
Domain-Famous	Webpages queried by Google search engine	Under 100 pages	Under 1,000 pages	Under 10,000 pages	Under 100,000 pages	Over 100,000 pages
DNS-Ranking	DNS rank supported by Rankey website	Under 10 %	Under 30 %	Under 50 %	Under 80	Over 90%
Website-Type	Type of a website	Company	School	Personal	Public Organization	Government
Security-Manager	Checks the security manager and security policy	No Security Manager	-	Security Manager	-	Security Manager, Policy

Fig. 4. Criteria for risk grade of the security risk elements

the country used by IP address. Domain-Life of the website is 5 years, and Domain-Age is 3 years. The number queried by Google, related with the website, is about 30. DNS ranking of the website supported by Rankey is under 30% among the registered websites. The website serves an electronic commerce company. Although, the company engages a security manager, there is no any security plan or policy.

Through Step 1 to Step 2 using above assumptions, the SREM is constructed as shown in Fig. 3.

In this paper, in order to measure the risk grade for each security risk elements, criteria as depicted in Fig.4 are used.

Through Step 1 to Step, the SREM is constructed as shown in Fig. 5.

Security Risk Element	Weight	Risk Grade					SRI
		4	3	2	1	0	
Server-Name	5					V	0
Domain-Country	2					V	0
Domain-Life	3			V			6
Domain-Age	3			V			6
Domain-Famous	3	V					12
DNS-Ranking	2		V				6
Website-Type	4	V					16
Security-Manager	1	V					4
TSI	-	-	-	-	-	-	50

Fig. 5. SERM through Step 1 to Step 3

Using the formula (3) presented in Step 4 in subsection 3.1, the TSI is calculated as follows:

$$TSI = \sum_{i=1}^{8} SRI_i = 0+0+6+6+12+6+16+4 = 50$$

Next, in Step 5, the MSI is calculated using the formula (5) as follows:

$$MSI = \sum_{i=1}^{8} SRI_i = \sum_{i=1}^{8} (Risk\ Grade_{\max} \times Weight)_i$$
$$= 20+8+12+12+12+8+16+4 = 92$$

Finally, the WSRI of website is calculated using the TSI and MSI as follows:

$$WSRI = \frac{TSI}{MSI} \times 100 = \frac{50}{92} \times 100 \cong 54$$

Website	Server-Name	Domain-Country	Domain-Life	Domain-Age	Domain-Famous	DNS-Ranking	Website-Type	Security-Manager	TSI	WSRI
Weight	5	2	3	3	3	2	4	1	MSI=92	
www.naver.com	0	0	0	0	0	0	4	0	16	17
			17 years	11 years	270,000	1	Com.			
www.daum.net	0	0	0	0	0	0	4	0	16	17
			13 years	11 years	228,000	3	Com.			
www.cyworld.com	0	0	1	0	0	0	4	0	19	21
			8 years	8 years	1,940,00	4	Com.			
www.deinside.com	0	0	0	0	0	0	4	2	18	20
			10 years	7 years	173,000	29	Com.			
www.nate.com	0	0	0	0	1	0	4	0	19	21
			11 years	11 years	38,000	2	Com.			
www.kisa.or.kr	0	0	0	0	1	0	1	0	7	8
			11years	11years	56,000	2273	Org.			
www.abuji.com	0	0	1	0	3	4	4	4	40	43
			8 years	6 years	406	267,000	Com.			
www.cqrity.net	0	0	2	1	4	4	2	4	41	45
			5 years	4 years	3	N/A	Per.			
livegame.any.to	0	4	4	4	1	4	4	4	63	68
			N/A	N/A	15,100	N/A	Com.			

Fig. 6. Experimental result for real websites

From above result, the website, which a user wants to contact, has high security risk level. That is, it is required that the user pay attention to contact with this website. In order to verify the proposed security risk evaluation method, we apply it to real website, and obtained following results.

From the result in Fig. 6, the famous websites such as Naver, Daum, and KISA (Korea Information Security Agency) [20] are recognized as safe website. However, other websites is estimated as suspicious websites like phishing websites.

4 Discussion

The major strength of the proposed approach is its ability to estimate the security risk of websites quantitatively based on the WWL database. This property is particularly useful for browser plug-in or applications against phishing attacks. However, in order to implement the real antiphishing applications and antiphishing systems using the proposed approach, the following issues should be considered:

First, information for several security risk elements proposed in this paper, such as Domain-Famous, and Domain-Raking, should be supported and implemented by the trusted third party. That is, the more precise result can be obtained with the more correct and trusted information about the security risk elements.

Second, the criteria for making decision on the risk grade of the security risk elements as depicted in Fig. 4 should be created with the more analyzed information. Particularly, in Step 2 in subsection 3.2, weight assignment the security risk elements should decide an appropriate method (a relative evaluation or an absolute evaluation) in some situation.

Finally, although only the eight security risk elements are proposed in this paper, the more extensible security features in the websites and the webpages, such as IP links, tricky links and external links, can be added.

5 Conclusion and Future Works

In this paper, we proposed the new method for evaluating the security risk of a website against the phishing attacks. In order to calculate the security risk, the eight security risk elements, and procedure for quantitative analysis were proposed. The proposed method can detect the suspicious websites containing phishing attacks and abnormal behavior, and warn users the security risk of website with the risk level. Furthermore, from the case study, we can discriminate the normal website from the phishing website.

However, as mentioned in Section 4, more research is needed to obtain a more precise result, such as: 'How should the appropriate threshold for the risk grade be decided?', and 'Which security risk element should be priorized first among the many security elements?'. Finally, a specific structure for implementing it to antiphishing system is required to be designed.

References

1. Anti-Phishing Working Group (APWG) (2008), http://www.antiphishing.org
2. Tygar, J.D., Dhamija, R., Hearst, M.: Why Phishing Works. In: Proc. of the Conference on Human Factors in Computing Systems (CHI 2006) (2006)
3. Microsoft, Sender ID Framework Overview (2008), http://www.microsoft.com
4. Yahoo: Yahoo! Anti-Spam Resource Center (2008), http://antispam.yahoo.com
5. Mutual Internet Practices Association, DomainKeys Identified Mail (DKIM) (2008), http://www.dkim.org
6. Dhamija, R., Tygar, J.D.: The Battle against Phishing: Dynamic Security Skins. In: Proc. of the 2005 symposium on Usable Privacy and Security (SOUPS 2005), pp. 77–88 (2005)
7. Dhamija, R., Tygar, J.D.: Phish and Hips: Human Interactive Proofs to Detect Phishing Attacks. In: Proc. of the Second International Workshop, pp. 127–141 (2005)
8. Fu, A.Y., Wenyin, L., Deng, X.: Detecting Phishing Web Pages with Visual Similarity Assessment Based on Earth Mover's Distance (EMD). IEEE Transactions on Dependable and Secure Computing 3(4), 301–311 (2006)
9. Liu, W., Deng, X., Huang, G., Fu, A.Y.: An Antiphishing Strategy Based on Visual Similarity Assessment. IEEE Internet Computing, 58–65 (2006)
10. Raffetseder, T., Kirda, E., Kruegel, C.: Building Anti-Phishing Browser Plug-Ins: An Experience Report. In: Proc. of third international workshop on Software Engineering for Secure Systems (SESS 2007) (2007)
11. Chou, N., Ledesma, R., Teraguchi, Y., Boneh, D., Mitchell, J.C.: Client-side Defense against Web-Based Identity Theft. In: Proc. of 11th Annual Network and Distributed System Security Symposium (NDSS 2004) (2004)
12. TrustWatch (2008), http://www.trustwatch.com
13. NetCraft (2008), http://www.netcraft.com
14. EarthLink (2008), http://www.earthlink.com
15. Microsoft, http://www.microsoft.com/mscorp/safety/technologies/antiphishing/
16. Wu, M., Miller, R.C., Little, G.: Web Wallet: Preventing Phishing Attacks by Revealing User Intentions. In: Proc. of Symposium On Usable Privacy and Security (SOUPS 2006), pp. 102–113. ACM Press, New York (2006)

17. Emigh, A.: Online Identity Theft: Phishing Technology, Chokepoints and Countermeasures. ITTC Report on Online. Identity Theft Technology and Countermeasures (2005)
18. Rankey (2008), http://www.ranky.com
19. Alexa the Web Information Company (2008), http://www.alexa.com
20. Korea Information Security Agency (KISA) (2008), http://www.kisa.or.kr

CyberIR – A Technological Approach to Fight Cybercrime

Shihchieh Chou[1] and Weiping Chang[2]

[1] Department of Information Management, National Central University, Chung-Li, 320 Taiwan
scchou@mgt.ncu.edu.tw
[2] Department of Student Affairs, Central Police University, kueishang, 333 Taiwan
una024@mail.cpu.edu.tw

Abstract. Fighting cybercrime is an international engagement. Therefore, the understanding and cooperation of legal, organizational and technological affairs across countries become an important issue. Many difficulties exist over this international cooperation and the very first one is the accessing and sharing of related information. Since there are no standards or unification over these affairs across countries, the related information which is dynamically changing and separately stored in free text format is hard to manage. In this study, we have developed a method and information retrieval (IR) system to relieve the difficulty. Techniques of vector space model, genetic algorithm (GA), relevance feedback and document clustering have been applied.

Keywords: Cybercrime, fighting cybercrime, high-tech crime, information retrieval.

1 Introduction

Cybercrime is usually deemed as any illegal activity conducted through a computer. Various types of cybercrime could be identified like Internet fraud, computer hacking, network intrusion, spreading of malicious code, cyber piracy, identity theft, electronic property theft, money laundering, cyber pornography, etc. And yet, new type of cyber crime still has been brewing. Since the taking place of these cybercrimes is not limited within a specific region, fighting cybercrime is an international engagement. Therefore, the understanding and cooperation of legal, organizational and technological affairs across countries become an important issue. However, there exist many difficulties over this international cooperation, and the very first one is the accessing and sharing of related information. There are several reasons: first, the information about legal, organizational and technological affairs is separated across countries; second, the information is usually stored in a free text format, third, the information is dynamically changing. Therefore, how to make this information transparent to world wide agencies responsible for cybercrime fighting becomes a big challenge. Our proposition for tackling this information sharing problem is the application of information retrieval (IR) techniques which can deal with the accessing

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 32–43, 2008.

of up-to-date information which is separately stored in free text format. To demonstrate the implementation of our proposition, this research has aimed to develop a method and IR system with the techniques of vector space model, genetic algorithm (GA), relevance feedback and document clustering applied. It will also experiment with the divergent embodiments of the IR system to provide some performance basis for advanced study.

In section 2, we will refer to the needs of the legal, organizational and technological affairs across countries and the application of the IR techniques. Next, the design of the IR system will be described. Then, the experiments on the divergent embodiments of the system will be presented. Finally, some conclusions will be made.

2 Fighting Cybercrime

To fight cybercrime, the human affairs of legal, organizational and technological approaches across countries are required. Since there are no standards or unification over these legal, organizational and technological affairs across countries, the related information becomes a huge and complicated one as we have presented in [7]. The proposed method to deal with this information sharing problem in this research is the application of IR techniques.

2.1 The Need of Legal, Organizational and Technological Information

The legal affairs are needed to restrict cybercrime activities. Many countries have created new laws or modified the current laws to fight cybercrime. World wide law enforcement agencies investigate cybercrime activities in accordance with these related laws. However, the related laws across countries are not in consistence. Activities deemed as illegal in some countries might not be treated as crime in many other countries. Therefore, conflicts between national law enforcement authorities might occur due to the territoriality. Since cybercrime is transnational, investigators processing cybercrime cases in one country might need to know the criminal liabilities of these cybercrime cases in the related countries. Therefore, sharing of the complicated legal information across countries is essential in the fighting against cybercrime.

In addition to sharing of the legal information, international cooperation is also highly required at the organizational level in terms of law enforcement. As the law enforcement agencies in one country investigate a cybercrime case, they might need to cooperate with the right law enforcement agencies of another country. They cooperate to trace the hacker or attacker, or to sweep the child exploitation, international money laundry, or illegal drug trafficking on Internet. Considering the communication speed utilized by the criminal, information to support the contact at the right time between the right organizations and agencies becomes urgent. Therefore, information about organizational affairs related to the law enforcement agencies in charge of cybercrime in all countries should be transparent to each other country at any time.

The criminals of cybercrime keep applying the novel techniques or developing new tools to commit cybercrime and to escape from being arrested. Therefore, law

enforcement agencies always need to keep the most new information and knowledge about computer technologies and investigation techniques to promote the effectiveness and efficiency of cybercrime analysis and investigation. Since different countries have different types of cybercrime, the knowledge of computer technologies and investigation techniques are usually developed locally. If the computer technologies and investigation techniques of all countries could be integrated, it would be very helpful in the fighting against cybercrime. Therefore, the sharing and reuse of technological information and knowledge should be very helpful to all countries in the enhancement of effectiveness and efficiency of cybercrime investigation.

2.2 The Application of IR Techniques

As aforementioned, the application of information retrieval (IR) techniques could provide great support to the using of the information since it has strong capability in the dealing of up-to-date information which is separately stored in free text format. In the study of IR, the basic problem to be solved is the relevance of retrieval. In the past, studies have applied many techniques to limit the amount and increase the relevance of information retrieved. There currently is not any solution suggested as a perfect and unique answer. In this research, the general concept of vector space model and genetic algorithm (GA) which have been proved to be useful will be applied; a specific document clustering and labeling algorithm which we have developed before [6] will be involved; and a method of relevance feedback will be developed.

2.2.1 Vector Space Model

The development of the IR system in this research has been based on the vector space model. According to the vector space model, the document and the query are both thought of as vectors in an n-dimensional space, where each dimension represents an index term and every index term has a weight. Tf-idf developed by Gerard Salton [22] is the most common term weighting approach. Having the document and the query represented as vectors, the cosine angle between the two vectors is calculated to measure the similarity between the two vectors. The advantage of the vector space model is the capability to produce a ranked list of documents in terms of similarity. However, the application of this model lies on the assumption that the terms are independent.

2.2.2 Genetic Algorithm

GA, invented by John Holland [12], is based on Darwin's theory of evaluation. It has been applied in solving optimization problems. The way of application is the performing of the set of problem population whose individuals are characterized by possessing a chromosome which is composed of a sequence of genes. Procedures of GA processing include chromosome reproduction, chromosome crossover, gene mutation, chromosome fitness and natural selection. With these procedures, GA can produce the best individual of the population or a combination of the best chromosome of that population. In IR, GA has been applied in many ways. For example, Michael Gordon used GA to search the acceptable set of terms and weights in representing a document [11]; Hsinchun Chen applied GA to search the keyword set that could effectively represent the example documents [4]; Praveen Pathak used

GA to adjust weights of similarity functions for documents and inquiry [20]; Zacharis Nick had GA keep refining the searching key words based on example documents, retrieved documents and feedback on retrieved documents to have the retrieved documents fit the user's requirement [17]. In this research, our application of GA will be identical to Nick's in general although the embodiment will be different.

2.2.3 Document Clustering

The ranked lists of retrieved snippets or documents by search engine are used as the major way of browsing. Usually, it is inefficient in the distinguishing of totally different concepts produced by poor or polysemous queries. Snippets or documents clustering are the effort endeavored to relieve the efficiency. In this research, we will have the clustering algorithm that we have developed before [6] applied in the IR system to generate hierarchical clustering for the retrieved snippets or documents with label attached.

2.2.4 Relevance Feedback

The very first study on the using of the user's relevance feedback to improve query performance was conducted by Rocchio [21]. The principle of Rocchio's study was to adjust query vector according to the user's relevance rating for the retrieved documents. The original formula proposed by Rocchio is as follows:

$$Q_1 = Q_0 + \beta \sum_{k=1}^{n1} \frac{R_k}{n_1} - \gamma \sum_{k=1}^{n2} \frac{S_k}{n_2}$$

Where

Q_1 = new query vector
Q_0 = initial query vector
R_k = vector for relevant document k
S_k = vector for non-relevant document k
n_1 = number of relevant documents
n_2 = number of non-relevant documents

β and γ = weight multipliers to control relative contributions of relevant and non-relevant documents.

In the formula, the terms are re-weighted by adding the weights from the actual occurrence of those query terms in the relevant documents, and subtracting the weights of those terms occurring in the non-relevant documents. While some of current retrieval systems rely on Rocchio's formula for query modification [1], many other researchers have continued to study the application of the relevance rating. Following are the main topics and example studies: (1) The β and γ parameters in Rocchio's original formula. Studies conducted by Yu et al. [27], Koster et al. [15] and Moschitti [16] have used different value for the β and γ parameters in Rocchio's original formula to achieve better accuracy. (2) The vector's weighting scheme. Studies by Buckley et al. [3] and Desjardins et al. [8] have developed weighting formula different from Rocchio's for use in their IR system. (3) Ranking of the relevance rating. Studies by Balabanovic et al. [2] or Nick and Themis [17] have collected the user's ranking of the relevance rating for the retrieved document and used it to modify the user's query interest. (4) Calculation of relevance degree.

Kim et al.'s [14] study measure the relevance by degree which was calculated by fuzzy inference using the information such as co-occurrence similarity, document frequency within the feedback documents and the inverse document frequency. The calculated relevance degree is used to re-weight the terms. (5) The strategy. In Okabe et al.'s [18] study, an approach for refining a document ranking by learning filtering rule sets through relevance feedback has been proposed. In Azimi-Sadjadi's [1] study, a learning mechanism which optimally maps the original query using relevance feedback from multiple expert users has been introduced.

Continued studies as mentioned have extended the application on the relevance feedback in the vector-space-modeled IR system. In this research, we have identified the information of term appearance situations (abbreviated as *tas* later) existed in the retrieved documents rated as relevant/non-relevant for application. Consider term appearance situations of the following: (1) A term can appear in relevant documents only and never appear in non-relevant documents (termed as *tas* 1 later). (2) A term can appear in non-relevant documents only and never appear in relevant documents (termed as *tas* 2 later). (3) A term can appear both in relevant and non-relevant documents (termed as *tas* 3 later). In the past, *tas*'s affection was not clearly identified and applied. Since studies have reported that both relevant and non-relevant feedback are useful in the enhancement of information retrieval [10,19,23,24,26], it is reasonable to suggest that terms belonging to *tas* 1,2,3 are all applicable but with different ways of application in the enhancement of retrieval effectiveness. In the system design presented in the following section, we will develop the vector's weighting strategy based on the information of *tas***.**

3 System Framework of CyberIR

To conquer the difficulty in the accessing of the information of legal, organizational and technological affairs across countries, we have developed an IR system – CyberIR. Figure 1 shows the system framework of CyberIR. To operate CyberIR, world wide organizations responsible for cybercrime fighting need just to maintain their own web servers.

In the system framework, the component Information Collector is the user interface; the five components - User Profile, Learning Agent, Query String Generator, Spider and Document Selector – together are used to retrieve the documents most relevant to the user based on the user's retrieval interest; the three component - Preprocessing, Label Extractor & Label Processor and Hierarchy Builder – together are used to extract labels and build hierarchy of labeled clusters. We detail each component as follows.

Information Collector: This component is the user interface used to input the original query keywords, the original example documents and the retrieved documents with the user's relevance feedback. The user's rating of relevance/non-relevance includes five categories: very relevant, relevant, not sure, non-relevant and very non-relevant. Except data input, this component also filters out useless information like punctuation and stop words.

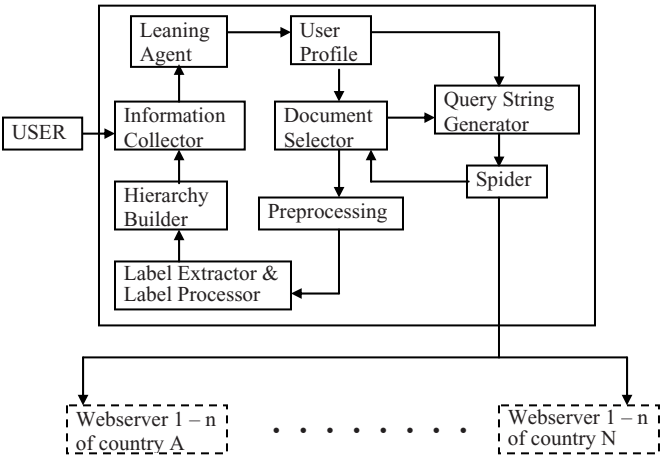


Fig. 1. System framework of CyberIR

User Profile: This component is used to maintain the user's interest/disinterest on information retrieval. There are two user profiles, the positive and the negative user profile; kept to record the system's learning of the user's interest/disinterest respectively. The positive user profile is a database table with the four attributes: 'Term', 'Frequency', 'Sensitivity' and 'Adjusted Frequency'. The negative user profile is a database table with the two attributes: N-Term and N-Frequency.

Learning Agent: This component is used to learn the user's interest/disinterest from the user's input, including the key words, the example documents, the retrieved documents and the user's rating of relevance/non-relevance for the retrieved documents. In detail, the learning is based on the following input information: terms, frequencies, ranking of relevance/non-relevance for the retrieved documents and term appearance situations.

With the input, the Learning Agent will construct and maintain the user profile. For the positive user profile, the value for the attribute 'Term' by each row is derived from the terms appeared in documents rated by the user as relevant/very relevant. The value for the attribute 'Frequency' by each term is determined together by formula 1 and TABLE 1. The value for the attribute 'Sensitivity" by each term is given according to the conditions shown in TABLE 2. The value for the attribute 'Adjusted Frequency' by each term is calculated from frequency and sensitivity as formula 2 shows. After the set of the values for the positive user profile, the rows of the positive user profile can be sorted by the values of 'Adjusted Frequency' in descending order. A certain number of top ranked terms together with the value of 'Adjusted Frequency' are then used to represent the user's interest. For the negative user profile, the value for the attribute 'N-Term' by each row is derived from the terms appeared 'only' in documents rated by the user as non-relevant/very non-relevant. The value for the attribute 'N-Frequency' by each term is determined together by formula 3 and TABLE 1. After the set of the values for the negative user profile, the rows of the negative user profile can be sorted by the values of 'N-Frequency' in descending

order. A certain number of top ranked terms are then used to represent the user's disinterest.

In TABLE 1, the C value used to weight the term's frequency is given according to the user's relevance rating for the document where the term appears. In TABLE 2, the sensitivity value used to weight the value of 'Frequency' is given according to the term's *tas*. The weighting scheme of C and sensitivity is an adoption of the adjustment strategy shown by many studies as feasible in the enhancement of information retrieval [2, 14, 17]. Like many researches that have set constants for their experiments [5, 9, 13, 25], we have conducted preliminary tests to detect the appropriate values that make CyberIR perform well and stable for C and sensitivity.

(1) $\text{Frequency}_i = \text{Frequency}_i + C_j \times F_{ij}$

Where

Frequency_i: the value for the attribute 'Frequency' by term_i in the positive user profile

C_j: adjustment value for F_{ij} according to the relevance rating for document_j formulated in Table 1

F_{ij}: frequency of term_i of the document_j rated as relevant/very relevant

$_i$: 1-n, n is the counting of the individual terms in the retrieved documents rated as relevant/very relevant

$_j$: 1-k, k is the counting of the retrieved documents rated as relevant/very relevant

(2) $\text{Adjusted frequency}_i = \text{Frequency}_i \times \text{Sensitivity}_i$

Where

$\text{Adjusted frequency}_i$: the value for the attribute 'Adjusted Frequency' by term_i in the positive user profile

Frequency_i : the value for the attribute 'Frequency' by term_i in the positive user profile

Sensitivity_i: adjustment value for Frequency_i according to *tas* classification of term_i formulated in Table 2

$_i$: 1-n, n is the counting of the individual terms in the retrieved documents rated as relevant/very relevant

(3) $\text{N-Frequency}_i = \text{N-Frequency}_i + C_j \times NF_{ij}$

Where

N-Frequency_i: the value for the attribute 'N-Frequency' by term_i in the negative user profile

C_j: adjustment value for NF_{ij} according to the relevance rating for document_j formulated in Table I

NF_{ij} : frequency of term_i of document_j where term_i belongs to *tas* 2

$_i$: 1-n, n is the counting of the individual terms belonging to *tas* 2

$_j$: 1-k, k is the counting of the retrieved documents with terms belonging to *tas* 2

Query String Generator: This component is used to generate a query string from the terms in the user profile and pass the query string to spider to retrieve documents. GA has been applied to optimize the using of the terms of the user profile in the forming of the query string. In our application of GA, each chromosome is represented by N bits. Among the N bits, M bits (M < N) is used to represent positive keywords connected by AND operator and M-N bits is used to represent negative key words connected by

Table 1. Adjustment value for relevance rating

Relevance rating for a retrieved document	C value
Very Relevant	1.2
Relevant	1
In-between	0
Non-relevant	1
Very Non-relevant	1.2

Table 2. Adjustment value for sensitivity

Conditions	Sensitivity value
Terms belonging to *tas* 1	1.2
Terms belonging to *tas* 3	1

NOT operator. The positive keywords are selected from the top M ranked terms in the positive user profile. The negative keywords are selected from the top M-N ranked terms in the negative user profile. In the chromosome, the value of the bit (1 or 0) represents the selection or not selection of one keyword.

Spider: This component is used to retrieve Cybercrime related information resided in the web servers of various organizations across countries.

Document Selector: This component is used to determine the matching of the retrieved document to the user's interest and to cooperate with GA in the Query String Generator to retrieve the most relevant documents to the user. It will compare the similarity between the retrieved document and the positive user profile first, then pass the similarity value to the GA in the Query String Generator, and finally generate K most relevant documents to the user. Similarity comparison in this component has been based on vector space model and the Nick's [17] term weighting scheme.

Preprocessing: This component is used to extract the information from the retrieved documents needed for labeling and clustering including steps like replace tags, identify start and end of sentence, identify part-of-speech information and stem.

Label Extractor & Label Processor: This component is used to extract candidate labels based on documents retrieved. The candidate labels could be single term, continuous collocation or interrupted collocation. Label Extractor encompasses two successive steps. The first step applies the string extraction method to extract single term and continuous collocation with variable-length. The second step uses the concept of lexical affinity to retrieve the interrupted collocations. After that, Label Processor is applied to filter out extracted candidate labels with poor grammatical form. More details could refer to our previous work in [6].

Hierarchy Builder: This component is used to build hierarchy of labeled clusters. The clustering algorithm developed here emphasizes the generation of non-redundant browsing hierarchy. In the redundant browsing hierarchy, a child label could be chosen by many parent labels. In some situations, parent labels could have almost the same set of snippets and thus would choose almost identical child labels. Our design of Hierarchy Builder can determine which labels are candidate child labels of specific

parent label first before the choosing of child labels to avoid redundancy. The detail of the design has been described in our previous work in [6].

4 Evaluation

We have conducted some formal tests to study the divergent embodiments of the IR system. Since the data servers of cybercrime related agencies are not currently available, we have turned to the web page on different servers as the data source considering the data characteristics of up-to-date and separately stored in free text format.

We have invited twenty persons possessing a minimum of a bachelor's degree and five years of web search experiences as the subjects. The subject's job was to input to the system the original searching key words, the example document, the retrieved documents and the relevance rating given to the retrieved documents. All our systems are implemented in C++ and performed on Windows XP. The process of the test is as follows:

1) The user inputs some keywords to CyberIR through Information Collector to retrieve some example document and input the example document to CyberIR.
2) CyberIR uses the example document to construct part of the positive user profile and output some documents most similar to the positive user profile to the user.
3) The user browses the retrieved documents and ranks each document as very relevant, relevant, not sure, non-relevant or very non-relevant and inputs the retrieved documents together with the relevance ratings to CyberIR through Information Collector.
4) The Learning Agent uses the retrieved documents and the relevance ratings to construct and maintain the complete positive user profile by modifying values of the attributes 'Term', 'Frequency', 'Sensitivity' and 'Adjusted Frequency', and to construct and maintain the negative user profile by modifying values of the attributes 'N-Term' and 'N-Frequency'.
5) The Learning Agent sorts both the positive user profile by 'Adjusted Frequency' and the negative user profile by 'N-Frequency' in descending order. Then, selects the top M ranked terms from positive user profile and the top N ranked terms from negative user profile and passes the selected "M positive, N negative" terms to the Query String Generator. The Query String Generator, the Document Selector and the Spider work together to produce optimal query string for document retrieval.
6) The Document Selector outputs the ranked list of documents to the user according to the positive user profile learned.
7) The user browses the retrieved documents and rates each document as relevant, not sure, non-relevant. This relevance rating for the retrieved documents is used to measure the performance of relevance retrieval.

As Table 3 shows, several factors that might affect the performance of the system have been studied. Figures 2-5 present the results. They can be summarized as follows. First, more example documents provided could increase the performance of the system. Second, the amount of terms that could make the system perform better is 100. Third, the best of the amount of key words represented by a chromosome is 25. However, between the amount of 10 and 30, the difference is not much. Forth, the best of the amount of negative terms used in the chromosome is 4. However, between the amount of 2-10, the difference is not much.

Table 3. The variables and embodiments

VARIABLE	VALUES OF THE VARIABLE
Amount of example documents provided by the user	1/ 3/ 5/ 7/ 10
Amount of terms of the user profile	30/ 50/ 100/ 150/ 200
Amount of key words represented by a chromosome	10/ 15/ 20/ 25/ 30
Amount of negative terms represented by a chromosome	2/ 4/ 6/ 10

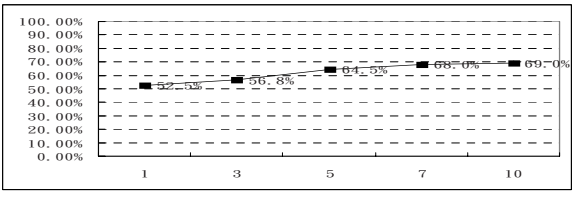


Fig. 2. The effect of the amount of example documents

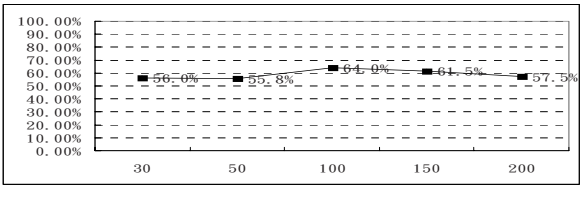


Fig. 3. The effect of the amount of terms of user profile

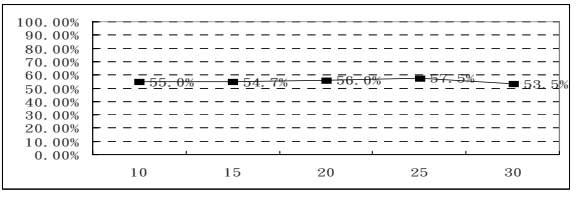


Fig. 4. The effect of the amount of key words

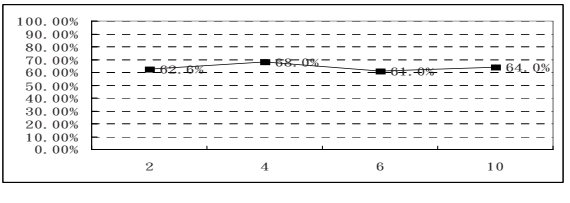


Fig. 5. The effect of the amount of negative term

5 Conclusion

Cybercrime is emerging as a major crime type nowadays. Owing to its characteristics, fighting cybercrime requires international cooperation. The human affairs of legal, organizational and technological approaches across countries are required. Since there are no standards or unification over these human affairs, the accessing and sharing of the related information is difficult. The major contribution of this research is the initiation and demonstration of a method and IR system in the solving of the problem. This beginning step could engender the serious consideration about international cooperation on cybercrime fighting. The other contribution is the examination on the factors that might affect the performance of the system. The data collected could provide more details for reference to the development of this kind of IR system. In the long run, the computer-related crime could be resolved only by computer-related technologies.

Acknowledgements

This work was supported in part by the National Science Council, Taiwan, under the Grant No. NSC96-2416-H-008-011-MY2.

The authors appreciate the anonymous referees for helpful comments and suggestions.

References

[1] Azimi-Sadjadi, M., Salazar, J., Srinivasan, S., Sheedvash, S.: An adaptable connectionist text retrieval system with relevance feedback. In: 2004 IEEE International Joint Conference on Neural Networks, vol. 1, pp. 309–314 (2004)

[2] Balabanovic, M.: An Adaptive Web Page Recommendation Service. In: Proceedings of the First International Conference on Autonomous Agents, New York, pp. 378–385 (1997)

[3] Buckley, C., Salton, G.: Optimization of relevance feedback weights. In: Proceedings of the Eighteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 351–357 (1995)

[4] Chen, H.: Machine Learning for Information Retrieval: Neural Network, Symbolic Learning, and Genetic Algorithms. Journal of the American Society for Information Science 46(3), 194–216 (1995)

[5] Choi, J., Kim, M., Raghavan, V.: Adaptive relevance feedback method of extended Boolean model using hierarchical clustering techniques. Information Processing and management 42(2), 331–349 (2006)

[6] Chou, S., Sun, C., Huang, S.: A Non-Redundant Hierarchical Web Snippet Clustering System to Enhance WWW Search. WSEAS Transactions on Information Science and Applications 4(2), 400–405 (2007)

[7] Chung, W., Chen, H., Chang, W., Chou, S.: Fighting cybercrime: a review and the Taiwan experience. Decision Support Systems 41(3), 669–682 (2006)

[8] Desjardins, G., Godin, R.: Combining Relevance Feedback and Genetic Algorithms in an Internet Information Filtering Engine. In: RIAO 2000 Conference Proceedings, vol. 2, pp. 1676–1685 (2000)

[9] Ekkelenkamp, R., Kraaij, W., Leeuwen, D.: TNO TREC7 Site Report: SDR and Filtering. In: Proceedings of the Seventh Text REtrieval Conference, Gaithersburg, Maryland, pp. 455–462 (1998)
[10] Fidel, R., Crandall, M.: Users' perception of the performance of a filtering system. In: Belkin, N.J. (ed.) Proceedings of the 20th annual international ACM/SIGIR conference on research and development in information retrieval, pp. 198–205 (1997)
[11] Gordon, M.: Probabilistic and Genetic Algorithms for Document Retrieval. Communications of the CAM 31(10), 1208–1218 (1988)
[12] Holland, J.H.: Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control and artificial intelligence. University of Michigan Press (1975)
[13] Justino, E., Bortolozzi, F., Sabourin, R.: A comparison of SVM and HMM classifiers in the off-line signature verification. Pattern Recognition Letters 26(9), 1377–1385 (2005)
[14] Kim, B., Kim, J., Kim, J.: Query term expansion and re-weighting using term co-occurrence similarity and fuzzy inference. In: IFSA World Congress and 20th NAFIPS International Conference, vol. 2, pp. 715–720 (2001)
[15] Koster, C., Beney, J.: On the Importance of Parameter Tuning in Text Categorization. In: Virbitskaite, I., Voronkov, A. (eds.) PSI 2006. LNCS, vol. 4378, pp. 270–283. Springer, Heidelberg (2007)
[16] Moschitti, A.: A Study on Optimal Parameter Tuning for Rocchio Text Classifier. In: Proceedings of the 25th European Conference on Information Retrieval Research, pp. 420–435 (2003)
[17] Nick, Z.Z., Themis, P.: Web Search Using a Genetic Algorithm. IEEE Internet Computing 5(2), 18–26 (2001)
[18] Okabe, M., Yamada, S.: Learning filtering rule sets for ranking refinement in relevance feedback. Knowledge-based systems 18, 117–124 (2005)
[19] Onoda, T., Murata, H., Yamada, S.: Non-Relevance Feedback Document Retrieval based on One Class SVM and SVDD. In: International Joint Conference on Neural Networks, Vancouver, BC, Canada, pp. 1212–1219 (2006)
[20] Pathak, P., Gordon, M., Fan, W.: Effective Information Retrieval using Genetic Algorithms based Matching Functions Adaptation. In: Proceedings of the 33rd Annual Hawaii International Conference on System Sciences, pp. 533–540 (2000)
[21] Rocchio, J.: Document retrieval systems – Optimization and evaluation. Unpublished doctoral dissertation, Harvard University, Cambridge, MA, USA (1966)
[22] Salton, G.: Automatic Information Organization and Retrieval. McGraw-Hill, New York (1968)
[23] Salton, G., Buckley, C.: Improving retrieval performance by relevance feedback. Journal of the American Society for Information Science 41(4), 288–297 (1990)
[24] Tao, D., Li, X., Maybank, S.: Negative Samples Analysis in Relevance Feedback. IEEE Transactions on Knowledge and Data Engineering 19(4), 568–580 (2007)
[25] Vires, A., Roelleke, T.: Relevance Information: A Loss of Entropy but a Gain for IDF? In: Proceedings of SIGIR 2005, Salvador, Brazil, pp. 282–289 (2005)
[26] Wallace, M., Karpouzis, K., Stamou, G., Moschovitis, G., Kollias, S., Schizas, C.: The electronic road: personalized content browsing. IEEE Multimedia 10(4), 49–59 (2003)
[27] Yu, C., Luk, W., Cheung, T.: A Statistical Model for Relevance Feedback in Information Retrieval. Journal of the ACM 23(2), 273–286 (1976)

The Banknote Anti-forgery System Based on Digital Signature Algorithms

Shenghui Su, Yongquan Cai, and Changxiang Shen

College of Computer Science, Beijing University of Technology,
100 Pingleyuan District Chaoyang, 100124 Beijing, P.R. China
{shsu,cyq,shenchx}@bjut.edu.cn

Abstract. Through the combination of digital signature algorithms and integrated circuit chips, the paper puts forwards a new banknote anti-forgery method which may be regarded as an information system. The paper conceives the preliminary design of the system which includes the banknote print-data editing, key management, signing-code injection, and signing-code verification four modules, and describes the function of every module. The paper analyzes the security and effectiveness of the new method, discusses curtly the digital signature algorithms with a small modulus, and points out that the thought of the banknote anti-forgery system may be extended to passport anti-forgery for antiterrorism.

Keywords: Digital signature, Anti-forgery, Preliminary design, Verification.

1 Introduction

In 1976, the concept of public key cryptography was first put forwards by Diffie and Hellman [1]. A public key cryptosystem contains a key generator, an encryption algorithm and a decryption algorithm. A public key which may be distributed to everyone in the domain, and is applied to encryption. A private key which is individually kept only by the user himself, and is applied to decryption. Furthermore, It is infeasible to infer a corresponding private key from a public key. By now, the applicability of public key cryptosystems has been extended to digital signature and identity verification for authenticating a sender and preventing messages from being denied by the sender or modified by attackers [2][3].

Although an electronic currency has already come forth, paper currencies, namely banknotes, as one of the age-oldest and most convenient trade mediums, still play an irreplaceable role today, and also will not disappear in the future. However, banknotes counterfeited enormously have brought great losses to world economies. Because existing physical and optical anti-forgery means can't suppress banknote counterfeit thoroughly, seeking for new and breaking-through anti-forgery techniques has become nondeferrable affairs of nearly every country's center bank. In this paper, a banknote anti-forgery system shortly called BAFS is proposed, and the effectiveness and advantage of BAFS are discussed.

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 44–49, 2008.

2 Basic Techniques for the Banknote Anti-forgery System

2.1 IC Chip Techniques

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 [4][5]. Such a size makes it possible to embed a chip into a banknote. Every IC chip is set a device number according to itself hardware circuit by the manufacturing plant. The IC device number is unique in the world and can't be modified forever. Therefore, a banknote embedded with an IC chip will possess one uniquely identifiable physical characteristic that is equivalent to a biological characteristic and especially important.

IC chips are divided into two sorts — touch-sensitive and non-touch-sensitive. Each of the IC chips that are used in BAFS belongs to the non-touch-sensitive, stores only data, has no CPU for computing, and exchanges information through radios or electromagnetic waves. After an IC chip is embedded into a banknote, it must be solidified specially in order that it will not be worn away while the banknote is in circulation. A banknote embedded with an IC chip is called an IC banknote.

2.2 Public Key Signature Algorithms

The public key cryptosystems RSA [6] and ECC [7] have already been employed in the many applications in recent years, where ECC is an analogue of ElGamal [8], and constructed in an ellipse curve group over a finite field.

In BAFS, we only need the key generation, signature and verification algorithms of a public key cryptosystem. In addition, a Hash one-way function is needed for obtaining the digest of a message printed on a banknote face. The signature algorithm converts the digest into a signing-code with a private key. The verification algorithm authenticates true or false banknotes with a public key. The security of the cryptosystem is the linchpin of BAFS playing really an anti-forgery role.

The Hash function has the following properties [3][9]: (1) Given the face message m of a banknote, to calculate the digest is very easy. (2) Given the message digest y, it is computationally infeasible to find a message m satisfying $\mathrm{Hash}(m) = y$. (3) It is computationally infeasible to find the two different face message m_1 and m_2 satisfying $\mathrm{Hash}(m_1) = \mathrm{Hash}(m_2)$. This condition requires Hash to be strongly collision-free. Therefore, the face message digest of a banknote is uniquely determined by Hash, and there is hardly any identical digest between arbitrary two banknotes.

3 Preliminary Design of the Banknote Anti-forgery System

BAFS is composed of hardware and software two parts. Hardware contains general computers, IC chip read-write machines and IC banknote checkers. Software contains the banknote print-data editing module, the key management module, the signing-code injection module, and the signing-code verification module.

3.1 Banknote Print-Data Editing Module

The banknote print-data editing module runs on a computer off-internet, and is employed by the national center bank. It should do the following tasks:

(1) Create a database file called BPD for storing banknote print-data: center bank name, banknote version, print order, par value, print date, printery, start currency number, end currency number, public key, etc. The database records should be sorted according to the banknote version and print order.
(2) Input all the field values to the database file. The print-data of banknotes per version and time corresponds to a record in the database file.
(3) Modify a record of the database via an operator authorized.
(4) Delete a record of the database via an operator authorized.

Notice that BPD needs to be preserved for a long time so as to be provided for management and decision of the center bank administrators.

3.2 Key Management Module

This module runs on a computer off-internet, is employed by the national center bank, and has a high secrecy. Its tasks should be as follows:

(1) Set the parameters of the selected public key cryptosystem.
(2) Produce a public key and a private key through the key generation algorithm.
(3) Output the private key to an USB flash disk, which is kept by a kernel official and must not be divulged.
(4) Store to BPD the public key which is provided for the signing-code injection module.

Notice that the IC banknotes having the different version numbers should use the different key pairs, and the IC banknotes having the different print orders but the same version numbers should also use the different key pairs.

3.3 Signing-Code Injection Module

Before the signing-code begins to be injected, the IC chip embedding and banknote printing should be finished in the banknote printery. The selected IC chip is non-touch-sensitive and writing-once-reading-many. The start and end currency numbers of the banknotes per version must be the same as the corresponding field values of the relevant record in BPD.

A combination of the bank name, par value, printing year and currency number on a banknote is called a face message. The first three items and the start number can be obtained directly from the relevant record in BPD, and we get the currency number by adding the start number to the sequence number. The sequence numbers begin with 0. This module is employed by the center bank, has a high secrecy, and runs off-internet. The computer has a copy of BPD and connects with an IC chip read-write machine. The module should do the following tasks:

(1) Capture the device number of the IC chip embedded in a banknote through the read-write machine.

(2) Read the values of the fields bank name, par value, printing year and start number of the record from BPD, according to a banknote version and printing order. Compute the currency number by adding the start number to the sequence number, and then let the sequence number increase 1.
(3) Calculate the digest of the face message and IC device number by the Hash function. Namely, Calculate Hash(face message, IC device number).
(4) Read from a flash disk the private key corresponding to the public key in BPD.
(5) Regard the message digest and private key as the parameters, and produce the signing-code of the banknote by the digital signature algorithm.
(6) Write the signing-code, public key and face message into the IC chip embedded in the banknote via the read-write machine.

Notice that when the signing-code is injected, we should sort the IC banknotes by the ascending order of currency numbers so as to warrant that a signing-code matches an IC banknote in currency number. The private key should to avoid being revealed.

3.4 Signing-Code Verification Module

This module as an embedded program runs in a banknote checker, which has an inductor for reading data from the IC chip and 16-bit CPU for calculating. The checker finishes verifications simultaneously while it counts IC banknotes.

The banknote checkers may be divided into two types. One of them is expensive. The other is cheap and portable, only can handle an IC banknote at a time, and is provided for small retailers and individuals. The module should do the following tasks:

(1) Capture the signing-code, face message, IC device number and public key from a static or moving IC banknote through the inductive head.
(2) Calculate the digest of the face message and IC device number by the Hash function. Namely, Calculate Hash(face message, IC device number).
(3) Regard the message digest, private key and signing-code as the parameters of the public key verification algorithm, and obtain the verification value.
(4) If the verification value satisfies the given condition, then the IC banknote is true. Otherwise the IC banknote is forged, and a sound alarm is sent to users.

The verification process of an IC banknote checker looks the same as that of an optical checker. Its verification speed rests with the public key verification algorithm.

4 Analyses of the Effectiveness of BAFS

We will see that the effectiveness of BAFS is irreplaceable. BAFS based on digital signature algorithms should be the thorough anti-forgery method.

4.1 Comparing Directly IC Device Numbers is Infeasible

It is known from section 2.1 that the device number of an IC chip is unique in the world. Since the IC chip is embedded in a banknote, the banknote also owns uniqueness in the world.

Then, why does the IC banknote need a signature by the national center bank yet? The reasons are as follows: (1) The banknote circulation is very vast. If IC device numbers are compared and checked directly, then the center bank must aggregate the device number of every IC banknote into a database provided for checking. This database will be huge and need an especially gigantic storage space. (2) When checking, a user needs to seek the device number of the IC banknote from the database. Obviously, the database is too large to be placed in a banknote checker, hence the user can only telnet to the database placed at the center bank. Telneting needs time. In addition, it also needs time to seek a specific IC device number from the huge database. (3) The IC device number database is open to every visitor so that its security can't be guaranteed. Checking work will be unable to proceed at all in case the data in the database is destroyed or divulged.

Therefore, it is infeasible in time and space directly to compare and check IC device numbers, and risky in security.

4.2 Adopting a Symmetric Key Cryptosystem Is Insecure

It is well known that commercial banks usually select relatively fast symmetric key cryptosystems to encrypt data or authenticate identities in financial applications based on IC cards.

Then, why don't we use a symmetric cryptosystem in BAFS? This is based on the following considerations: (1) Banknotes are circulatable and exchangeable while IC cards are inexchangeable and can only be held fixedly by users. (2) The user terminals ─ ATM for example of IC card system are owned and supervised generally by the commercial bank itself, which can protect the security of the symmetric cryptosystem. However, IC banknote checkers are held either by commercial banks or by individuals. If the symmetric key is placed in a banknote checker, it is fully possible that the symmetric key is disclosed. (3) If a symmetric key for signatures is changed, its counterpart in every banknote checker also needs to be replaced correspondingly. Because the banknote checkers owned by users are extremely numerous, and the human resources of the checker manufacturer are limited, it is almost impossible for the manufacturer to finish all the replacement work. (4) The symmetric key appears in both signatures and verifications, and is disclosed potentially at each work stage. It is very difficult for the center bank to supervise the symmetric key effectually.

So, the difficulties in distributing and managing keys make it infeasible to use a symmetric key cryptosystem in BAFS.

4.3 A Public Key Cryptosystem Is an Inevitable Selection

Comparatively, a public key cryptosystem has the following advantages:

Firstly, the large-scale database for verifying signing-codes isn't necessary since signing-codes are stored in the IC chips of the banknotes respectively. Secondly, the signing and verifying are to use two different keys, it is convenient to distribute and management them. The security of the system is upgraded to the greatest degree. Lastly, the public key is stored in the IC chip of a banknote, and is irrelevant to an IC

banknote checker. Even if the public key is changed according to the private key by the center bank, all the banknote checkers don't need to be changed or maintained.

A public key cryptosystem has a slower execution speed than a symmetric key cryptosystem does, but the execution speed of BAFS will be able to meet the users' practicable requirement if we select a suitable cryptosystem and its parameters.

5 Prospects

For the center bank, the cost of IC banknotes come principally from IC chips, but the center bank issuing IC banknotes saves the cost of the physical and optical anti-forgery technologies at the same time. The two costs are equivalent by and large. Like personal computers, with the popularization and slathering of IC chips, the price of an IC chip will decline further.

The integration degree of an IC chip is more and more high, and is striding into a nanometer scale from a micron scale. At present, an IC chip scaled in microns is fit to be embedded in a banknote. Thus, the physical size of an IC chip will not become a problem of developing BAFS.

The modular length has the great influences on banknote checking, which requires us to manage to design out a public key cryptosystem which bears a relatively small modulus ─ 128-bit modulus for example. If so, we may conceive a spare signature scheme. The spare signing-code will be used only when the IC chip in an IC banknote is damaged. Moreover, the damaged IC banknote must be submitted to the national center bank by its holder for security consideration, and the verification of the spare signing-code can only be performed by the center bank.

Obviously, the thought of and the technique for BAFS may be extended to passport anti-forgery which is an urgent demand from international antiterrorism.

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. Rabaey, J.M., Chandrakasan, A., Nikolic, B.: Digital Integrated Circuits, 2nd edn. Prentice-Hall, New Jersey (2002)
5. Martin, K.: Digital Integrated Circuit Design. Oxford University Press, Oxford (1999)
6. 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)
7. Yan, S.Y.: Number Theory for Computing, 2nd edn. Springer, New York (2002)
8. ElGamal, T.: A Public-key Cryptosystem and a Signature Scheme Based on Discrete Logarithms. IEEE Transactions on Information Theory 31(4), 469–472 (1985)
9. Stallings, W.: Cryptography and Network Security: Principles and Practice, 2nd edn. Prentice-Hall, New Jersey (1999)

Sequence Matching for Suspicious Activity Detection in Anti-Money Laundering

Xuan Liu[1], Pengzhu Zhang[1], and Dajun Zeng[2]

[1] Department of Management Information System
Shanghai Jiao Tong University, Shanghai, China
{amethyst,pzzhang}@sjtu.edu.cn
[2] Department of management information systems
University of Arizona, Tucson, Arizona
zeng@email.arizona.edu

Abstract. Developing effective suspicious activity detection methods has become an increasingly critical problem for governments and financial institutions in their efforts to fight money laundering. Previous anti-money laundering (AML) systems were mostly rule-based systems which suffered from low efficiency and could can be easily learned and evaded by money launders. Recently researchers have begun to use machine learning methods to solve the suspicious activity detection problem. However nearly all these methods focus on detecting suspicious activities on accounts or individual level. In this paper we propose a sequence matching based algorithm to identify suspicious sequences in transactions. Our method aims to pick out suspicious transaction sequences using two kinds of information as reference sequences: 1) individual account's transaction history and 2) transaction information from other accounts in a peer group. By introducing the reference sequences, we can combat those who want to evade regulations by simply learning and adapting reporting criteria, and easily detect suspicious patterns. The initial results show that our approach is highly accurate.

Keywords: Anti-money laundering, suspicious activity detection, SARs, Euclidean distance, sequence matching.

1 Introduction

Money laundering (ML) has been recognized as a critical problem with serious economic and social ramifications. ML is always accompanied with by crimes, such as drug trafficking, financial terrorism, corruption, and so on. Federal agencies estimate that as much as $300 billion is laundered annually, worldwide [1].

To combat ML, governments and financial institutions, including financial information agencies, various regulatory agencies, and law enforcement agencies have been working together, and a number of different approaches have been attempted. [2]. Among these approaches, suspicious activity detection plays an important role as it generates an initial set of potential money laundering cases for further investigation. In

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 50–61, 2008.

addition, it is helpful in summarizing money laundering trends and patterns for prevention [3]. During suspicious activity detection, domain experts in each financial institution are required to extract suspicious activity reports (SARs) according to the suspicious criteria, and deliver those reports to a supervision center. Experts in the supervision centers combine reports from different financial institutions along with other information to further isolate highly suspicious ML related activities. Recently AML systems have been developed in most financial institutions to help extract SARs, however, major technical gaps exist when developing an effective and efficient method to detect suspicious activities from a large quantity of transactional information streams. Surveys from China have shown that until recently most of the AML systems running for SARs were based on human-set thresholds, which suffered from both low false-positives and low true-positives for reporting SARs. They have also had difficulty explaining the meaning of the suspicious transactions.

This paper focuses on developing a computational approach to enable financial institutions to classify normal and suspicious sequences from their large volume of transactions. To achieve it, we introduce a sequence matching method, define the target sequences as 'query sequences', and introduce two kinds of information as 'reference sequences': historical information in the same accounts' transactions, and transactions of other accounts in the same peer group. There are two major advantages to this approach: first of all, by viewing transactions as temporal sequences, it is much easier to characterize them as fraudulent than it is to characterize an individual's single transaction as fraudulent. For example, a single deposit of just under $10,000 may not be suspicious, but multiple such deposits are and a large deposit may not be suspicious, but followed closely by withdrawals it will be quite suspicious. Secondly, by comparing query sequences with historical information it is possible to detect suspicious activity according to its own transaction trend or temporal pattern rather than by simply picking out those transaction points exceeding human-set thresholds. Thus it is possible to reduce the number of incorrectly detected cases since a human-set threshold is usually determined statistically and it is only suitable for average cases. Furthermore, introducing peer group information (a widely used technique in other fraud detection application domains such as computer intrusion and telecommunication fraud) to AML, can unveil laundering techniques such as constructing "structure transactions" or "training their behavior patterns" to hide anomaly transactions.

The organization of this paper is as follows: Section 2 introduces the existing suspicious activity detection methods as well as algorithms used in time series analysis. Section 3 proposes our framework for suspicious activity detection. Section 4 and Section 5 summarize the experiment design and results. The Section 6 concludes the paper and discusses future work.

2 Literature Review

2.1 Suspicious Activity Detection

Suspicious activity detection can be viewed as an outlier detection problem, in which, most of the methods are categorized into threshold-based detection and state-based detection. In AML realities, most AML detection systems use human-set thresholds to filter those suspicious transactions according to survey result from China[2]. Among

the more elaborate AML systems, the American National Association of Securities Dealers, Inc. uses break detection technology to detect abnormal stock transactions[4] and the U.S. Financial Crimes Enforcement Network AI System (FAIS) uses suspicious scores to flag certain types of transactions and activities, and it also utilizes link analysis technology to detect related crimes[5, 6].

Some researchers focus on using machine learning methods to help detect suspicious activities. For example, Tang [7] used SVM to deal with labeled transaction data in order to find out suspicious activity and Wang [8] introduced decision trees to customer AML risk assessment using manually labeled examples to train the trees. Also there are papers focusing on finding suspicious accounts whose transaction tendencies were differentiated from other customers in the same industry[9, 10].

Most papers that we reviewed focused on suspicious customer and account detection, while for AML suspicious retail transactions must be reported. At the same time, both supervised and unsupervised learning methods are used, but because of a lack of truly suspicious data, evaluations in those papers are weak.

2.2 Time Series Analysis and Sequence Related Work

Time series analysis has been used in many financial problems. Control charts, breakpoint analysis [11], and change-point detection [12] are among the popular methods designed to find bursts in a single time series.

Different from those three methods mentioned above, sequence matching is a method focusing on a comparison between sequences in time series databases, and it has been used in stock analysis and computer intrusion [13-15]. It helps to identify other days in which stock X had movements similar to the current movements, and it then further classifies objects that have similar temporal patterns into different groups. There are two kinds of sequences in sequence matching: query sequences and reference sequences, and these can have different meanings depending on the application domain. One of the main jobs in sequence matching is defining similarity measures [16]. Claudia discussed different similarity measures for sequences [17], and the most used in time-domain continuous representations is the Euclidean distance, which is determined by viewing each sub-sequence with n points as a point in R^n.

Although used in other applications as mentioned, to our knowledge, sequence matching has not been used in suspicious transaction detection. In this paper, we use the classical Euclidean similarity distance to define similarity between different sequences, while using a threshold-based method to classify normal and suspicious sequences.

3 A Framework for Suspicious Activity Detection Using Sequence Matching

Our algorithm differs from existing machine-learning methods used in AML in that it focuses on assisting financial institutions to detect suspicious sequences as required by supervision centers, rather than by detecting suspicious accounts or customers. Meanwhile, compared with other time series analysis methods, our method combines sequence matching and classification methods (see details in Fig. 1). In order to map an AML problem into a sequence learning problem, we form the query sequences by

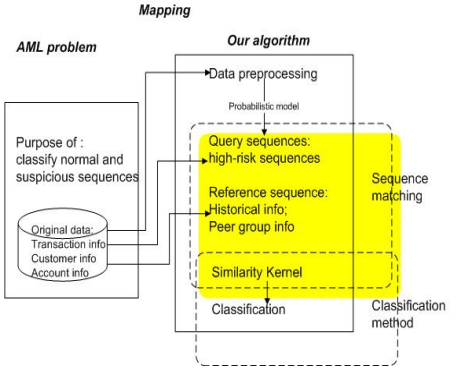


Fig. 1. Relations between our algorithm and sequence matching and classification methods

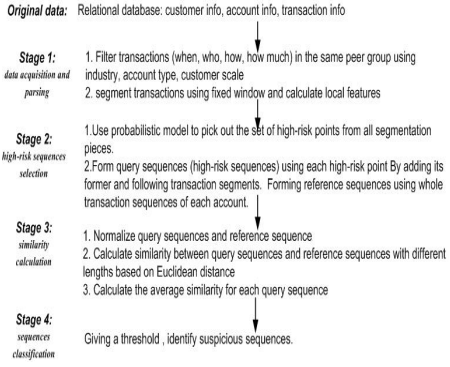


Fig. 2. Framework for suspicious activity detection using sequence matching

selecting high-risk sequences in an account's transaction sequences using a probabilistic model. To create reference sequences, we use historical information and peer group information because we identify a suspicious activity based on the hypothesis that these transactions are different from normal accounts' transaction activities in the same peer group, as discussed in the first section. For similarity measures we use classic Euclidean distance of local features pairs for different sequences. After calculating the similarities between different sequences, we can further classify those sequences into normal and suspicious.

Our algorithm contains four steps: data acquisition and parsing, high-risk sequences selection, similarity calculation between high-risk sequences and reference sequences, and sequences classification. See Fig. 2.

During the data acquisition and parsing stages, transaction data within the same peer group are filtered from the financial database. Following the previous literature, we use those categorized attributes such as industry, account type, and company size to determine the peer group. The resulting transaction activities include the following information: who (transaction account), when (transaction time), how (operation type) and how much (the amount of money). Then for all accounts in the same peer group, we use a window with fixed size to segment the original transaction information pieces and to extract local features for further stages of analysis. In our study, we chose a relatively small widow and segmented transactions into a daily level. For each piece of an account $A_i \in A$ (A is the set of accounts in database), the local features $F = < f_1, f_2, \dots f_k >$ (k is the dimension of local features),which could be user-defined by domain knowledge, prior AML literature usually includes: <frequencies of transactions, frequencies of flow-in transactions, frequencies of flow-out transactions, amount of transactions, amount of flow-in transactions, amount of flow-out transactions >. Transaction temporal activities of A_i after segmentation are as follows $< t_1, q_{1i1}, q_{2i1}, \dots q_{ki1} > \cdots < t_l, q_{1il}, q_{2il}, \dots q_{kil} > \cdots < t_{l_i}, q_{1il_i}, q_{2il_i}, \dots q_{kil_i} >$, for different accounts, the total number of segment pieces could be different because of the customers' registration period, let l_i represents the number of segmentations pieces for A_i.

In the high-risk sequence selection stage, we use two steps to form the high-risk sequences. 1) The first step is to select high-risk pieces using a threshold-base probabilistic model. Sequence matching in a large scale database is usually time consuming, thus the main purpose of this step is to select the most likely high-risk points for forming query sequences. For each feature $f_j \in F$, given a threshold α_j, using all the pieces formed by stage one $S_{f_j} = \{q_{jil} | 0 < i \leq |A|, 0 \leq l \leq l_i\}$ (in which $|A|$ represents the cardinality of set A), we can get the set of high-risk piecesSP_{f_j} according to following equations (1a) (lb):

$$SP_{f_j} = \left\{ \langle i, l \rangle | 0 < j \leq k, 0 < i \leq |A|, 0 \leq l \leq l_i, q_{jil} \geq T_{f_j} \right\} \tag{1a}$$

$$P\left(q_{jil} \geq T_{f_j}\right) = 1 - \alpha_j \tag{1b}$$

By setting different thresholds for different features, we can easily adjust the weights for each feature. Let SP be the set of candidate pieces to form high-risk sequences, by calculating the union of all SP_{f_j}, see (2a). Every point in SP has potential to be recognized as suspicious just because their transactions at the segmentation point are statistically abnormal, either their transaction amount is relatively large or because they exhibit relatively high frequencies within a short time. Nevertheless, we need to view each piece in the context of temporal sequences to confirm the intention behind the activities. Thus for further using sequence matching method, we partition SP according to accounts. For each $A_i \in A$, we can get certain high-risk segments dispersed within the whole transaction sequences, see (2b).

$$SP = \cup_{j=1}^{k} SP_{f_j} \tag{2a}$$

$$SP = SP_A = \cup_{i=1}^{|A|} SP_{A_i} \tag{2b}$$

2) The second step is to form query sequences and reference sequences. On one hand, we form query sequence according to high risk points in setSP. In this study, we use a simple method to form high-risk sequences for each piece in SP by adding adjacent transactions. Given a sequence length $2e + 1$ for query sequences, for each account $A_i \in A$, each point $\langle i, l \rangle \in SP_{A_i}$, the forming query sequence could be represented as $SEQ(A_{il})$, which is a $k * (2e + 1)$ metrics, see equations (3a) (3b):

$$SEQ(A_{il}) = \begin{pmatrix} SEQ_1(A_{il}) \\ \vdots \\ SEQ_j(A_{il}) \\ \vdots \\ SEQ_k(A_{il}) \end{pmatrix}_{k*(2e+1)} \tag{3a}$$

$$SEQ_j(A_{il}) = (q_{ji(l-e)} \quad \cdots \quad q_{jil} \quad \cdots \quad q_{ji(l+e)}) \tag{3b}$$

On the other hand, we form reference sequences using all the accounts in set A. For $SEQ(A_{il})$, which is a certain segment of A_i, the reference sequences includes two kinds, 1) using A_i to get reference sequence $SEQ'(A_i)$, it represents the transaction history of the account itself; 2) using $A_j \left(j \neq i, 0 A_j \in A\right)$ to get reference sequences $SEQ'\left(A_j\right)$, we can further compare query sequences

$SEQ(A_{il})$ with the transaction sequences in its peer group. Thus for each A_{il}, there are a total of $|A|$ reference sequences, see (4a), (4b), (4c).

$$\cup_{j=1}^{|A|} SEQ'(A_j) = \left(\cup_{j=1,j\neq i}^{|A|} SEQ'(A_j)\right) \cup SEQ'(A_i) \tag{4a}$$

$$SEQ'(A_r) = \begin{pmatrix} SEQ_1{}'(A_r) \\ \vdots \\ SEQ_j{}'(A_r) \\ \vdots \\ SEQ_k{}'(A_r) \end{pmatrix}_{k*l_r} \tag{4b}$$

$$SEQ_j{}'(A_r) = (q_{ji1} \quad \cdots \quad q_{jil} \quad \cdots \quad q_{jil_r}) \tag{4c}$$

In the similarity calculation stage, we calculate similarities between each query sequence $SEQ(A_{il})$ and each reference sequence $SEQ'(A_r)$. Before calculating sequence similarities, we normalize all the query sequences and reference sequences to eliminate noise or short-term fluctuation as stated in [15, 18], then we get the normalized reference sequence $(SEQ'(A_r))_{norm}$ and query sequence $(SEQ(A_{il}))_{norm}$. See (5a) (5b) (5c) (5d).

$$(SEQ(A_{il}))_{norm} = \begin{pmatrix} SEQ_1(A_{il}))_{norm} \\ \vdots \\ (SEQ_j(A_{il}))_{norm} \\ \vdots \\ (SEQ_k(A_{il}))_{norm} \end{pmatrix}_{k*(2e+1)} \tag{5a}$$

$$SEQ_j(A_{il}))_{norm} = \left(\frac{q_{ji(l-e)}-\mu_{ij}}{\sigma_{ij}} \quad \cdots \quad \frac{q_{jil}-\mu_{ij}}{\sigma_{ij}} \quad \cdots \quad \frac{q_{ji(l+e)}-\mu_{ij}}{\sigma_{ij}}\right) \tag{5b}$$

$$(SEQ'(A_r))_{norm} = \begin{pmatrix} (SEQ_1{}'(A_r))_{norm} \\ \vdots \\ ((SEQ_j{}'(A_r))_{norm} \\ \vdots \\ (SEQ_k{}'(A_r))_{norm} \end{pmatrix}_{k*l_r} \tag{5c}$$

$$(SEQ_j{}'(A_r))_{norm} = \left(\frac{q_{ji1}-\mu_{rj}}{\sigma_{rj}} \quad \cdots \quad \frac{q_{jil}-\mu_{rj}}{\sigma_{rj}} \quad \cdots \quad \frac{q_{jil_r}-\mu_{rj}}{\sigma_{rj}}\right) \tag{5d}$$

After that, Euclidean distance measure is used to calculate similarities between query sequences and reference sequences. There are two kernels which need to be defined in similarity calculation. 1) Kernel to calculate similarities of sequences with different lengths. 2) Kernel to combine all the features. For the former, because query sequences are with the length of $2e+1$, while different reference sequences may have distinct lengths l_r, we need to define a similarity between sequences with different lengths. Let $D((SEQ(A_{il}))_{norm}, (SEQ'(A_r))_{norm})$ present the similarity between two normalization metrics $(SEQ(A_{il}))_{norm}$,$(SEQ'(A_r))_{norm}$, see equation (6a).

$$D((SEQ(A_{il}))_{norm}, (SEQ'(A_r))_{norm}) = \begin{cases} Euclidean(SEQ(A_{il}))_{norm}, (SEQ'(A_r))_{norm}), \text{if } 2e+1 = l_r \\ \frac{\sum_{n=1}^{l_r-2e} D\left(SEQ(A_{il}))_{norm}, ((SEQ'(A_r))_{norm})_{[n..(n+2e)]}\right)}{l_r / 2e+1}, \text{if } 2e+1 < l_r \\ \frac{\sum_{n=l-k}^{l+k-l_r} D\left((SEQ(A_{il}))_{norm})_{[n...(n+l_r)]}, (SEQ'(A_r))_{norm}\right)}{l_r / 2e+1}, \text{if } 2e+1 > l_r \end{cases} \quad (6a)$$

$$Euclidean(SEQ(A_{il}))_{norm}, (SEQ'(A_r))_{norm}) = \sqrt{\sum_{j=1}^{k} w_j ((SEQ_j(A_{il}))_{norm} - (SEQ_j{}'(A_r))_{norm})^2} \quad (6b)$$

As for the later, we use different weight for w_j different feature in calculating the similarity f_j. In this study, we assume each feature to have the same weight in determining the similarity. Finally, we calculate average similarities between $SEQ(A_{il})$ and all reference sequences of accounts set A, named $F_{sim}(SEQ(A_{il}), A)$, see (7):

$$F_{sim}(SEQ(A_{il}), A) = \frac{1}{r}\sum_{r=1}^{r=|A|} D((SEQ(A_{il}))_{norm}, (SEQ'(A_r))_{norm}) \quad (7)$$

In the final stage sequence classification, we order the similarity of each candidate sequence, giving a threshold T_s, choose the relatively higher ones, and flag them in the original database as suspicious.

4 Experiment Design

4.1 Research Tested

In order to examine the proposed algorithm, we conducted two experimental studies: in the first experiment we use human-set thresholds suggested by financial institutions, and in the second experiment we use our algorithm following which we compare the results of two experiments. The data we use in the experiments are real financial data from a Chinese financial institution. We select transaction data for individual customers and view them as in the same peer group. The data are labeled with two categories: normal and suspicious. For normal data we randomly select 640 accounts and all their transactions from Year 1995 to Year 2001. Data fields include transaction time, account number, transaction direction, and transaction amount, with a total of 120,986 original transaction records. The suspicious category includes those reported from the

commercial bank and identified by the supervision center as suspicious. There are a total of 64 accounts and 1,940 transaction records, and each account with transactions is from a different time period, see Fig. 3 and Fig. 4. The transferring days refer to the number of total days transaction activities occurred for certain accounts.

In the first experiment, we use human-set thresholds which are utilized in most of the current AML detection systems in China. We use on thresholds of quantity criteria to extract suspicious transactions as our benchmark. In the second experiment, we carefully select parameters to conduct the experiment aiming to get higher performance. See more details in Section 5.2.

4.2 Evaluation Metrics

We use standard classification performance metrics, sensitivity/recall and specificity to evaluate the experiment result. In which true positives are those suspicious records and correctly detected as suspicious and false positives represent those normal transactions

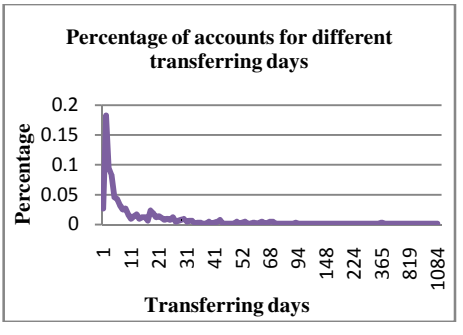


Fig. 3. Distribution of normal accounts with different transferring days

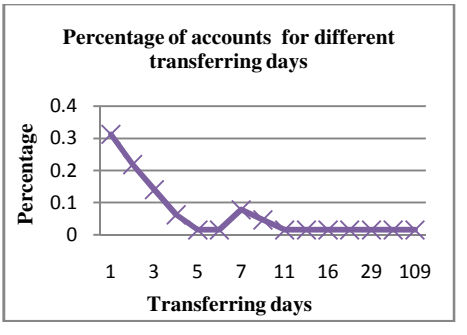


Fig. 4. Distribution of suspicious accounts with different transferring days

which are incorrectly detected as suspicious ones, the higher the sensitivity and specificity are, the better the performance of the algorithm. See (8a) (8b).

$$\text{sensitivity} = \frac{\text{number of true positives}}{\text{number of true postivites+number of false negatives}} \tag{8a}$$

$$\text{specificity} = \frac{\text{number of true negatives}}{\text{number of true negatives+number of false positives}} \tag{8b}$$

5 Experiment and Result Analysis

5.1 Overall Performance Comparison

Using the criteria suggested by experts in financial institutions, we select transactions from two perspectives: 1) single transaction exceeds certain threshold T_1 ; 2) the total

transaction amount exceed T_2 within a day. The results are displayed in Table 1. From the detection results we can see that the main problem of existing AML detection methods used for suspicious transaction extraction is the low specificity- large volume of false positive transactions reported. This is followed by simple quantity criteria, which imposes a large burden on experts in supervision centers to eliminate those useless reports after experts combine reports from different financial institutions and further isolate highly suspicious ML related activities

Table 1. Result of benchmark

(T_1, T_2)(RMB)	Sensitivity	Specificity
(200,000, 200,000)	0.988	0.332
(500,000, 500,000)	0.954	0.442

In the second experiment on the algorithm we proposed, we select daily segmentation as our algorithm, and after the preprocessing, we obtain 31,127 segments, in which 439 are suspicious ones. In stage two, we use six features <total amount, flow in amount, flow out amount, total frequency, flow in frequency, flow out frequency> to select the high-potential segments, and for computational reason, we select two thresholds for each feature: $\alpha_j = 0.8$ and $\alpha_j = 0.9$. In stage three, we choose $e = 2$ so that each query sequence has the length of 5 days. In stage four, we calculate similarity, using a different similarity threshold T_s . The results are shown in Table 2. The lower sensitivity compared with the benchmark has two causes: 1) in stage two, we chose a relatively high threshold, and thus not enough high-risk points are selected for further suspicious sequences detection; when choosing a lower threshold, the sensitivity will be higher. 2) Using $\alpha_j = 0.8$ $T_s = 0.6$ as an example, most of the undetected sequences, when we check their accounts, we found that all of them not longitudinal enough- they all have transaction information occurred in no more than 2 days.

Table 2. Result for our algorithm when α_j =0.8 and α_j =0.9

$T_s(\alpha_j = 0.8)$	Sensitivity	Specificity	$T_s(\alpha_j = 0.9)$	Sensitivity	Specificity
0.8	0.759	0.704	0.8	0. 659	0.732
0.6	0.830	0.648	0.6	0. 769	0. 628

5.2 Discussion and Parameters Selection in Our Algorithm

The performance of our algorithm can be further improved by adjusting parameters used in different stages in our algorithm. In the second stage, threshold α_j can influence the time complexity for similarity calculations. The lower the thresholds are, the larger the cardinality of the set SP (see details in Table 3), cause the similarity complex $O(|SP| * |A|)$. Meanwhile, the other parameter $2e + 1$ represents the

length for the query sequences, this parameter can be regarded as the smallest period needed to accomplish certain phrase of money laundering cycle (placement, layer and integration) .

In stage three, we can further change the similarity kernel for better performance. In this paper we calculate the similarity without taking into consideration the influence of previous similarities. Meanwhile, to simplify the problem, we give each feature the same weight when define similarity. One of the possible ways of improving the algorithm performance is to assign different weights to different attributes, and a second is to calculate similarity based on Bayesian probability model – update similarity using previous similarity info.

Table 3. Different threshold in stage two and their extraction results

α_j	|SP|	True negative segments	False positive segments
0.9	9606	271	9335
0.8	11675	384	11291
0.6	27629	433	27196

6 Conclusion and Future Work

In this paper, we propose a novel algorithm that can be used in suspicious activity detection. We use historical information and customer information to help detect suspicious sequences, formulating the problem as a temporal sequence matching problem. In order to balance the efficiency and the correctness of the detection, we first use a probabilistic model to get the high-risk transaction segments which provides input to the subsequential sequence matching stages. We then use Euclidean distance to define similarities between two sequences with different lengths. The algorithm we proposed is quite flexible and allows for setting different parameters in each step without changing the entire working flow. The experiment results show that by selecting different detection features and adjusting thresholds in different steps, we get high sensitivity and specificity.

Future work includes 1) further study of the influences of different features on suspicious activity detection, 2) extension of our approach to include a network of accounts for suspicious activity detection.

Acknowledgements

This work is supported by the NNSFC #70533030, #60621001, #60573078, MOST #2006CB705500, #2004CB318103, #2006AA010106, and CAS #2F05N01, #2F07C01. During the writing and editing process, Xin Li from University of Arizona, and Dr. Dick Solie provides us with valuable feedback.

References

1. U.S. Congress, Information technologies for the control of money laundering. Office of Technology Assessment, Report OTA-ITC-630, U.S. Government Printing Office, Washington, DC (1995)
2. Liu, X., Zhang, P.: Research on Constraints in Anti-Money Laundering (AML) Business Process in China Based on Theory of Constraints. In: Hawaii International Conference on System Sciences, Proceedings of the 41st Annual, p. 213 (2008)
3. Force, F.A.T.: Report on Money Laundering Typologies, 2000-2001 (2002) English ed. Accessed
4. Senator, T.E.: Ongoing management and application of discovered knowledge in a large regulatory organization: a case study of the use and impact of NASD Regulation's Advanced Detection System (RADS). In: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 44–53 (2000)
5. Goldberg, H., Senator, T.E.: Restructuring databases for knowledge discovery by consolidation and link formation. In: Proceedings of 1998 AAAI Fall Symposium on Artificial Intelligence and Link Analysis (1998)
6. Senator, T.E., et al.: The financial crimes enforcement network AI system(FAIS): identifying potential money laundering from reports of large cash transactions. The AI magazine 16(4), 21–39 (1995)
7. Tang, J., Yin, J.: Developing an Intelligent Data Discriminating System of Anti-Money Laundering Based on SVM. In: Proceedings of 2005 International Conference on Machine Learning and Cybernetics, 2005 (2005)
8. Wang, S.N., Yang, J.G.: A Money Laundering Risk Evaluation Method Based on Decision Tree. In: International Conference on Machine Learning and Cybernetics, 2007 (2007)
9. Zhu, T.: Suspicious Financial Transaction Detection Based on Empirical Mode Decomposition Method. In: IEEE Asia-Pacific Conference on Services Computing, 2006. APSCC 2006, pp. 300–304 (2006)
10. Zhu, T.: An Outlier Detection Model Based on Cross Datasets Comparison for Financial Surveillance. In: IEEE Asia-Pacific Conference on Services Computing, 2006. APSCC 2006, pp. 601–604 (2006)
11. Sankoff, D., Blanchette, M.: Multiple Genome Rearrangement and Breakpoint Phylogeny. Journal of Computational Biology 5(3), 555–570 (1998)
12. Gombay, E.: Sequential change-point detection with likelihood ratios. Statistics and Probability Letters 49(2), 195–204 (2000)
13. Gavrilov, M., et al.: Mining the stock market: Which measure is best. In: Proc. of the 6th ACM SIGKDD (2000)
14. Lane, T., Brodley, C.E.: Temporal sequence learning and data reduction for anomaly detection. ACM Transactions on Information and System Security (TISSEC) 2(3), 295–331 (1999)
15. Rafiei, D., Mendelzon, A.: Similarity-based queries for time series data. In: Proceedings of the 1997 ACM SIGMOD international conference on Management of data, pp. 13–25 (1997)
16. Banerjee, A., Ghosh, J.: Clickstream clustering using weighted longest common subsequences. In: Proc. of the Workshop on Web Mining, SIAM Conference on Data Mining, p. 33 (2001)

17. Antunes, C.M., Oliveira, A.L.: Temporal data mining: An overview. In: KDD Workshop on Temporal Data Mining, pp. 1–3 (2001)
18. Agrawal, R., et al.: Fast Similarity Search in the Presence of Noise, Scaling, and Translation in Time-Series Databases. In: Proceedings of the 21th International Conference on Very Large Data Bases, pp. 490–501 (1995)

Data Protection in Memory Using Byte Reordering

Hyun Jun Jang, Dae Won Hwang, and Eul Gyu Im

College of Information and Communications
Hanyang University, Seoul, 133-791
Republic of Korea
{jhj0104,hhboy82,imeg}@hanyang.ac.kr

Abstract. With the explosive uses of Internet, personal information can be exposed to attackers through various attacks. API hooking based mechanisms or virtual machine based mechanisms are proposed to protect personal data stored in the memory. But these mechanisms require another application program and attackers can get around by developing their own device drivers.

We propose a new memory reorganizing mechanism to protect memory. In our proposed mechanism, a user can define his or her own data types and can store data using his or her own data types so that data formats stored in actual memory have different forms.

Keywords: Memory protection, Data Protection, Information Security.

1 Introduction

As Internet popularity grows explosively, people use various Internet services. As a consequence, information sent through the Internet is vulnerable to be exposed to attackers. Recently active attacks such as memory scanning attacks targeted for data of Internet services occur increasingly. In addition, most of DRM(digital rights management) mechanisms do not provide data protection methods against memory scanning attacks.

As for Internet banking services, keyboard hacking protection solutions provide data protection between input devices and the Internet banking processes. But, Internet banking services are still vulnerable to data(e.g. account numbers and PIN) exposure through memory scanning attacks. Memory scanning attacks can also acquire contents protected by various DRM mechanisms.

In this paper, we propose a mechanism to protect against memory scanning attacks. Our proposed mechanism defines its own data types to substitute normal data types, so data formats stored in memory are different from actual user inputs.

The rest of paper is organized as follows: Section 2 addresses related work, and Section 3 explains the Windows memory structure. An attack example is shown in Section 4. Our proposed model is explained in Section 5, followed by conclusions and future directions in Section 6.

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 62–68, 2008.

2 Related Work

2.1 Memory Protection

Memory Protection mechanisms can be divided into two categories: mechanisms to detect or to prevent buffer overflow attacks [1,5], and memory temper resistance mechanisms to protect illegal changes of memory data. In a large sense, our proposed mechanism belongs to the second category, so we will address previous work of the second category.

Previous memory temper resistance mechanisms usually use memory related API(application programming interface) hooking to protect memory data [2,3,4]. API(Application Programming Interface) is a interface between the operating system and application programs, and it is usually used for file controls, windows controls, display processing, and so on.

API hooking mechanisms change parameters passed to memory related APIs to protect data from attackers. To implement API hooking mechanisms, kernel-level privileges are required. One way of acquiring kernel-level privileges is to implement a program as a device driver using WDM (Windows Driver Model).

API hooking based mechanisms tried to find illegal memory accesses by placing special marks in memory. A problem with this mechanism is that a separate protection program is required and attackers can get around the protection program through their own device driver programs. Since device drivers run in a kernel level, it is quite complicated to detect illegal accesses from various device drivers.

2.2 DRM

DRM(Digital Rights Management) provides mechanisms to protect illegal uses of digital contents, and DRM includes both technologies and services to protect digital rights. DRM includes DOI (Digital Object Identifier) which is a content identifier, INDECS which is used to store e-commerce related data, and digital watermarking which is used to prevent and track illegal copies of digital contents.

Even though digital contents are protected by various protection mechanisms, memory scanning attacks can acquire contents illegally by access physical memory directly.

3 The Windows Memory Structure

Figure 1 shows the structure of Windows virtual memory [6,7]. Windows virtual memory has several areas: the kernel code and data area, the stack area, the heap area, the global data area, the program code area, and the PCB(Process Control Block) area. Among these areas, the stack area, the heap area, and the global data area are needed to be protected because these areas have user input data.

Application programs usually use stack or heap areas to store data, so that attackers scan these areas to get or to change other users' input data.

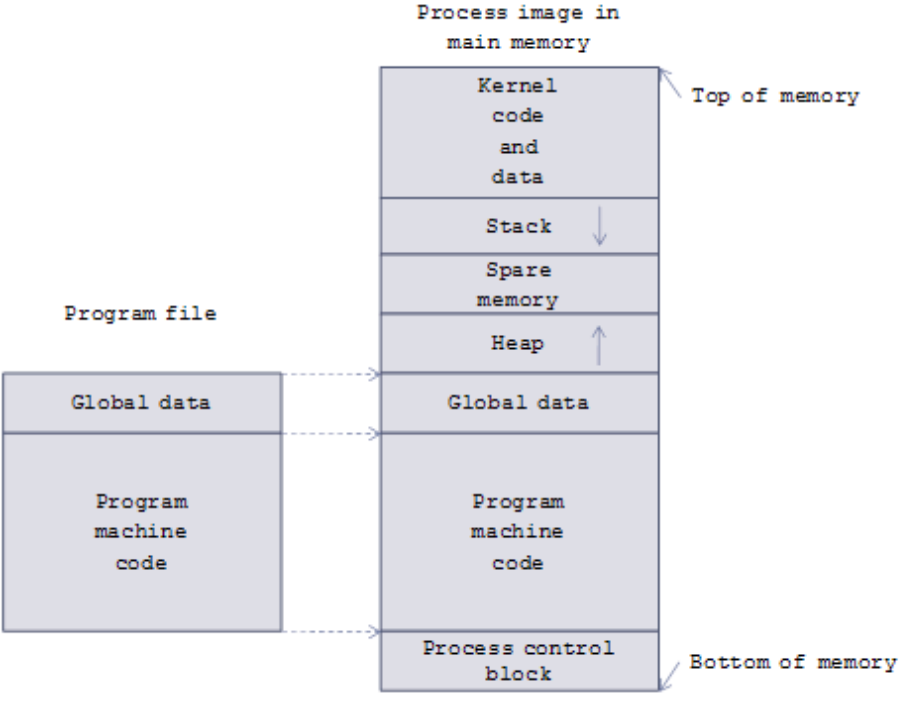


Fig. 1. Windows Memory Structure

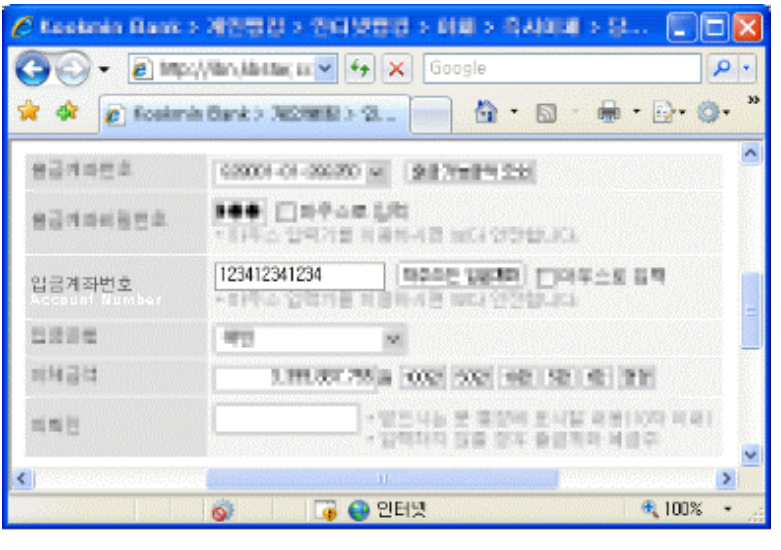


Fig. 2. Screen Shot of the Internet Banking Service

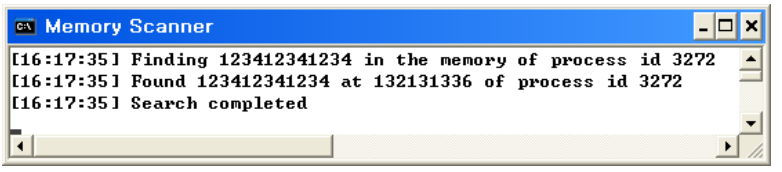


Fig. 3. Results of a Memory Scanner

4 Memory Scanning Attacks

Figure 2 shows a screen shot that a user tries to transfer money through an Internet banking service. When a user enters an account number as shown in Figure 2, attackers can get account information using hacking tools, such as a memory scanner (see Figure 3). If an attacker examines memory used by web browsers or Internet services, he or she may get information passed to web browsers or the Internet services by users. In addition, an attacker even changes information passed by user, to direct money transferred to his own account.

As for DRM applied digital contents, such as audio files, the digital contents are played after the original contents are restored. Therefore, an attacker can get the original contents, i.e. DRM-free digital contents from memory, by scanning memory used by a player program, such as Windows Media Player.

5 Our Proposed Memory Reorganization Mechanism

To countermeasure memory scanning attacks, our proposed mechanism is to substitute normal data types to user defined data types, and to recompile the applications. For this purpose, we implemented a C++ library, called MPL(Memory Protector Library) which is our memory protection library for Windows applications. In the current version of MPL, values of the variable type *int*, which is the most widely used variable type, are changed to a different format. Future extension of MPL will be also able to handle other types, such as *float, double, char, long*, and so on.

The following source codes a part of implementation of MPL using operator overloading.

```
#include <stdio.h>
typedef unsigned short WORD, *PWORD;

class new_int {
public:
  new_int() { m_nValue = 0; }
  new_int(int nValue) {
    *((PWORD)&m_nValue) = (WORD)*(PWORD)&nValue;
    *((PWORD)&m_nValue + 1) = (WORD)*((PWORD)&nValue + 1);
    nValue = 0xCCCCCCCC; }
```

```
  operator int() {
    int nValue;
    *((PWORD)&nValue) = (WORD)*(PWORD)&m_nValue;
    *((PWORD)&nValue + 1) = (WORD)*((PWORD)&m_nValue + 1);
    return nValue; }
  int m_nValue; };
```

To reorganize byte ordering in memory, MPL divides 4 byte integer into two parts, higher 16 bits and lower 16 bits, and exchange these two parts. For example, if a stored value is 0x12345678, then the *new_int* class stores it as 0x56781234 in memory.

The newly defined class, *new_int*, can be used as follows:

```
int main() {
  ...
  new_int my_int = 'KEY!';

  printf("%d\n", my_int);
  ...
}
```

Figure 4 and Figure 5 show results of a memory scanner with or without the new class, *new_int*. When 'KEY!' is assigned to an *int* type variable, it is stored as '!YEK' because the Intel architecture stores data in the little endian method.

In Figure 4, the application program uses the normal *int* type, so '!YEK' is found in two memory locations: one is the location of the source code and the other location is the address where actual data is stored.

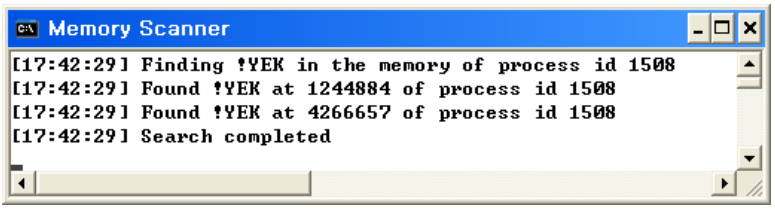


Fig. 4. Results of a Memory Scanner

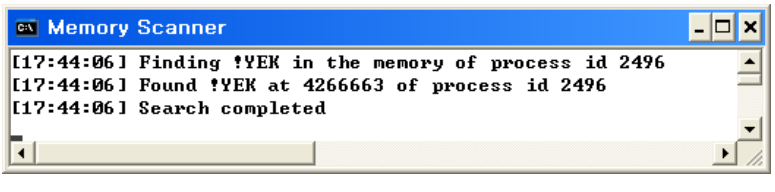


Fig. 5. Results of a Memory Scanner

In contrast, in Figure 5, the *new_int* class was used in the application, and the string '!YEK' was found at only one location (in the program code area) because '!YEK' was stored after byte ordering was changed and the program source codes still have the string. The *new_int* class changes 'KEY!' to 'Y!KE' through byte reordering, and the Intel's little endian method stores it as 'EK!Y' in the actual memory.

In Figure 4, attackers can find the memory location that the value is stored (in this example, 1244884), so that data can be acquired or altered by attackers.

6 Conclusions and Future Directions

With the explosion of Internet, people use various Internet services, including Internet shopping, Internet banking, and so on. While using Internet services, people input sensitive information, such as account numbers or PIN. Attackers may get these kinds of information by scanning memory used by the Internet services.

This paper introduced a new memory protection mechanism by changing data formats stored in the memory. The proposed mechanism can be used to protect data stored in the memory. We implemented the *new_int* class to substitute the normal *int* data type in C++, so that *int* type data can be stored in a different way.

Our proposed mechanism has some limitations:

- Since our mechanism defines new data types, application programs must be recompiled with new data types.
- If an attacker analyzes executable files and finds data formats stored in memory, our proposed mechanism cannot protect data stored in the memory. As a future direction, to solve this problem, a key can be used to determine data formats stored in the memory, and the key can be provided or generated dynamically.

References

1. Uppuluri, P.: Preventing Race Condition Attacks on File-Systems. In: Proceedings of the Symposium on Applied Computing (2005)
2. Witchel, E., Cates, J., Asanovic, K.: Mondrian Memory Protection. ACM SIGARCH Computer Architecture News archive (2002)
3. Clause, J., Doudalis, I., Orso, A., Prvulovic, M.: Effective Memory Protection Using Dynamic Tainting. In: Proceedings of the 22nd IEEE/ACM International Conference on Automated Software Engineering (ASE) (November 2007)
4. Kharbutli, M., Jiang, X., Solihin, Y., Venkataramani, G., Prvulovic, M.: Comprehensively and Efficiently Protecting the Heap. In: Proceedings of the 12th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), pp. 207–218 (October 2006)
5. Bhatkar, S., Sekar, R., DuVarney, D.C.: Efficient techniques for comprehensive protection from memory error exploits. In: Proceedings of the 14th conference on USENIX Security Symposium (2005)

6. The Virtual Memory Manager in Windows NT,
http://msdn2.microsoft.com/en-us/library/ms810616.aspx
7. Virtual Memory Architecture,
http://www.ece.cmu.edu/~ece548/handouts/05vmarch.pdf

Highly Efficient Password-Based Three-Party Key Exchange in Random Oracle Model

Hung-Yu Chien[1] and Tzong-Chen Wu[2]

[1] Department of Information Management, National Chi-Nan University, Taiwan 545, R.O.C.
[2] Department of Information Management, National Taiwan University of Science and Technology, R.O.C.

Abstract. A password-based three-party encrypted key exchange (3PEKE) is a protocol enables any pair of two registered clients to establish session keys via the help of a trusted server such that each client shares only one password with the server. This approach greatly improves the scalability of key agreement protocol in distributed environments, and provides great user convenience. This paper proposes a new password-based 3PEKE scheme with only four message steps, which is the minimum among the published works. The proposed scheme is secure in the random oracle model.

Keywords: authentication, guessing attack, impersonation attack, key agreement, random oracle.

1 Introduction

The two-party encrypted key exchange protocols (2PEKE) like [18, 19] allows two entities to establish session keys without the help of trusted third party. However, this approach owns poor scalability because each entity needs to keep n-1 secrets when there are n entities in the system. On the contrary, a three-party encrypted key exchange (3PEKE) protocol allows any pair of registered clients, where each registered client keeps only one secret with a trusted server, to establish authenticated session keys via the help of the server. This approach greatly improves the scalability and the maintenance cost. Due its convenience for users, low-entropy password is still very popular as the shared secret used for authentication and key exchange in distributed environments. However, these password-based schemes like Kerberos [8] and KryptoKnight [14] are vulnerable to various password guessing attacks- on-line detectable guessing attack, on-line un-detectable guessing attack, and off-line guessing attack [7, 12].

Because the interaction in the three-party case is much more complicated, many existing 3PEKE schemes like [12, 13, 16, 17] were found in-secure or the claimed security is still un-proven [4, 9, 10, 11, 20]. In addition to the security, many previous works like [4, 6, 9-13, 15-17, 20] tried to reduce the number of message steps to improve the communication performance, but the most efficient ones like [11, 13, 20] still require 5 message steps. Therefore, it is desirable to design secure and efficient 3PEKE protocol which is resistant to various attacks, and is optimal in terms of communication

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 69–76, 2008.

performance. In this paper, we shall propose a new password-based 3PEKE protocol which requires only 4 message steps in four rounds, which is the minimum among the published works. This result disputes the previous lower bound set by Gong [20, 21].

Related works

To deter the password guessing attacks, Steiner et al. proposed a password-based 3PEKE protocol (called STW-3PEKE for short) [16]; However, Ding and Horster [7], Lin et al. [11], and Sun et al. [17] respectively showed the scheme was vulnerable to various password guessing attacks. Since then, there are many published works aimed to improve either the security or the communication performance [4, 6, 9, 10, 12, 13, 15]. In evaluating the communication performance of the 3PEKE protocols, the number of *message round*s and the number of *message step*s are important criteria.

Message step: it denotes one transmission step by which one entity sends the data to another entity (or several entities in broadcast environments) in a single step.

Message round: it denotes the integration of one or more message steps of which there is no data dependency between these steps and they can be executed in parallel to save communication time. Thus an entity can simultaneously send different messages to different entities in one round, and so can multiple entities sends messages in one round.

The previous password-based 3PEKE protocols can be classified into two types-those with server's public key [11, 17] and those without server's public key [4, 6, 10, 12, 13, 15]. In Lin-Sun-Hwang's protocol (hereafter referred to as LSH-3PEKE protocol) [11] and Sun-Chen-Hwang's protocol (hereafter referred to as SCH-3PEKE) [17], the server is equipped with a public key and the clients use the public key to securely transmit their passwords and keying material. This approach is more efficient in terms of the number of message steps. The other approach like Lin-Sun-Steiner-Hwang's protocol (called LSSH-3PEKE) [12], Chang-Chang's scheme (called CC-3PEKE) [4], Lee-Hwang-Lin's scheme (called LHL-3PEKE) [10] and Lu-Cao's schemes (called LC-3PEKE) [13] do not use the server's public key.

Even though many works have tried to improve either the security or the communication performance of 3PEKE protocols, only Gong [21] had studied the communication lower bounds of 3PEKE protocols. He classified 3PEKE protocols into several categories according to the following properties: time-stamp-based (TB) vs. nonce-based (NB), authentication-only (AO) vs. authentication with handshake (AH), and server choosing the session key (SO) vs. one client choosing the key (CO) vs. both clients choosing the key (CC). The authentication-only (AO) means that the clients has properly received the session key but they are not ensured whether the communicating party has the same key, and the authentication with handshake (AH) means that the clients are also ensured they share the same key. To be more descriptive and compatible with the previous works, we would like to use the terms key distribution (KD), key confirmation (KC) to replace the terms authentication-only (AO), authentication with handshake (AH) respectively in the rest of this paper. Gong has discussed the lower bounds under different combinations of the properties. This paper focuses on the NB+KC+CC case because it is the main stream of recent works [4-7, 9-13, 15, 16, 17] and this setting is more suitable to the current communication environments, where

strict time synchronization cannot be ensured or costly, and the clients are capable of choosing good keying material for establishing secure session keys. The *previous* lower bounds set by Gong [21] for the NB+KD+CC case, the NB+KC+CC are 5 message steps, 6 message steps respectively. However, we think that they should be 3 message and 4 messages respectively. We will prove this observation in the full version.

In this paper, we shall propose a highly efficient password-based 3PEKE protocol which requires only four message steps, which is the minimum among the published works. Our contributions are two-fold: the proposed protocol requires the least number of interactions among the published works, and we prove its security in a formal model. The rest of this paper is organized as follows. Section 2 proposes a new password-based 3PEKE scheme to enhance both the security and the communication performance. The scheme requires only four message rounds in four steps, which is the minimum among those published works. Section 3 proves its security in the random oracle model, and evaluates the performance. Finally, we draw a brief conclusion in section 4.

2 A New Password-Based 3PEKE with 4 Message Steps

The proposed scheme is based the computational Diffie-Hellman problem (CDHP) and the IND-CCA2 secure public key encryption scheme.

2.1 Preliminaries

The notations are introduced as follows.

(G, g, p): a finite cyclic group G generated by an element g of order p. In the following, we omit the $\bmod p$ operation when the semantic is clear.

C_A, C_B: the two entities want to establish session keys via the help of S.

S, Y_S : the trusted server, the public key of the server.

P_A, P_B **:** the passwords of C_A and C_B respectively.

$E_{Y_S}[.]$: the probabilistic public key encryption using the key Y_S .

$h()$, $h^t()$: two secure one-way hash functions. They will be modeled as a random oracle in the security proof.

Definition 1. The Computational Diffie-Hellman problem (CDHP) over Z_p^* is defined as follows: Given $g^x \bmod p$, $g^y \bmod p$ and g , where x and y are random numbers and g is a generator for the group, compute $g^{xy} \bmod p$ is believed to be a hard problem.

Definition 2. Indistinguishable against adaptive chosen cipher text attack (IND-CCA2) for Public Key Encryption Scheme [1, 5]. Here, it means that no probabilistic polynomial time attacker can tell which plaintext is encrypted, given a ciphertext corresponding to one of two possible plaintexts against adaptive chosen

cipher text attack. Let $ENC_{pub} = (K_{gen}, E_{K_{pub}}(), D_{K_{priv}}())$ be a public key encryption scheme, where K_{gen} is the key generation algorithm which outputs the public key K_{pub} and the private key K_{priv}, and $E_{K_{pub}}()$, $D_{K_{priv}}()$ respectively denotes the public key encryption algorithm and the decryption algorithm. Let A_{ENC} be an adversary for IND-CCA2. A_{ENC} be composed of a find-stage algorithm A_1 and a guess-stage A_2. Let $l \in N$ be a security parameter and *st* be state information. A specification for the experimental algorithm is as follows.

Experiment. $Cca2Exp^{ind-cca2}_{ENC_{pub}, A_{ENC}}(l)$

$K_{pub}, K_{priv} \leftarrow_R K_{gen}(l)$, $(m_0, m_1, st) \leftarrow A_1^{K_{pub}, E_{K_{pub}}(), D_{K_{priv}}()}(l, find)$,

$b \leftarrow_R \{0,1\}$, $c \leftarrow E_{K_{pub}}(m_b)$

$b' \leftarrow A_2^{K_{pub}, E_{K_{pub}}(\cdot), D_{K_{priv}}()}(l, guess, m_0, m_1, c, st)$

if [b'= b] and [A_2 never queries the oracle $D_{K_{priv}}()$ *with c]*

then return 1 else return 0

Now let $Succ^{ind-cca2}_{ENC_{pub}, A_{ENC}}(l) \overset{def}{=} 2\Pr[\ Cca2Exp^{ind-cca2}_{ENC_{pub}, A_{ENC}}(l) = 1] - 1$, and the advantage function of IND-CCA2 for the public key encryption ENC_{pub} is defined as $Adv^{ind-cca2}_{ENC_{pub}}(l, t, q_e, q_d) \overset{def}{=} \max_{A_{ENC}}\{Succ^{ind-cca2}_{ENC_{pub}, A_{ENC}}(l)\}$, where the maximum is taken over all A_{ENC} with execution time t, q_e numbers of queries to the encryption oracle and q_d numbers of queries to the decryption oracle, made by A_{ENC} during the attack. ENC_{pub} is IND-CCA2 secure if $Adv^{ind-cca2}_{ENC_{pub}}(l, t, q_e, q_d)$ is a negligible function in l for any adversary A_{ENC} whose time complexity and number of queries are polynomial in l.

2.2 The Proposed Protocol

Now we are ready to present our protocol as follows, where "*sid*" denotes the session identifier used to uniquely identify a session from other sessions.

1. $C_A \rightarrow C_B : sid,\ C_A, C_B, N_A = g^x,\ E_{Y_S}[C_A \| C_B \| P_A \oplus N_A \| R_A]$

C_A randomly chooses two integers x and R_A, computes $N_A = g^x$, and then uses the server public key to encrypt the data $C_A \| C_B \| P_A \oplus N_A \| R_A$, where the public key-based encryption should be a secure probabilistic encryption that satisfies the in-distinguishability against adaptive chosen ciphertext attack (IND-CCA2) [1, 5]. It finally sends the data in Step 1 to C_B.

2. $C_B \rightarrow S : sid,\ C_A, C_B, N_A = g^x,\ E_{Y_S}[C_A \| C_B \| P_A \oplus N_A \| R_A],\ N_B = g^y,$

$E_{Y_S}[C_A \| C_B \| P_B \oplus N_B \| R_B \| h(sk"_{A,B})]$

C_B randomly chooses two integers y and R_B, computes $N_B = g^y$, and then encrypts the data $P_B \oplus N_B \| R_B \| h(sk"_{A,B})$, where $sk_{A,B} = h(sid \| C_A \| C_B \| | N_A{}^y) = h(sid \| C_A \| C_B \| g^{x \cdot y})$ and $sk"_{A,B} = h^t(sid \| C_A \| C_B \| g^{x \cdot y})$. $sk_{A,B}$ is the session key. Finally, it sends the data in message 2 to S.

3. $S \rightarrow C_A :\ N_B, M_1 = h(sid \| C_A \| C_B \| N_A \| N_B \| h(sk"_{A,B}) \| R_A),\ M_2 = h(sid \| C_A \| C_B \| N_A \| N_B \| R_B)$

Upon receiving the request, S first decrypts the data $E_{Y_S}[C_A \| C_B \| P_A \oplus N_A \| R_A]$ to get the plaintext and checks whether $P_A \oplus N_A$ exits in the plaintext. If so, S accepts this as a valid message from C_A and extracts R_A, the second part of the plaintext. Likewise, S decrypts $E_{Y_S}[C_A \| C_B \| P_B \oplus N_B \| R_B \| h(sk"_{A,B})]$ and checks whether the data $P_B \oplus N_B$ exists in the plaintext, and extracts the values R_B and $h(sk"_{A,B})$. If both verifications succeed, it accepts the request and sends the data in Step 3 to C_A; otherwise, it rejects the request.

4. $C_A \rightarrow C_B :\ M_2 = h(sid \| C_A \| C_B \| N_A \| N_B \| R_B),\ M_3 = h(sid \| C_A \| C_B \| N_A \| N_B \| h(sk"_{A,B}))$

After receiving the response from S, C_A uses the local values to verify M_1. If the verification succeeds, it accepts the connection, accepts the key $sk_{A,B}$, and computes M_3. Finally, it sends the data in Step 4 to C_B; otherwise, C_A rejects the connection.

Upon receiving the data in Step 4, C_B uses its local values to verify M_2 and M_3. M_2 is used to verify the authenticity of the message from C_A via the help of the server, and M_3 is used to verify the session key known by C_A. If all the verifications succeed, C_B accepts this session key.

3 Security Proof and Performance Evaluation

3.1 Security Proof

The security requirements of a secure password-based 3PEKE should consider the possible password guessing attacks in addition to the security requirements of general key agreement protocols. We prove its security in the following theorem.

Theorem 1. The proposed password-based 3PEKE protocol is secure, in the random oracle model, if the Computational Diffie-Hellman problem is hard and the underlying public key encryption scheme is IND-CCA2 secure.

Proof: due to page limitation, we skip the detailed proof in this version.

For the NB+KC+CC case, we think the lower bound is four message steps. However, due to page limitation, we shall prove this point in the full version.

Table 1. Performance comparison of 3PEKE protocols

	LSSH [15]			LHL [12]			[2]SCH2 [19, 21]			LSH [14]			Our		
Rounds/ Steps	7/7			6/6 [4/10] [1]			5/5[4/6][2]			5/5			4/4		
Server's public key	No			No			Yes			Yes			Yes		
weakness	unproved			unproved			Impersonation attacks			unproved			Provable security		
	C_A	C_B	S	C_A	C_B	S	C_A	C_B	S	C_A	C_B	S	C_A	C_B	S
# of Asym. Encrypt.	0	0	0	0	0	0	1	1	2	1	1	2	1	1	2
# of Sym. Encrypt.	1	1	2	1	1	2	2	2	0	2	2	2	0	0	0
# of random number	1	1	2	1	1	2	1	1	4	3	2	0	2	2	0
Hash/pseu-do random	5	5	4	6	6	5	0	0	0	1	1	0	4	4	2
Expon.	3	3	4	3	3	4	2	2	4	2	2	0	2	2	0
Modular multip.	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0

1. Lee et al. has another round efficient version which requires 4 runs in 10 steps.

2. Sun-Chen-Hwang proposed 3 versions of their scheme. SCH-3PEKE2 requires 5 steps in 5 runs; while SCH-3PEKE3 is a run-efficient version which requires 4 runs in 6 steps.

3.2 Performance Evaluation

This section evaluates the performance of various 3PEKE protocols in terms of computations, the number of message rounds and the number of message steps. The comparisons among 3PEKE protocols are summarized in Table 1 (in the end of this paper). In the table, we can see that our proposed scheme owns the minimum number of message rounds and the minimum number of steps. In the table, some schemes favor the computation of clients while others favors that of server. Among those 3PEKE schemes with server public key, [11, 15, 17], our scheme also requires the least computational overhead.

4 Conclusions

In this paper, we have proposed a new password-based 3PEKE scheme. The scheme not only enhances the security but also requires only four message steps. We have proved its security in the random model. So, an interesting future work is to extend our result to the setting without the server's public key.

Acknowledgments. This research is partially supported by the National Science Council, Taiwan, R.O.C., under contract no NSC 95-2221-E-260 -050 -MY2.

References

1. Bellare, M., Desai, A., Pointcheval, D., Rogaway, P.: Relations among Notations of Security for Public Key Encryption Schemes. In: Krawczyk, H. (ed.) CRYPTO 1998. LNCS, vol. 1462, pp. 26–45. Springer, Heidelberg (1998)
2. Bellare, M., Pointcheval, D., Rogaway, P.: Authenticated Key Exchange Secure against Dictionary Attacks. In: Preneel, B. (ed.) EUROCRYPT 2000. LNCS, vol. 1807, pp. 139–155. Springer, Heidelberg (2000)
3. Bellare, M., Rogaway, P.: Provably Secure Session Key Distribution: The Three Party Case. In: 27th ACM Symp. on the Theory of Comput., pp. 57–66. ACM Press, New York (1995)
4. Chang, C.C., Chang, Y.F.: A Novel Three-Party Encrypted Key Exchange Protocol. Computer Standards and Interfaces 26(5), 471–476 (2004)
5. Chien, H.Y.: Selectively Convertible Authenticated Encryption in The Random Oracle Model. The Computer Journal (January 17, 2008) (2008) doi:10.1093/comjnl/bxm090
6. Chung, H.R., Ku, W.C.: Three Weaknesses in a Simple Three-Party Key Exchange Protocol. Information Sciences 178(1), 220–229 (2008)
7. Ding, Y., Horster, P.: Undetectable On-Line Password Guessing Attacks. ACM Operating Systems Review 29(4), 77–86 (1995)
8. Kohl, J., Neuman, C.: The Kerberos Network Authentication Service (V5). Internet Request for Comments 1510 (1993)
9. Ku, W.C., Chiang, M.H., Chang, S.T.: Weaknesses of Yoon-Ryu-Yoo's Hash-Based Password Authentication Scheme. ACM Operating Systems Review 39(1), 85–89 (2005)
10. Lee, T.F., Hwang, T., Lin, C.L.: Enhanced Three-Party Encrypted Key Exchange without Server Public Keys. Computers and Security 23(7), 571–577 (2004)

11. Lin, C.L., Sun, H.M., Hwang, T.: Three Party-Encrypted Key Exchange: Attacks and a Solution. ACM Operating System Review 34(4), 12–20 (2000)
12. Lin, C.L., Sun, H.M., Steiner, M., Hwang, T.: Three-Party Encrypted Key Exchange without Server Public-Keys. IEEE Commun. Lett. 5(12), 497–499 (2001)
13. Lu, R., Cao, Z.: Simple Three-Party Key Exchange Protocol. Computers Security 26(1), 94–97 (2007)
14. Molva, R., Tsudik, G., Van Herreweghen, E., Zatti, S.: KryptoKnight Authentication and Key Distribution System. In: Deswarte, Y., Quisquater, J.-J., Eizenberg, G. (eds.) ESORICS 1992. LNCS, vol. 648, pp. 1–16. Springer, Heidelberg (1992)
15. Nam, J., Kim, S., Won, D.: Attack on the Sun-Chen-Hwang's Three-Party Key Agreement Protocols Using Passwords. IEICE Trans. on Fund. of Electronics, Communications and Computer Sciences E89-A(1), 209–212 (2006)
16. Steiner, M., Tsudik, G., Wainder, M.: Refinement and Extension of Encrypted Key Exchange. ACM Operation Systems Review 29(3), 22–30 (1995)
17. Sun, H.M., Chen, B.C., Hwang, T.: Secure Key Agreement Protocols for Three-Party against Guessing Attacks. The Journal of Systems and Software 75, 63–68 (2005)
18. Chien, H.Y., Wang, R.C., Yang, C.C.: Note on Robust and Simple Authentication Protocol. The Computer Journal 48(1), 27–29 (2005)
19. IEEE P1363.2: Password-Based Public-Key Cryptography, http://grouper.ieee.org/ groups/ 1363/passwdPK/index.html
20. Gong, L.: Optimal Authentication Protocols Resistant to Password Guessing Attacks. In: The 8th IEEE Workshop on Computer Security Foundations, p. 24 (1995)
21. Gong, L.: Lower Bounds on Messages and Rounds for Network Authentication Protocols. In: The 1st ACM Conference on Computer and Communications Security, pp. 26–37 (1993)

A Pairwise Key Pre-distribution Scheme for Wireless Sensor Network

Hui-Feng Huang

Department of Information Management,
National Taichung Institute of Technology, 404 Taichung, Taiwan
phoenix@ntit.edu.tw

Abstract. Wireless sensor networks have been widely used in a variety of domains which include military sensing and tracking, patient monitoring and health care, etc. When sensor networks are deployed in a hostile environment, security becomes extremely important, as they are vulnerable targets to different types of malicious attacks. Due to the wireless property of sensor devices, the sensor networks may easily be compromised by attackers who modify messages or provide misleading information to other sensor nodes. Therefore, the security of the sensor networks is very important. This paper will present a new efficient key pre-distribution scheme for secure wireless sensor networks. It provides an approach that any pair of sensor nodes can find a common pairwise secret key between them with simple calculation. Compared with previously proposed key pre-distribution schemes, the proposed method could reduce large amounts of computations and communications in both the key pre-distribution step and finding a common secret key for any pair of nodes to achieve secure connectivity.

Keywords: sensor network, pairwise, key pre-distribution.

1 Introduction

Recent advancement in wireless communication and electronic technologies has enabled the development of low-cost wireless sensor networks. Sensor networks are usually comprised of one or more base stations and a large number of sensor nodes. These tiny sensor nodes consist of sending, data processing, and communication components [1]. They use their processing abilities to locally carry out simple computations and transmit only the required and partially processed data. The position of sensor nodes need not be engineered or predetermined. This allows random deployment in inaccessible terrain or disaster relief operations. Therefore, the sensor networks are being deployed for a wide variety of applications, including military sensing and tracking, patient monitoring, environmental monitoring, airport and home security [1,4,8,10].

Due to the wireless property of sensor devices, the sensor networks may easily be compromised by attackers who modify messages or provide misleading information

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 77–82, 2008.

to other sensor nodes. To prevent information and communication systems from illegal delivery and modification, message authentication and identification needs to be examined through certificated mechanisms. Most previous schemes proposed for the security of distributed sensor networks have used asymmetric cryptography such as Diffie-Hellman [5] key agreement or the RSA cryptography system [9]. However, these traditional security techniques used in traditional networks are not suitable for sensor networks due to computation capability, energy resources of sensor nodes, dynamic networks, and bandwidth, etc [1,8]. Therefore, new ideas are needed. Today, the practical option for the distribution of keys to the sensor nodes of wireless sensor networks rely on key pre-distribution [2,3,4,6,7,8]. That is, keys are pre-installed in the sensor nodes and the nodes having a common key are provided with a secure connection between them. One solution is all the nodes carry a master secret key. Any pair of nodes can use this master key to achieve key agreement and obtain a new pairwise key [7]. This procedure does not exhibit network resiliency: if the master secret key is compromised by one node, the security of the entire sensor network will be compromised. Some existing research [2] suggests storing the master key in tamper-resistant hardware to reduce the risk. However, it would increase the cost and energy consumption of each node.

Recently, Eschenauer and Gligor [6] proposed a random key pre-distribution scheme. In their scheme, each node receives a subset of random keys from a pool of keys before deployment. It provides an approach that any two nodes have one common pairwise key within their respective subsets for their secure communication. Using this concept, Choi and Youn [4] proposed a key pre-distribution scheme which provides that any pair of sensor nodes can find a common shared key between them with simple calculation. However, it has a shortcoming because the time overhead is high for performing the key pre-distribution step. To improve Choi and Youn's scheme, in 2005, Youn et al. [8] proposed a new scheme which could significantly reduce the time overhead for performing the key pre-distribution step while preserving the same property. When any pair of nodes wants to derive a common secret key between them, the above mentioned schemes [4,8] have to exchange some information for computing a common key; however, the bandwidth consumption is quite demanding and likely to bottleneck in many applications when the number of nodes is large, as required in wireless sensor networks.

This article will present a new pairwise key pre-distribution scheme suitable for power and resource constrained sensor nodes. The proposed scheme could reduce large amounts of computations and communications for both the key pre-distribution step and finding a common secret key for any pair of nodes wanting to communicate. Especially, it does not transmit any information for any pair of nodes to derive a common secret key between them. Compared with most previously proposed schemes [2,3,4,6,7,8] for secure wireless sensor networks, the proposed method could significantly improve the performance and energy efficiency of the sensor nodes.

The rest of this paper is organized as follows. In the next section, we will present a new pairwise key pre-distribution scheme for secure sensor networks. The security analyses and the performances of the proposed scheme are discussed in Section 3. And some conclusions will be made in the last section.

2 The Proposed Scheme

In this section, we will develop a novel key pre-distribution scheme for secure sensor networks. It guarantees that any two sensor nodes can find a common shared key between themselves with simple calculation. This shared key is pairwise. A shared key refers to the relationship between the node and one of its direct neighbors. The neighborhood nodes are predetermined relationships in the sensor network. We now explain the basic idea as follows.

Suppose there are a number of k neighborhood nodes $\{N_1, N_2, \cdots, N_k\}$, the system firstly generates a number of k secret polynomials $\{f_1(x), f_2(x), \cdots, f_k(x)\}$ of degree $(k-1)$ such that $f_i(j) = f_j(i)$ and assigns the secret polynomial $f_i(x)$ to the node N_i, for $i = 1, 2, \cdots, k$. Then, the common key of nodes N_i and N_j is $f_i(j) = f_j(i)$. In other words, any pair of nodes N_r and N_s could compute $f_r(s)$ and $f_s(r)$ by using their secret polynomial $f_r(x)$ and $f_s(x)$, respectively. Therefore, because $f_r(s) = f_s(r)$, it is their pairwise shared key. For the efficiency, the system could perform the following steps to generate a number of k secret polynomials $\{f_1(x), f_2(x), \cdots, f_k(x)\}$ of degree $(k-1)$ for the key pre-distribution.

1. First, the system randomly generates a pool of secret keys $f_i(j)$ such that $f_i(j) = f_j(i)$ for $i = 1, 2, \cdots, k$ and $j = 1, 2, \cdots, k$.
2. According to the Lagrange interpolating polynomial [11], the system could construct the secret polynomial $f_i(x)$ of degree $(k-1)$ of the form $f_i(x) = a_{i,0} + a_{i,1}x + a_{i,2}x^2 + \cdots + a_{i,k-1}x^{k-1}$ for $i = 1, 2, \cdots, k$, by using these k distinct secret points. By applying Lagrange's Formula, the system could obtain $f_i(x) = \sum_{j=1}^{k} f_i(j)(\Pi_{r=1, r\neq j}^{k})\frac{(x-r)}{(j-r)} = a_{i,0} + a_{i,1}x + a_{i,2}x^2 + \cdots + a_{i,k-1}x^{k-1}$.
3. Finally, the system distributes the secret polynomial $f_i(x)$ to each node N_i.

By means of the key pre-distribution $f_i(x)$, a common shared key is guaranteed to be found between two nodes wanting to communicate and mutual authentication is supported. Using a common key, the authenticated channel can be achieved with authentication techniques and we omit it in the description. Here, we present an example of the proposed scheme as follows.

Example 1: Assume there are three neighborhood sensor nodes $\{N_1, N_2, N_3\}$.
Step 1: Generate a pool of keys as is listed in the following:

$f_i(j)$	N_1	N_2	N_3
N_1	1	2	9
N_2	2	3	8
N_3	9	8	9

The above table shows that $f_1(1) = 1$, $f_1(2) = 2$, and $f_1(3) = 9$. Similarly, we have $f_2(1) = 2$, $f_2(2) = 3$, $f_2(3) = 8$, $f_3(1) = 9$, $f_3(2) = 8$, and $f_3(3) = 9$. It is clear that $f_i(j) = f_j(i)$.

Step 2: According to the above keys, we could derive the secret polynomial $f_1(x) = 3x^2 - 8x + 6$ by using these three points $(1, 1), (2, 2)$, and $(3, 9)$. Similarly, we obtain $f_2(x) = 2x^2 - 5x + 5$ and $f_3(x) = x^2 - 4x + 12$. Finally, the system assigns the secret polynomial $f_i(x)$ to each sensor node N_i, for $i = 1, 2, 3$. Now, assume that nodes N_2 and N_3 want to communicate or authenticate with each other, then N_2 and N_3 could compute $f_2(3) = 8$ and $f_3(2) = 8$ by using their secret key pre-distribution $f_2(x)$ and $f_3(x)$, respectively. Hence, any pair of nodes N_i and N_j, can derive the shared key $f_i(j) = f_j(i)$ for their securing communications.

3 Discussions

In this section, we are going to explore the securities and performances of our scheme.

3.1 Security Analysis

Assume there are a number of k neighborhood nodes $\{N_1, N_2, \cdots, N_k\}$, then the system generates a number of k secret polynomials $\{f_1(x), f_2(x), \cdots, f_k(x)\}$ of degree $(k-1)$ such that $f_i(j) = f_j(i)$; and distribute the secret key distribution $f_i(x)$ to the node N_i, for $i = 1, 2, \cdots, k$. The security of the presented scheme is based on the secret polynomial $f_i(x)$ for $i = 1, 2, \cdots, k$. By applying Lagrange interpolating polynomial, the polynomial f of degree $(k-1)$ requires at least k distinct points, namely $(x_i, f(x_i))$, to reconstruct the polynomial f. In other words, $k-1$ or fewer points cannot reconstruct the polynomial f. In our scheme, the system uses k distinct secret points $(1, f_i(1)), (2, f_i(2)), \cdots, (k, f_i(k))$ to create the secret polynomial $f_i(x)$ of degree $(k-1)$ for each node N_i, where $f_i(x) = a_{i,0} + a_{i,1}x + a_{i,2}x^2 + \cdots + a_{i,k-1}x^{k-1}$ for $i = 1, 2, \cdots, k$. On the other hand, our scheme provides $f_j(i) = f_i(j)$. In this situation, even if all $(k-1)$ sensors of $\bigcup_{j=1, j\neq i}^{k} N_j$ reveal their secret shared key $f_j(i)$ for $j = 1, 2, \cdots, k$, and $j \neq i$, without knowing $f_i(i)$, it also cannot derive the secret key distribution $f_i(x)$ of node N_i.

In the proposed scheme, a shared common key between two nodes is pairwise. In this case, if the common key $f_i(j)$(or $f_j(i)$) of nodes N_i and N_j is compromised, without knowing other common keys of any two nodes, the security of the entire sensor network will not be significantly affected. It could provide more secure connectivity between these sensor nodes.

3.2 Performance

Suppose there are a number of k neighborhood nodes $\{N_1, N_2, \cdots, N_k\}$, the system firstly generates a number of k secret polynomials $\{f_1(x), f_2(x), \cdots, f_k(x)\}$ of degree $(k-1)$; such that $f_i(j) = f_j(i)$, and assigns the secret key distribution $f_i(x)$

Table 1. Comparisons of computation and transmission for two schemes

Schemes	Park et al.'s scheme	The proposed scheme
Compute the key pre-distribution for each node	k^2T_m	k^2T_m
Compute a common key for each node	kT_m	$(k-1)T_m$
Number of transmissions for each node to compute a common key	k	0

to the node N_i, where $f_i(x) = a_{i,0} + a_{i,1}x + a_{i,2}x^2 + \cdots + a_{i,k-1}x^{k-1}$ for $i = 1, 2, \cdots, k$. Then, it has

$$\begin{aligned} & f_i(x) = a_{i,0} + a_{i,1}x + a_{i,2}x^2 + \cdots + a_{i,k-1}x^{k-1} \\ & = ((\cdots((a_{i,k-1}x + a_{i,k-2})x + a_{i,k-3})x + \cdots)x + a_1)x + a_0. \end{aligned} \tag{1}$$

Therefore, from Equation (1), any pair of nodes N_r and N_s could compute their common key $f_r(s) = f_s(r)$ by using their secret polynomial $f_r(x)$ and $f_s(x)$, respectively. In this situation, it only requires $(k-1)$ multiplication computations and $(k-1)$ addition computations to obtain a common key for each node. In the transmissions, it does not transfer or exchange messages for any two sensor nodes N_r and N_s to obtain a common key $f_r(s) = f_s(r)$. Therefore, the proposed method could significantly reduce the overhead of communication and computation for sensor nodes to achieve secure connectivity.

With regard to efficiency and communications (transmissions), we compare our scheme with Park et al.'s [8] which is more efficient than previously proposed schemes, to the best of our knowledge. For convenience, the notation T_m means the time for one multiplication computation. Note that the times for computing addition and subtraction are ignored, since they are much smaller than T_m. We summarize the comparisons of our scheme with Park et al.'s in Table 1. As shown in Table 1, in the computational complexities of the key pre-distribution step and finding a common key between two nodes, the proposed scheme is as efficient as Park et al.'s. However, Park et al.'s scheme is required to deliver k messages for any sensor node to compute a common key, where k is the number of neighborhood nodes. The bandwidth consumption is quite demanding and that is likely bottleneck in many applications. The proposed scheme does not transfer any message to compute a common key. Therefore, the proposed key pre-distribution scheme presents a significant improvement in performance and energy efficiency of the sensor nodes.

4 Conclusions

This paper proposed a new pairwise key pre-distribution scheme suitable for power and resource constrained sensor nodes. It significantly reduces the overhead by avoiding modular exponentiation and inverse computations. Compared with previously proposed schemes, the proposed method reduces large amounts of computations and communications for both the key pre-distribution step and finding a

common secret key for any pair of nodes to have secure communication between them. It is noteworthy that no information transfer is required to compute a common key between any two nodes to achieve secure connectivity. Therefore, the proposed method is very adoptable for the sensor nodes that are limited in power, computation capability, and bandwidth.

References

1. Akyildiz, I.F., Su, W., Sankarasuramaniam, Y., Cayirci, E.: A Survey on Sensor Networks. IEEE Communications Magazine 40(8), 102–114 (2002)
2. Anderson, R., Kuhn, M.: Tamper Resistance- a cautionary note. In: Proceedings of the Second Usenix Workshop on Electronic Commerce, pp. 1–11 (1996)
3. Chan, H., Perrig, A., Song, D.: Random Key Pre-distribution Schemes for Sensor Networks. In: IEEE Symposium on Security and Privacy, pp. 197–213 (2003)
4. Choi, S.J., Youn, H.Y.: An Efficient Key Pre-distribution Scheme for Secure Distributed Sensor Network. In: Yang, L.T., Amamiya, M., Liu, Z., Guo, M., Rammig, F.J. (eds.) EUC 2005. LNCS, vol. 3824. Springer, Heidelberg (2005)
5. Diffie, W., Hellman, M.E.: New Directions in Cryptography. IEEE Transactions on Information Theory 22(6), 644–654 (1976)
6. Eschenauer, L., Gligor, V.D.: A Key-management Scheme for Distributed Sensor Networks. In: Proceedings of the 9th ACM Conference on Computer and Communication Security, pp. 41–47 (2002)
7. Liu, D., Ning, P.: Establishing Pairwise Keys in Distributed Sensor Networks. In: Proceedings of the 10th ACM Conference on Computer and Communications Security, pp. 52–61 (2003)
8. Park, C.W., Choi, S.J., Youn, H.Y.: A Novel Key Pre-distribution Scheme with LU Matrix for Secure Wireless Sensor Networks. In: Hao, Y., Liu, J., Wang, Y.-P., Cheung, Y.-m., Yin, H., Jiao, L., Ma, J., Jiao, Y.-C. (eds.) CIS 2005. LNCS (LNAI), vol. 3801, pp. 494–499. Springer, Heidelberg (2005)
9. Rivest, R.L., Shamir, A., Adleman, L.M.: A Method for Obtaining Digital Signatures and Public Key Cryptosystems. Communications of the ACM 21, 120–126 (1978)
10. Sara, R.M., Lopez de, A.M., Marco, M.P., Damia, B.: Biosensors for Environmental Monitoring: A Global Perspective. Talanta 65(2), 291–297 (2005)
11. Shamir, A.: How to Share a Secret. Communications ACM 22, 612–613 (1979)

Attacks on SVD-Based Watermarking Schemes

Huo-Chong Ling[1], Raphael C.-W. Phan[2], and Swee-Huay Heng[3]

[1] Centre for Cryptography and Information Security, Faculty of Engineering,
Multimedia University, 63100 Cyberjaya, Malaysia
hcling@mmu.edu.my
[2] Electronic & Electrical Engineering,
Loughborough University, LE11 3TU, United Kingdom
R.Phan@lboro.ac.uk
[3] Centre for Cryptography and Information Security,
Faculty of Information Science & Technology,
Multimedia University, 75450 Melaka, Malaysia
shheng@mmu.edu.my

Abstract. One major application of a watermarking scheme is to protect the copyright of a content owner by embedding the owner's watermark into the content. Recently, two watermarking schemes were proposed by Chang et al. [1, 2], which are based on singular value decomposition (SVD). In this paper, we present attacks on these watermarking schemes and show how the designers' security claims can be invalidated, namely related to robustness and protection of rightful ownership. Our results are the first known attacks on these SVD-based watermarking schemes.

Keywords: watermarking, singular value decomposition, attacks, robustness, proof of ownership.

1 Introduction

Most information, documents and contents these days are stored and processed within a computer in digital form. However, since the duplication of digital content results in perfectly identical copies, the copyright protection issue is a main problem that needs to be addressed. A watermarking scheme [1-9] is one where it is desired to protect the copyright of a content owner by embedding the owner's watermark into the content.

In order to prove the ownership of the watermarked content, the owner takes the case of ownership claim to the authority, and proves ownership by performing the watermark detection process on the claimed content to extract her watermark.

In this paper, we concentrate on a singular value decomposition (SVD)-based watermarking scheme. Singular value decomposition (SVD) is a linear algebra scheme that can be used for many applications, particularly in image compression [10]. Using SVD, it is possible to get an image that is indistinguishable from the original image, but only using 45% of the original storage space [10]. Suppose an N by N image matrix A with rank r <= N. The SVD of A is defined as:

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 83–91, 2008.

$$
\begin{aligned}
A &= USV^T . \\
&= [\, u_1, u_2, \ldots, u_N \,] \begin{bmatrix} s_1 & & & \\ & s_2 & & \\ & & \ddots & \\ & & & s_N \end{bmatrix} [\, v_1, v_2, \ldots, v_N \,]^T . \\
&= \sum_{i=1}^{r} u_i s_i v_i^T .
\end{aligned} \tag{1}
$$

where S is an N by N diagonal matrix containing singular values, and U and V are N by N orthogonal matrices. V^T is the adjoint (transpose and conjugate) of the N by N matrix V. s_i's are singular values satisfying $s_1 >= s_2 >= \ldots >= s_r > s_{r+1} = \ldots = s_N = 0$. Since the singular values are arranged in decreasing order, the last terms will have the least affect on the overall image.

In past years, several SVD-based watermarking schemes have been proposed [1, 2, 4, 5, 7, 9]. The most notable is due to Liu and Tan [9]. They proposed to insert the watermark into the SVD domain of the cover-image, and demonstrated its high robustness against image distortion. However, their scheme is vulnerable to attacks as proven by Rykaczewski [11] and Zhang and Li [12]. Recently, two very similar SVD-based watermarking schemes were proposed [1, 2], the latest being published in ICIC International [2]. In this paper, we present attacks in these two schemes that undermine their security and hence show that their scheme are not as robust as claimed and do not guarantee that the content owner's copyright is protected.

In section 2, we review the schemes proposed by Chang et al. [1, 2]. We then present attacks on these schemes in section 3. Experimental results on the attacks are shown in section 4, and section 5 concludes this paper.

2 Two SVD-Based Watermarking Schemes

In this section, we briefly review the two recently proposed SVD-based watermarking schemes [1, 2]. Both schemes were proposed by Chang et al. and were based on the SVD of image blocks. The difference is that one [2] can restore the watermarked image with high quality while the other [1] is not able to do so. For the ease of discussions, we will refer the scheme proposed in [1] as scheme A and the scheme proposed in [2] as scheme B.

2.1 Chang et al. [1] Scheme (Scheme A)

We first describe the watermarking scheme presented in [1]. The watermark embedding steps are as follows:

E1.1. Denote cover image I as an N by N matrix, and a binary watermark W as a P by P matrix. I is divided into non-overlapping 4 x 4 blocks B_j ($1 <= j <= N/4$ x $N/4$). In order to achieve high robustness, W is copied 3 times to generate P x P x 3 bit streams.

E1.2. Apply a one-way hash function [13] which is based on Rabin's scheme [14] to decide the positions (x_i , y_i) of the embedding blocks B_j ($j = x_i$ ($N/4$) + y_i and $1 <= j <= N/4$ x $N/4$) for each watermark bit, W_i ($1 <= i <= P$ x P x 3).

E1.3. Let $i = 1$.

E1.4. Perform SVD on B_j of (x_i , y_i):

$$B_j = U_j S_j V_j^T . \tag{2}$$

Assume that $S_j = \begin{bmatrix} s_1 & & & \\ & s_2 & & \\ & & s_3 & \\ & & & s_4 \end{bmatrix}$

E1.5. $s'_3 = s_2$. (3)

E1.6. If $s_1 < s_2 + \delta W_i$, then $s'_1 = s_2 + \delta W_i$, else $s'_2 = s_2 + \delta W_i$
where δ is a constant. (4)

The modified $S'_j = \begin{bmatrix} s'_1 & & & \\ & s'_2 & & \\ & & s'_3 & \\ & & & s_4 \end{bmatrix}$

E1.7. Perform SVD on the watermarked block, BW_j as:

$$BW_j = U_j S'_j V_j^T . \tag{5}$$

E1.8. Increment i by 1 and go to step E1.4. until $i = P$ x P x 3, i.e. all the watermark bits have been embedded into the coefficient of S_j .

The watermark extraction steps are as follows:

X1.1. Let $i = 1$.

X1.2. Perform SVD on the watermarked image block, BW_j of (x_i , y_i):

$$BW_j = UW_j SW_j VW_j^T . \tag{6}$$

Assume that $SW_j = \begin{bmatrix} sw_1 & & & \\ & sw_2 & & \\ & & sw_3 & \\ & & & sw_4 \end{bmatrix}$

X1.3. If $sw_2 - sw_3 > \delta/2$, then $WT_i = 1$, else $WT_i = 0$. (**7**)
Note that from equation (4), the difference between sw_2 and sw_3 is either 0 or δ.

X1.4. Increment i by 1 and go to step X1.2. until $i = P$ x P x 3.
X1.5. Let $i = 1$.
X1.6. If $WT_i + WT_{i+PxP} + WT_{i+PxPx2} >= 2$, then $W_i = 1$, else $W_i = 0$. (8)
X1.7. Increment i by 1 and go to step X1.6. until $i = P$ x P.

2.2 Chang et al. [2] Scheme (Scheme B)

We next describe the watermarking scheme presented in [2]. The scheme is actually an extension of the scheme presented in section 2.1. with a minor modification on the embedding and extraction steps. Each binary value of the watermark is embedded into the second non-zero coefficient of the singular values of the image's block, and the extra information required for recovering the image is embedded into the fourth non-zero coefficient of the singular values of the image's blocks. The purpose of the scheme is to remove the hidden watermark so that authorized user can restore the watermarked image with high quality for later usage after the ownership of the image have been verified. The watermark embedding steps are as follows:

E2.1. Perform steps E.1.1. through E1.3.
E2.2. Perform SVD on B_j of (x_i , y_i):

$$B_j = U_j S_j V_j^T . \tag{9}$$

Assume that $S_j = \begin{bmatrix} s_1 & & & \\ & s_2 & & \\ & & s_3 & \\ & & & s_4 \end{bmatrix}$

E2.3. If $s_1 > s_2$, then proceed to the next step, else increment i by 1, then go to step E2.2.
E2.4. If $|s_2 - s_3|$ will not change the order of non-zero coefficients, then $s'_4 = |s_2 - s_3|$, else $s'_4 = - |s_2 - s_3|$.
E2.5. $s'_2 = s_2 + \delta W_i$, where δ is a constant. (10)

The modified $S'_j = \begin{bmatrix} s_1 & & & \\ & s'_2 & & \\ & & s_3 & \\ & & & s'_4 \end{bmatrix}$

E2.6. Perform SVD on the watermarked block, BW_j as:

$$BW_j = U_j S'_j V_j^T . \tag{11}$$

E2.7. Increment i by 1 and go to step E2.2. until $i = P$ x P x 3, i.e. all the watermark bits have been embedded into the coefficient of S_j .

The watermark extraction steps are as follows:

X2.1. Let $i = 1$.
X2.2. Perform SVD on the watermarked image block, BW_j of (x_i , y_i):

$$BW_j = UW_j \, SW_j \, VW_j^T . \tag{12}$$

Assume that $SW_j = \begin{bmatrix} sw_1 & & & \\ & sw_2 & & \\ & & sw_3 & \\ & & & sw_4 \end{bmatrix}$

X2.3. If $sw_1 <= sw_2$, then increment i by 1 and go to step X2.2., else proceed to the next step.
X2.4. If $sw_2 - sw_3 > \delta/2$, then $WT_i = 1$, else $WT_i = 0$. (13)

X2.5. Increment i by 1 and go to step X2.2. until $i = P$ x P x 3.
X2.6. Let $i = 1$.
X2.7. If $WT_i + WT_{i+PxP} + WT_{i+PxPx2} >= 2$, then $W_i = 1$, else $W_i = 0$. (14)
X2.8. Increment i by 1 and go to step X2.7. until $i = P$ x P.

Both of the methods [1, 2] show promising results. Chang et al. [1, 2] claims that both schemes are robust since every binary value of watermark W is embedded in three different blocks of image I, and secure pseudorandom number generator is used to decide the embedding positions of the watermark. They further concluded that their schemes are suitable for protection of rightful ownership of digital images.

However, their methods are vulnerable to attacks which are described in the next section.

3 Attacks

In this section, we show how attacks can be mounted to invalidate the security claims made by the designers of the two watermarking schemes in Section 2.

3.1 Attacks on Robustness

Since scheme B's embedding steps are extended from scheme A, the first attack can be applied to both schemes. This attack invalidates the designers' claim that the schemes are robust. The steps of the attack are as follows:

A1.1. Denote watermarked image, I_w as an N by N matrix. I_w is divided into non-overlapping 4 x 4 blocks B_j ($1 <= j <= N/4$ x $N/4$).
A1.2. Let $j = 1$.
A1.3. Perform SVD on B_j

$$B_j = U_j \, S_j \, V_j^T . \tag{15}$$

Assume that $S_j = \begin{bmatrix} s_1 & & & \\ & s_2 & & \\ & & s_3 & \\ & & & s_4 \end{bmatrix}$

A1.4. Let $s_2 = s_3$. (16)

A1.5. Perform SVD on the modified block, BM_j as:

$$BM_j = U_j S'_j V_j^T . \quad (17)$$

where S'_j is the modified singular values.

A1.6. Increment j by 1 and go to step A1.3. until all the blocks have been processed, i.e. until $j = N/4$ x $N/4$.

During the extraction steps, the watermark will fail to be extracted from the modified watermarked image due to equation (16). This is because based on the watermark extraction steps X1.3. or X2.4., the watermark bit is always 0 since the condition is always false. We do not even need to know the secret key pairs to determine the positions of the embedding blocks. Thus, robustness is invalidated.

3.2 Attacks on Proof of Ownership

The second attack works on both schemes also, and it shows that both schemes cannot be used for proof of ownership claims.

To give the intuition for the attack, consider for an illustrative example a scenario whereby Alice is the owner of an image I, and thus she embeds her watermark W into the image I using scheme A, to obtain the watermarked image I_w. An attacker Bob, successfully obtains I_w and he repeats the process of embedding his own watermark WB into I_w using scheme A, to obtain the watermarked image $I_{wb.}$ This in fact eliminates Alice's watermark W in the watermarked image $I_{w.}$ A dispute arises when Alice claims that she is the real owner of I but she cannot extract her own watermark W from the watermarked image I_{wb}, since it has been tampered by Bob. On the other hand, Bob can extract his own watermark WB from I_{wb}, since he has overwritten Alice's watermark W with his own watermark WB using the same scheme. This attack also works for Scheme B.

Both of the attacks in section 3.1 and section 3.2 work because the embedding space in which the watermark has been embedded, can easily be modified to change the embedded watermark.

Therefore, both of these attacks directly invalidate the designers' claims that their schemes are robust and can be used for protection of rightful ownership of the image.

4 Experimental Results

In this section, experiments are carried out to prove that the attacks in section 3 are feasible. Figure 1 shows four different gray images with the size 200 x 200, and Figure 2 shows the images after the attacks in section 3.1 are launched on the images in Figure 1.

(a) (b) (c) (d)

Fig. 1. (a) Baboon image. (b) Lena image. (c) Pepper image. (d) Plane image.

(a) (b) (c) (d)

Fig. 2. (a) Baboon image. (b) Lena image. (c) Pepper image. (d) Plane image.

In order to evaluate the image quality of the original images in Figure 1 and the modified images in Figure 2, peak signal-to-noise ratio (PSNR) is used. PSNR is defined in equation (18).

$$\text{PSNR} = 10 \text{ x } \log_{10} (255^2 / \text{MSE}) . \tag{18}$$

where MSE is the mean square error between the original and the corresponding modified pixel values. Table 1 shows the PSNR of the modified images. It is observed that the modified images can preserve a good image quality.

Therefore, even though the singular values of the image blocks have been modified, the modified images still resemble the original images as proven in the experiment.

Table 1. PSNR of the modified images

Modified image	***PSNR***
Baboon	31.1063
Lena	34.1663
Pepper	33.5445
Plane	32.2124

For the second attack in section 3.2., we use the Lena image in Figure 1(b), and the binary watermark in Figure 3(a). Figure 4(a) shows the watermarked image of Lena, and Figure 4(b) shows the modified watermarked image after the attacker's watermark in Figure 3(b) is embedded into the watermarked image in Figure 4(a) using Scheme A. Figure 3(c) is the extracted watermark from Figure 4(a), whereas Figure 3(d) shows the extracted watermark from Figure 4(b), after the attack in Section 3.2 is launched.

From the experiment, the original watermark in Figure 3(c) failed to be extracted from Figure 4(b). This means that Chang et al's schemes are not suitable for protection of rightful ownership of images as claimed.

(a) (b) (c) (d)

Fig. 3. Binary watermark with the size 25 x 25. (a) Original watermark. (b) Attacker's watermark. (c) Extracted watermark from Figure 4(a). (d) Extracted watermark from Figure 4(b).

(a) (b)

Fig. 4. (a) Watermarked Lena image. (b) Modified watermarked Lena image after the attack.

5 Conclusions

We have presented attacks on two watermarking schemes which are based on SVD. These attacks work because the designers did not take into consideration the embedding space in which the watermark has been embedded. We show that the embedding space can easily be modified to change the embedded watermark, and we do not even need to know the secret key pairs which are used to determine the positions of the embedding blocks. Our attacks directly invalidate the security claims made by the designers, namely robustness and protection of rightful ownership of digital images. Experimental results have proven that the attacks are feasible and our results are the first known attacks on both schemes.

Acknowledgements. We thank God for His many blessings: "For there is nothing hidden except to be made visible; nothing is secret except to come to light" [Mark 4:22]. Thanks to Jean-Philippe Aumasson for making the connection between the relevance of this verse and information concealment.

References

1. Chang, C.C., Hu, Y.S., Lin, C.C.: A Digital Watermarking Scheme based on Singular Value Decomposition. In: Chen, B., Paterson, M., Zhang, G. (eds.) ESCAPE 2007. LNCS, vol. 4614, pp. 82–93. Springer, Heidelberg (2007)
2. Chang, C.C., Lin, C.C., Hu, Y.S.: An SVD Oriented Watermark Embedding Scheme with High Qualities for the Restored Images. International Journal of Innovative Computing, Information and Control 3(2), 609–620 (2007)
3. Lu, C.S.: Multimedia Security: Steganography and Digital Watermarking Techniques for Protection of Intellectual Property. Idea Group Publishing (2004)
4. Chandra, D.V.S.: Digital Image Watermarking using Singular Value Decomposition. In: 45th IEEE International Midwest Symposium on Circuits and Systems, Tulsa, Oklahoma, vol. 3, pp. 264–267 (2002)
5. Huang, F., Guan, Z.H.: A Hybrid SVD-DCT Watermarking Method based on LPSNR. Pattern Recognition Letters 25, 1769–1775 (2004)
6. Cox, I.J., Miller, M.L., Bloom, J.A.: Digital Watermarking. Morgan Kaufmann, San Francisco (2001)
7. Chung, K.L., Shen, C.H., Chang, L.C.: A Novel SVD- and VQ-based Image Hiding Scheme. Pattern Recognition Letters 22, 1051–1058 (2001)
8. Arnold, M., Wolthusen, S.D., Schmucker, M.: Techniques and Applications of Digital Watermarking and Content Protection. Artech House Publishers (2003)
9. Liu, R., Tan, T.: An SVD-based Watermarking Scheme for Protecting Rightful Ownership. IEEE Transactions on Multimedia 4(1), 121–128 (2002)
10. Andrews, H.C., Patterson, C.L.: Singular Value Decomposition (SVD) Image Coding. IEEE Transactions on Communications 24(4), 425–432 (1976)
11. Rykaczewski, R.: Comments on An SVD-based Watermarking Scheme for Protecting Rightful Ownership. IEEE Transactions on Multimedia 9(2), 421–423 (2007)
12. Zhang, X.P., Li, K.: Comments on An SVD-based Watermarking Scheme for Protecting Rightful Ownership. IEEE Transactions on Multimedia 7(2), 593–594 (2005)
13. Hwang, M.S., Chang, C.C., Hwang, K.F.: A Watermarking Technique based on One-way Hash Functions. IEEE Transactions on Consumer Electronics 45(2), 286–294 (1999)
14. Rabin, M.O.: Digitalized Signatures and Public Key Functions as Intractable as Factorization. Technical Report MIT/LCS/TR212, Laboratory for Computer Science, Massachusetts Institute of Technology, Cambridge, MA (January 1979)

Trigger Based Security Alarming Scheme for Moving Objects on Road Networks

Sajimon Abraham[1] and P. Sojan Lal[2]

[1] Marian College, Department of Computer Science, Kuttikkanam, Idukki, Kerala, India
[2] M.G University, School of Computer Sciences, Kottayam, Kerala, India
sajimabraham@rediffmail.com, sojanlal@gmail.com

Abstract. The advent of modern monitoring applications such as location based services, presents several new challenges when dealing with continuously evolving spatio-temporal information. Spatio-Temporal data analysis plays a central role in many security-related applications including those relevant to transportation infrastructure, border and inland security. This paper reviews a novel binary encoding scheme to store location information and proposes a trigger based security alarming scheme when an object enters into a sensitive area with proper messages to the security people.

Keywords: binary encoding scheme, database trigger, location based services, moving objects on networks, spatio-temporal databases.

1 Introduction

The safety of citizens is the most important task of the government. Security informatics is the study of development and evaluation of advanced information technologies and systems for national security related applications. Spatio-Temporal data analysis plays a central role in many security-related applications including those relevant to transportation infrastructure and border security[16]. Location based services (LBS), one major area in spatio-temporal data management, which involve the ability to find the geographical location of mobile devices and provide services based on this location information [3]. Location Based Services can offer tremendous benefits in the security informatics area. For example tracking the location of a person who needs urgent help, or of a criminal who is wanted or giving proper alarm when vehicle is entering into critical area such as insurgency, terrorist attack, communal violence etc. In the case of emergency calls (eg terrorist attack, robbery, murder etc) it is obvious that if the call responders have the information concerning the location of the people making the call then the response time can be reduced.

The spatio-temporal data analysis frame work and related computational methods discussed in [1] are relevant to many security-related applications in the context of transportation systems. For instance, recent crime analysis in security informatics has discovered that criminals are increasingly utilizing various types of vehicles to assist their criminal activities. FBI has written in their oversight report, “modern transportation and modern technology give terrorist’s abilities unheard of only a few

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 92–101, 2008.

years ago." Mining and analyzing spatial, temporal or spatio-temporal interaction patterns form transportation data can provide many valuable insights to facilitate crime fighting and counter terrorism efforts. For example, spatio-temporal analysis can help to identify emerging hotspots of crime activities in a sensitive area, which in turn provides useful information to help border control agencies efficiently allocate patrol resources. At the same time by identifying such hotspot area, if the system could give sufficient alarm to vehicles moving in that area it would be an important security measure for people moving in that vehicle.

One of the popular services under heavy demand is the location-based service (LBS) that exploits the spatial information of moving objects per temporal changes. In order to support LBS well, in this paper, we investigate how spatio-temporal information of moving objects can be efficiently stored by reducing the data into one dimension and also how this dimension reduction is advantageous in identifying sensitive areas on road networks and how to provide security alarm to objects currently in that area using the trigger concept available in relational data base systems.

In this paper we assume that the database stores the complete history of moving objects through time and must answer queries about any time in the history of objects. We assume that each database record has the format (oid, location, time*),* where oid identifies an object, location is the spatial coordinates(x,y) represented as binary string, and time indicates the time in which the object remained at position *(x, y)*. A typical domain where such a model fits is mobile device tracking, e.g., of GPS, PDA, or wireless phone devices. Unlike the trajectory model [17], our data model does not assume anything about the movement of objects between records. The model reflects a real-world application constraint where assuming an object follows a linear trajectory between data points may lead to incorrect, and unacceptable, assumptions. For example, in security/monitoring applications, a person could be mistakenly assumed to have entered a restricted area, instead of gone around it, because his/her movement was interpolated. Our model can be viewed as a step-wise interpolation instead of a linear interpolation. That is, as long as the object's position is not updated in the database it is assumed to remain stationary in its last observed position.

Our proposal contains a trigger which will fire on updating the moving object data base by an object traveling on road network when it enters/crosses a pre-defined sensitive area. The trigger will provide alarming message to the object traveling on the vehicle as well us it will provide information to highway police. There are many products for vehicle tracking system which either uses GPS receivers to identify user's location and managing data in multidimensional way using GIS packages. We propose a system which maintains a remote moving object data base where information are stored in one-dimensional way in conventional relational data base techniques. As we follow a dimension reduction, the data management becomes easier and the system will provide fast response.

This paper is organized as follows. Section 2 contains the related work on storing location information on road network using coordinates and dimension reduction. Section 3 describes the necessary method available to express location information using hierarchical administrative district as of binary string. In section 4 the proposed method for security alarm is described with necessary logical system diagram and needed algorithms. The paper concluded with a note on future extension in section 5.

2 Related Work

The moving point objects in many cases do not move freely in the 2D plane but rather within spatially embedded networks such as roads. We can then represent movements relative to the network rather than to the 2D space. So instead of describing position of an object by geographic coordinates we can describe its as being at kilometer 220.30 on a particular highway [4],[5]. Research using road network [6] limits the movable boundary of objects to roads so that it represents location information represented as a pair of (the identifier of the nearest road, the distance from the nearest road). In [7], Papadias et al. propose an index suitable for such road coordinates and discuss an algorithm that uses the network distance, instead of the Euclidean distance, to measure the distance between two coordinates. In [6], Gupta et al. propose a scheme where nodes on the road network are systematically converted into binary strings, and thus various queries can be efficiently handled by utilizing simple operations on the binary strings (e.g., Hamming distance). In this scheme, however, the length of binary strings is proportional to the number of nodes in the road network, and the relative locations of roads and moving objects are not easily obtainable from the binary strings of road network nodes. Refs. [11,12,13,14] are some of the representative works that aim at reducing the dimensions of indices to improve query processing performance. These three methods express the locations of moving objects effectively by exploiting the observation that moving objects seldom change their locations much per time. However, they still have to use every information of moving objects, which are at least three-dimensional, for query processing.

The method proposed in [8],[9] transform two-dimensional locations into one dimensional representation which uses Hilbert Curve[10]. Since this method does not consider hierarchical administrative district during the transformation, it does not fit well the real world situations. The binary string representation proposed [18] however, is based on the information of hierarchical administrative district and thus can be easily converted into address formats that are easier for human users to interpret. In this paper the researcher proposes an extension of this method that could be very useful in security informatics for easy identification of sensitive areas and also alarming objects moving in that area. Since each district or road can easily identified by a unique pattern in the new encoding scheme and since the multi-dimensional spatio-temporal data has been converted into one-dimension, query processing will be easier and the PL/SQL and trigger features can be effectively used to implement the scheme.

3 Binary Encoding Method to Store Roads and Relative Locations

In conventional approaches, the location information of moving objects were expressed as a geometric coordinate (x,y) in two-dimensional space. However, instead, [18] propose to express location information using both hierarchical administrative district and road network in one-dimensional space that fits real world better. For instance, if a moving object is in a building at a coordinate of latitude = 125.58 and longitude = -37.34, then it can be expressed as a set of fields according to an administrative district such as city, road-name, road-block (e.g., Seoul, Main road, 165th block). Furthermore, by converting the fields into a binary string that has efficient ways to process queries, we focus the following advantages.

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

The dimension reduction of spatio-temporal data management [18] discusses two algorithms for binary encoding process, one for administrative district, road and location encoding and the second, for converting a position represented as geometric co-ordinate into an equivalent binary string. Since the proposed alarming scheme is based on this encoding method these basic algorithms are briefly discussed below.

Algorithm 1 describes how an administrative district can be represented and stored as a set of binary string. The method is a recursive procedure which will successively divide the entire region into sub-regions and finally map each district into a two-dimensional space and then assign a binary string to each district. To provide the relative position of districts the mapping is based on space-filling curves such as Z-ordering[15].

Algorithm 1. Mapping administrative districts into binary strings

1. Compute the centroid of each district.
2. Divide the region into two sub-regions, south and north, so that the numbers of centroids in both south and north are Similar.
3. If region south has more than one centroid, divide it into two sub-regions, South-east and south-west, so that the numbers of centroids in both south-east and South-west are similar.
4. Do the same for region north symmetrically.
5. For each sub-region obtained from Steps 3 and 4, if it contains more than one centroid, repeat Steps 2–4.
6. Considering the division process undergone, map each district onto a two-dimensional space.
7. Using a Z-ordering, assign a binary string to each district.

Fig 1 shows the division process of a country having 8 districts. Each district is labeled by its centroid (a, b,..h). The hierarchical division process is numbered as 1,2 3 and 4. Fig 2 shows the mapping of the regions into two-dimensional space. Binary string conversion of each district is shown in fig 3 where the relative ordering is based on Z-ordering.

The proposed encoding based on Z-ordering produces more informative binary strings. Let us consider two moving objects, one located at the district 00 000 and other located at 00 001. In addition to the fact that two objects are in the same country but in different cities, we can infer that (1) since the first two bits for cities are all 00, the cities are located at southwest area of the country, and (2) since the last bits for cities are different, the city where the first object is located is south of the city where the second object is located.

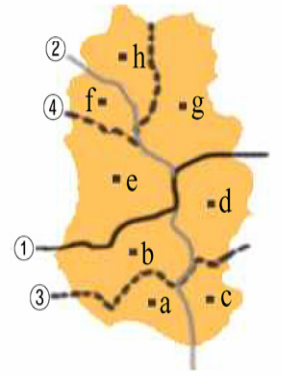


Fig. 1.

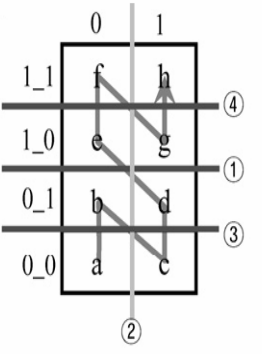


Fig. 2.

a: 000
b: 001
c: 010
d: 011
e: 100
f: 101
g: 110
h: 111

Fig. 3.

Algorithm 1 can be used to map road within each district into binary string by considering region for district and road as line. When a line is too curvy so that the centroid finding will be difficult then it can be partitioned and may be treated as different lines with separate code strings. Once all roads are partitioned into sub-roads, their mappings to binary strings are performed using Algorithm 1. The code for road will be concatenated with the binary code for the district. For example in the binary string 00000010 represents a road in district A(code 00000) where code for the road is 010.

To encode the location on road, we first partition a road into 2^n-1 units of the same size, and then represent each boundary as an n-bit binary string. Then we choose the boundary nearest from an object and use its binary string as the location of the object on the road.

Algorithm 2[18] will fix the position of a moving object represented by geometric co-ordinates as usually supplied by the GPS, onto roads. The algorithm uses an R-tree for roads to quickly convert a two-dimensional point into equivalent binary string.

Algorithm 2. Utilizing an R-tree to quickly convert a two-dimensional point, (x,y), into the equivalent binary String

---

1. Generate the rectangle uMBR by expanding x to its left and right by uR, and expanding y up and down by uR. uMBR is then expressed as ([x _ uR,x + uR], [y _ uR,y + uR]). Here, uR is a system parameter used for determining the nearness of roads from a two- dimensional point.
2. Search the R-tree for the roads whose MBRs overlap uMBR.
3. From the roads obtained in Step 2, select the road R whose Euclidean distance to (x,y) is the smallest.
4. Project (x,y) onto the road R. Let (x_0,y_0) denote the coordinate of (x,y) after the projection.
5. Using the relative position of (x_0,y_0) on the road R, calculate the binary string for (x_0,y_0).
6. Concatenate bit string(R) and the binary string for (x_0,y_0)

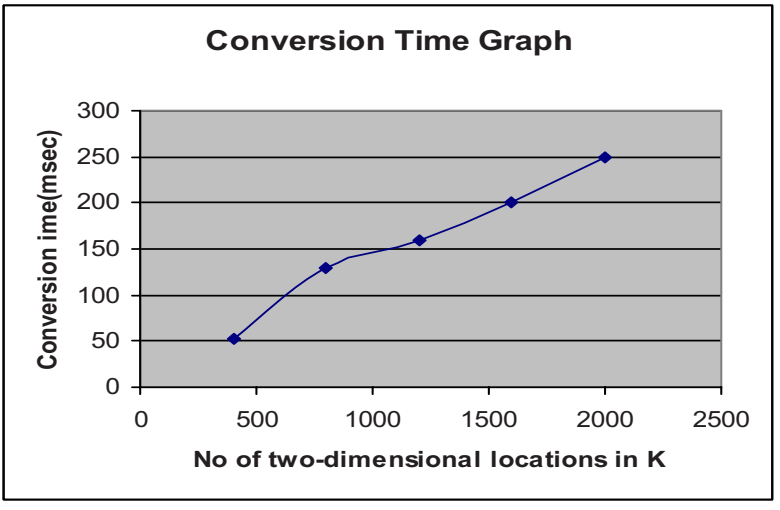


Fig. 4. Elapsed time to convert two-dimensional locations into equivalent binary string

Experimental validation [18] proved that the overhead in converting (x,y) into binary string will be negligible for a typical LBS environment and tolerable even for an environment with a huge number of locations. Fig. 4 shows that the total time spent for the conversion increases almost linearly with the number of two-dimensional locations and more than seven million locations can be processed per second.

The characteristics of the binary encoding scheme which made base line for the proposed alarming scheme in [18] are (i) It will be easy to find out the lowest common administrative district by extracting the longest common prefix of a given set of binary strings, and (ii) a district containing a set of lower districts can be represented by the range of binary strings; for example, county ''A'' in Fig. 1 is represented by the range [00000, 00111]. These advantages make it easy in addressing the whole country, whole district or of a single road by identifying the common prefix in the binary string representing the location of the object.

For example if the object's location on the road encoded as 001010000110010 where first 2 bits for the country, 4 bits for the district, 5 bits for road and 4 bits for relative location on the road. Then in all locations if first 2 bits are 00 then they all belongs to the same country.

4 Proposed Trigger Based Alarming Scheme

In the proposed scheme, moving object data base is a relational database having the following table structures.

a) **Current-objects (object-id, location bit sequence, Time):** Contains the details of the moving objects currently on roads under consideration. The object will update this table at frequent time intervals on the assumption that the system will have a

continuous network connectivity and sufficient capacity to handle large amount of data. Our proposed trigger will act on this table upon an update operation.

b) Sensitive Districts(District binary code, ts, te, status): Contains the details of district which are sensitive due to a terrorist attack, communal violence, natural calamity, emergency declared by local administration etc. Each record contains the equivalent binary string of the district obtained as described in section 3, ts and te are time start and time end respectively to denote the period of disturbance[2] and status denotes whether to set the area sensitive or not.

c) Sensitive Roads (Road binary code, ts, te, status): Similar to Sensitive districts table but contains details of sensitive roads.

The environment in which the alarming system will function is shown in fig 5. The mod will be informed about the sensitive area by highway police through authorized messages. The police people will inform the sensitive area in real-life address and the algorithm discussed in section 3 can be used to convert that into binary string notation and will set active status in mod. Alarming will work for objects already within the sensitive area and will be triggered in the next immediate update so that they can be vigilant or changing their routes or stop the movement. Advantage of this scheme is that a complete district, complete city or the whole single road can be set or reset to sensitive state easily as in the binary encoding scheme discussed earlier has a fixed substring to represent a district or city or road or part of a road. When an object crosses a sensitive area or being in such area, an update of current location will automatically triggered and message will be send to the object with sensitive location information in real-life address. A similar message will pass to patrolling police with information of object id and the location it has reached in the last update.

The scheme consists of three algorithms as explained below.

Algorithm 3. Marking sensitive area on MOD

---

1. Read the address of the area in real-life notation and duration from security people
2. Check the authority of the message.
3. Search on the B-tree for district/road names to find out the equivalent binary string.
4. Write the obtained binary string into district-sensitive table or road-sensitive table and set the status active

---

Algorithm 4. MOD updating by a moving object at fixed intervals

---

1. Get the location(x,y) of the moving object through the mobile set having GPS facility and time of updating.
2. Convert the location(x,y) into binary string using algorithm explained in section 3 (Algorithm 2).
3. Update the moving object data base (objects table) with record consisting of (object- id, binary string representing location, time)
4. During updating if the location is in any of the sensitive area a trigger "security-alarm" will be fired as detailed in the following algorithm.

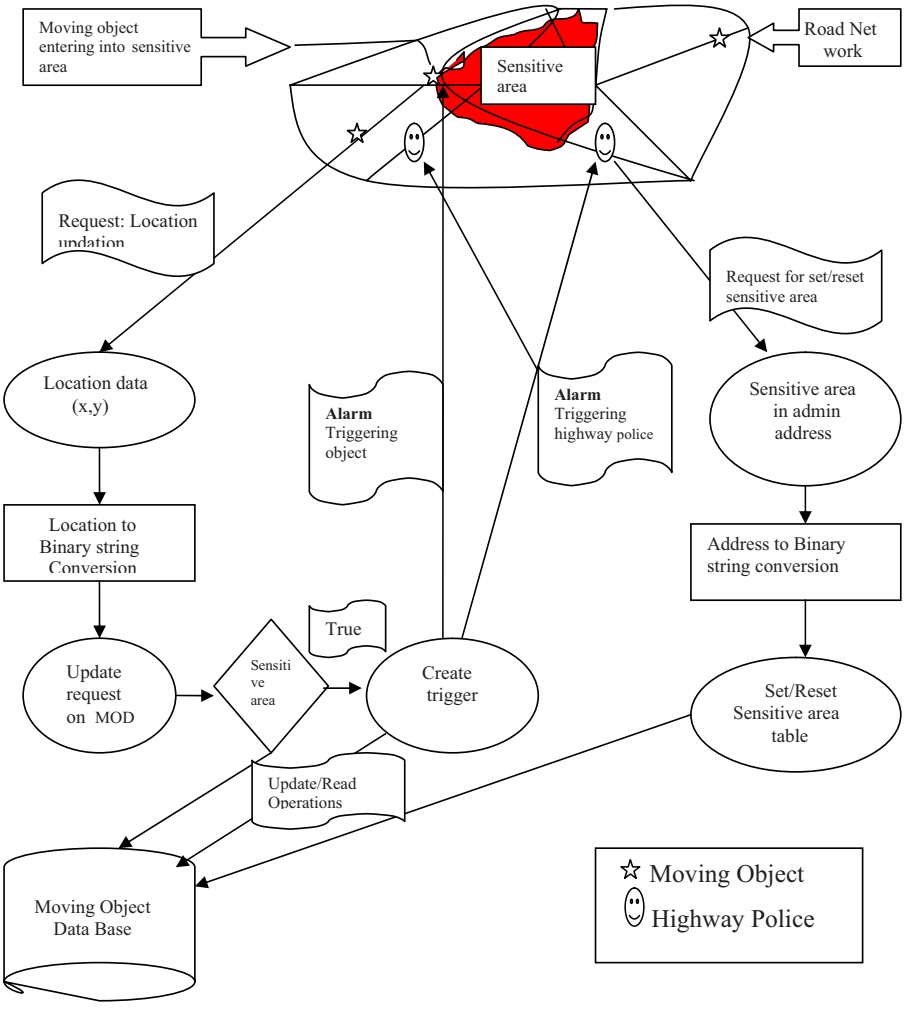


Fig. 5. Alarming System Environment

Algorithm 5: Trigger "security-alarm"

Create or alter trigger **security-alarm** before update of Objects for each row
Begin

1. If district binary substring (or road binary substring) in the location code is in District-sensitive table (or road-sensitive table)
2. f status column in district sensitive table (or sensitive-road table) is "yes"
3. If time to update match with the disturbance time from sensitive-district table (Or sensitive-road table)
4. Search the B-tree for address of the disturbed area where binary code is the key

5. Send a security-alarm to object through object-id with message containing Real-life address of location, city, road, part of road affected
6. Send a security-warning to highway police through Police-id containing the object-id , road/district entered into, time of entry
   End.

---

5 Conclusion

Dimension reduction is one of the challenging problems in multidimensional data management which has many applications in area that handle large volumes of data. Security informatics is such an area where huge spatio-temporal data is to be managed. The proposed alarming scheme for moving object on road network uses one of the dimension reduction methods over large scale spatio-temporal data with the concept of binary encoding techniques. The method could be extended to other domain such as health informatics in identifying and managing area of spreading infectious diseases.

References

1. Zeng, D., Cahng, W., Chen, H.: Clustering based Spatio- Temporal Hotspot Analysis Techniqes in Security Informatics. IEEE Transactions on Intelligent Transportation Systems
2. Trajcevski, G.: Context-Aware Optimization of Continuous Range Queries Maintenance for Trajectories (2005)
3. Prasad, M.: Location Based Services (2002), `http://www.GISdevelopment.net`
4. Jensen, C.S., Pedersen, T.B., Speicys, L., Timko, I.: Data Modeling for Mobile services in the Real World. In: Proceedings of the Eighth International Symposium on Spatial and Temporal Databases (2003)
5. Hage, C., Jensen, C.S., Pedersen, T.B., Speicys, I., Timko, I.: Integrated Data Management for Mobile services in the Real World. In: Proceedings of the Twenty Ninth International Conference on Very Large Databaes (2003)
6. Gupta, S., Kopparty, S., Ravishankar, C.: Roads, codes, and spatiotemporal queries. In: Proc. 22th ACM SIGACT-SIGMODSIGART Symposium on Principles of Database Systems, San Diego, California, pp. 115–124 (2004)
7. Papadias, D., Zhang, J., Mamoulis, N., Tao, Y.: Query processing in spatial network databases. In: Proc. 29th International Conference on Very Large Databases, Berlin, pp. 802–813 (2003)
8. Pfoser, D., Jensen, C.S.: Indexing of network constrained moving objects. In: 11th ACM Symposium on advances in Geographic Information Systems, New Orleans, Louisiana, pp. 25–32 (2003)
9. Pfoser, D., Jensen, C.S.: Trajectory indexing using movement constraints. In: V. Springer Science and Business Media B.V., GeoInformatica, pp. 25–32 (2005)
10. Faloutsos, C., Roseman, S.: Fractals for secondary key retrieval. In: 8th ACM SIGACTSIGMOD-SIGART Symposium on Principles of Database Systems, Philadelphia, Pennsylvania, pp. 247–252 (1989)

11. Almeida, V., Guting, R.H.: Indexing the trajectories of moving objects in networks. In: 16th International Conference on Scientific and Statistical Database Management, Santorini Island, pp. 115–118 (2004)
12. Almeida, V., Guting, R.H.: Indexing the trajectories of moving objects in networks. In: V. Springer Science and Business Media B.V., GeoInformatica, pp. 33–60 (2005)
13. Chakka, V.P., Everspaugh, A., Patel, J.M.: Indexing large trajectory data sets with SETI. In: 1st Conference on Innovative Data Systems Research, Asilomar, CA, pp. 164–175 (2003)
14. Frentzos, E.: Indexing objects moving on fixed networks. In: 8th International Symposium on Spatial and Temporal Databases, Santorini Island, pp. 289–305 (2003)
15. Orenstein, J.A., Merrett, T.H.: A class of data structures for associative searching. In: 3rd ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, Waterloo, Ontario, pp. 181–190 (1984)
16. Chang, W., Zeng, D., Chen, H.: Prospective Spatio-Temporal Data Anaysis for security Informatics, Department of management Information Systems at the University if Arizona, Tueson
17. Pfoser, D., Jensen, C.S., Theodoridis, Y.: Novel approaches in query processing for moving object trajectories. In: VLDB, pp. 395–406 (2000)
18. Lee, S., Park, S., Kim, W.-C.: An efficient location encoding method for moving objects using hierarchical administrative district and road network. Information Sciences 177, 832–843 (2007)

Reducing False Alarm of Video-Based Smoke Detection by Support Vector Machine

Chan-Yun Yang[1], Wei-Wen Tseng[2], and Jr-Syu Yang[3]

[1] Department of Mechanical Engineering, Technology and Science Institute of Northern Taiwan, No. 2 Xue-Yuan Rd., Beitou, Taipei, Taiwan, 112
cyyang.research@gmail.com
[2] Department of Fire Science, Central Police University, No. 56 Shuren Rd., Guishan Shiang, Taoyuan County, Taiwan, 333
weiwen.tseng@mail.cpu.edu.tw
[3] Department of Mechanical and Electro-Mechanical Engineering Tamkang University Taipei, Taiwan, 251
096034@mail.tku.edu.tw

Abstract. Techniques used in video smoke detection systems have been discussed noticeably in past few years. With the advantage of early fire alarm in large or specific spaces such as studio and tunnels, the video-based smoke detection systems would not have time delay as conventional detectors. In contrast, how to reduce false alarm and increase the generalization ability is the key issue for such state-of-the-art systems. In this paper, examples consisting of features extracted from a real time video are collected for the training of a discriminating model. A prototype of support vector machine (SVM) is therefore introduced for the discriminating model with the capability in small sample size training and the good generalization ability. In order to reduce the false alarm, the prototype is then extended to a class-imbalanced learning model to deal with rarity of the positive class. A number of assuming data are used for imbalanced test to cope with the real world of fire safety. The technique is optimistic to enhance accuracy and reduce false alarm in video-based smoke systems.

Keywords: Margin, Video-based Smoke detection, False alarm, Pattern recognition, Support vector machines.

1 Introduction

Conventional spot thermal and smoke detection techniques, there are some drawbacks to be solved. A spot detector takes charge of a limited area in space. In addition, the area may be affected by environment like forced ventilation, so the smoke or fire would not be detected or have a long time delay. Therefore, those conventional detection systems are not suitable in some large or specific spaces such as hangers, studio, atriums, tunnels, storage, and offshore platform. Video-based fire systems are developed to solve above deficiencies in past few years. However, the most of video-based techniques to give fire alarm are aim at flame detection. For early alarm purpose, the

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 102–113, 2008.

researches and techniques on video-based smoke detection are more and more paid attention. However, such state-of-the-art system has great technical challenges in terms of detection rate and false alarm rate [1]. Based on recent proposed methods of video capture in smoke signals, Chen [2] used the combination of a chromaticity-based statistic and a diffusion-based dynamic characteristic to reorganize smoke generation. Kopilovi [3] proposed irregularities in motion due to non-rigidity of smoke. Toreyin et al. [4] extracted image features by using the methods of background subtraction, temporal wavelet transformation, and spatial wavelet transformation which is the algorithm of smoke detection used in this paper.

The study reviewed the video-based techniques for smoke detection. As a framework of an intelligent early alarming system, the study proposed a classification post-processor to produce the alarm automatically. Based on different sensing devices, there are several types of artificial neural networks (ANNs) have been developed as the post-processor [5]. Since the method of support vector machine (SVM) [6-7] has demonstrated good generalization ability in classification [8], a post-processor based on the SVM is studied and presented in this paper. Unlike the ANNs minimize only the empirical risk of the training data which may incur an unexpected overfitting, the SVM seeks to maximize classification "margin" in the solution in which good generalization ability is accessible.

Basically, the discriminant function coming from the post-processor shall be supervised learned with some positive (real fire) and negative (normal condition) examples in advance. One difficulty arose when collecting examples for learning. Whatever the techniques are applied to acquire the smoke signals, the learning of the post-processor is still suffered from the rarity of the real positive learning examples. Machine learning based on the rare positive examples leads to a biased discriminant function and conducts generally the system a higher rate of false alarm. As known, the false alarm degraded the generalization ability of the intelligent early alarming system.

With the arguments, the post-processor has to employ further a technique of class imbalanced learning to deal with the false alarm. In machine learning, topics of class imbalanced learning can mainly be categorized into two levels, the data and algorithmic levels of modification [9]. At the data level, strategies of up-sampling and down-sampling are often used to deal with the imbalanced datasets [9-10]. The down-sampling eliminates the samples in majority class while up-sampling duplicates those in minority-class. Both techniques reduce the degree of imbalance. At the algorithmic level, people introduce the cost-sensitive learning as a solution for imbalanced class learning. This kind of strategies gives higher learning cost to the samples in the minority-class to counter balance the degree of imbalance [11-12]. A general practice is to exploit the misclassification costs of identifying the majority-class to outweigh those of identifying the minority-class. The reweighing scheme is generally merged into the common edition of classification algorithms [12].

2 Smoke Detection Approach

This paper follows the approach of Toreyn et al., [4] and Shuai [13], i.e., Smoke detection algorithm consisting of using background estimate methods to determine

moving pixels and regions in the video, examining temporal wavelet sub-signals, analyzing spatial wavelet features, and checking *U*/*V* channels of the images to collect related parameters.

Moving pixels and regions in the video are determined by using a background estimation method developed by Collins et al., [14]. In this method, a background image $\mathbf{B}_{n+1}$ at time instant n+1 is recursively estimated from the image frame $\mathbf{I}_n$ and the background image $\mathbf{B}_n$ of the video as follows:

$$\mathbf{B}_{n+1}(s,t) = \lambda \mathbf{B}_n(s,t) + (1-\lambda)\mathbf{I}_n(s,t) \quad \text{stationary,} \tag{1}$$

$$\mathbf{B}_{n+1}(s,t) = \mathbf{B}_n(s,t) \quad \text{moving,} \tag{2}$$

where $\mathbf{I}_n(s, t)$ represent a pixel in the n^{th} video frame $\mathbf{I}_n$, and λ is a parameter between 0 and 1. Moving pixels are determined by subtracting the current image from the background image and shareholding.

From temporal wavelet transform, the signal $u_n(s, t)$ is the luminance (*Y* component) of the image. To examine the wavelet sub-signals $d_n(s, t)$ and $e_n(s, t)$ at 5 Hz image capture rate. In a stationary pixel, values of these two sub-signals should be equal to or very close to zero because of high-pass filters used in sub-band analysis. If there is an ordinary moving object going through pixel (s, t) then there will be a single spike in one of these wavelet sub-signals because of the transition from the background pixel to the object pixel. If the pixel is part of a smoke boundary then there will be several spikes in one second due to transitions from background to smoke and smoke to background. Therefore, if $|e_n(s, t)|$ and/or $|d_n(s, t)|$ exceed a threshold value several times in a few seconds then an alarm is issued for this pixel. By using spatial transform, let

$$\omega_n(u,v) = |LH_n(u,v)|^2 + |HL_n(u,v)|^2 + |HH_n(u,v)|^2, \tag{3}$$

represent a composite image containing high-frequency information at a given scale. Wavelet sub-images *LH*, *HL* and *HH* contains horizontal, vertical and diagonal edges of the original image, respectively. This sub-band image is divided into small blocks of size (s_1, s_2) and the energy $e(t_1, t_2)$ of each block is computed as follows

$$e(t_1,t_2) = \sum_{(x,y)\in R_i} \omega_n(u + t_1 s_1, v + t_2 s_2), \tag{4}$$

where R_i represents a block of size (s_1, s_2) in the wavelet sub-image.

Color information is also used for identifying smoke in video. Initially, when the smoke starts to expand, it is semi-transparent thus it preserves the direction of the RGB vector of the background image. This is another clue for differentiating between smoke and an ordinary moving object. By itself, this information is not sufficient because shadows of moving objects also have the same property. As the smoke gets thicker, however, the resemblance of the current frame and the background decreases and the chrominance values *U* and *V* of the candidate region in the current frame become smaller than corresponding values in the background image. Only those pixels with lower chrominance values are considered to be smoke.

The parameters of the derived models in this section are used as input feature vectors to train an SVM classifier, which is then used to detect the presence of steam in

the video image. A flowchart is provided in Fig. 1 to better illustrate the proposed algorithm. From the proposition, an eight-dimensional input vector is formed as the input of the SVM:

$$\mathbf{x} = [\lambda, d_n, e_n, w_n, e, Y, U, V]^T. \tag{5}$$

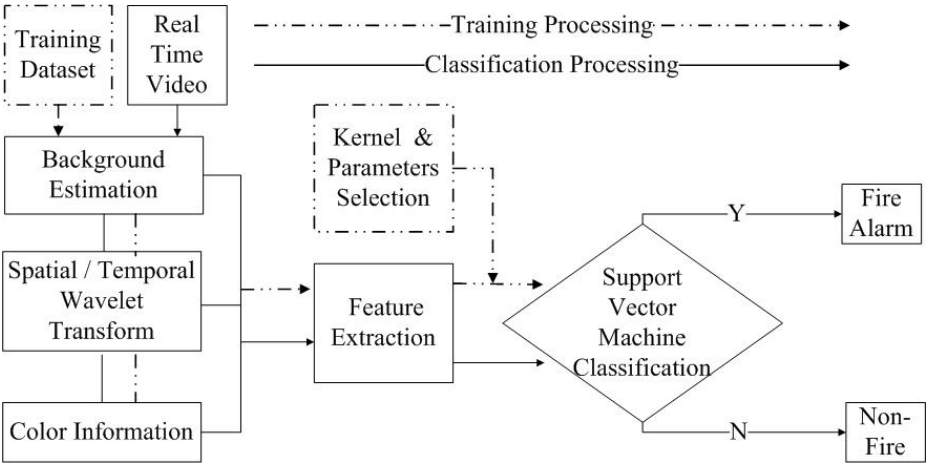


Fig. 1. Flowchart of techniques for the video-based smoke detection with SVM classification

3 Support Vector Machine Discriminating Post-processor

3.1 Elementary Support Vector Machine

The discriminating post-processor involves constructing a discriminant function to classify video-based signals for prediction. With a set of labeled learning examples $S = \{(\mathbf{x}_i, y_i) \mid i = 1, 2, \ldots, l\}$ where y_i's are known classifications, $y_i \in \{+1, -1\}$ the discriminant function can be built by the post-processor. As a kind of the kernel methods [8, 15], SVM [6-7] employs a feature map, $\varphi(\mathbf{x})$: $R^d \rightarrow R^H$, implicitly mapping the d-dimensional learning examples from the input space R^d into a higher dimensional Hilbert feature space R^H, to manipulate linearly a sophisticated non-linear separating problem. With the kernel method, most of real world applications known generally as non-linear problems can be tackled. With the map $\varphi(\mathbf{x})$, SVM learns from the learning set S and builds a linear discriminant function f in the feature space R^H:

$$f(\mathbf{x}) = \mathbf{w}^T \varphi(\mathbf{x}) + b, \tag{6}$$

where $\mathbf{w}$ is a weight vector denoting a d-dimensional transposed normal vector to the separating hyperplane, and b is a bias term denoting a scalar for offsetting the hyperplane.

A key to guarantee that the SVM is capable of achieving a good generalization performance is to maximize the distance ρ, termed as "margin," between the separating hyperplane and those learning examples lying closest to it [6, 8]. The basic ideal of

SVM programming is sought to maximize the margin of the separating hyperplane. With a consideration of the errors produced by heterogeneous samples which might come from a contaminated or noisy non-separable learning set, a vector of slack variables $[\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_l]^T$, $\varepsilon_i \geq 0$, is introduced to participate the optimization programming. An eventual revision of soft-margin SVM is given as [16]:

$$\min \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{l}\varepsilon_i, \tag{7}$$

subject to constraints:

$$y_i(\mathbf{w}^T\varphi(\mathbf{x}_i) + b) \geq 1 - \varepsilon_i, \text{ and} \tag{8}$$

$$\varepsilon_i \geq 0, \text{ for } i = 1, 2, ..., l, \tag{9}$$

where C is a regularization parameter to adjust the ratio of both objective terms in the optimization. The model with slack variables renders more the SVM as a general model to deal with the real world applications since its capability in dealing noisy input signals [8]. The capability is essentially achieved by the penalization term in (7):

$$R_{\text{emp}} = \sum_{i=1}^{l}\varepsilon_i. \tag{10}$$

Here, R_{emp}, called empirical risk, states classification risk in the programming. In the statistical learning theory [6-7], a major assumption often drawn in a learning machine is that there is no distribution information about experimental data known in advance, but only experimental data themselves. The empirical risk, instead of expected risk, is taken to assess performance. The penalized objective function (7) is aimed to minimize not only the inversion of ρ but also the classification risk. The race of the two minimizing operations is regularized by C. In general, penalty constant C scales the learning cost of those samples with non-zero ε, and controls therefore the optimization to achieve a good generalization performance. In the study, C is also an important key to achieve the goal of reducing false alarm.

A vector of Lagrange multipliers $[\alpha_1, \alpha_2, \ldots, \alpha_l]^T$ is therefore introduced to convert the primal formulations of soft-margin SVM (7) - (9) into a dual problem:

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j \kappa(\mathbf{x}_i, \mathbf{x}_j), \tag{11}$$

subject to:

$$\sum_{i=1}^{l} y_i\alpha_i = 0, \text{and} \tag{12}$$

$$0 \leq \alpha_i \leq C, \text{ for } i = 1, 2, ..., l, \tag{13}$$

where $\kappa(\cdot,\cdot)$ is a kernel function that is given by

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \varphi(\mathbf{x}_i)^T\varphi(\mathbf{x}_j). \tag{14}$$

The dual form of (11) renders the classification problem easier to be solved by some existing convex quadratic programming algorithms [17]. In the above expression, the kernel function, taking inner-product of pair-wise high dimensional input feature map, is a feasible solution for manipulating the similarity between two input patterns [8, 15]. Since the direct operation of the inner product of feature maps may be infeasible if the dimension of feature space is very high due to the rapidly increasing computation complexity. The kernel function provides alternatively a practical way to compute implicitly the inner-product without mapping into such a high-dimensional space. Types of kernel function for SVM are [17]:

linear kernel: $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$,

polynomial kernel: $\kappa(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^T \mathbf{x}_j + 1)^d$, and

RBF kernel: $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2 / \sigma^2)$,

where d is a positive integer as the polynomial order of the polynomial kernel, and σ denotes the width parameter for the RBF basis function in the RBF kernel.

After the quadratic programming of (11) - (13), a set of exact solution of α_i^* associated with the learning examples is found. According to the sparseness of the KKT (Karush-Kuhn-Tucker) conditions [8], only those examples having non-zero α_i^*, called support vectors, are preserved to support the discriminant function [6]. The crucial property of the sparseness in support vector expansion reduces the computing time of the subsequent discriminant function, and is beneficial to an on-line prediction system. As illustrated, the discriminant function f can be fully specified using only the minor subset of support vectors:

$$f(\mathbf{x}) = \sum_{\alpha_i^* \neq 0} \alpha_i^* y_i \kappa(\mathbf{x}, \mathbf{x}_i) + b. \tag{15}$$

Once the discriminant function established, the intelligent early alarming system can employ the function for on-line prediction with instantaneous video-based signals.

3.2 Class Imbalanced Learning

In order to solve the problem of false alarm in the smoke detection system, the elementary SVM shall consider further the issue of class imbalanced learning. Paper survey shows the development of class imbalanced learning in SVM shares the same merits of two-level techniques which we have described in section I. Known as the techniques in data level , variant sampling strategies have been employed in solving some real applications [18]. Akbani et al., [19] have explained why the down-sampling strategy is not the best choice for SVM. They developed a method incorporating synthetic minority up-sampling technique (SMOTE) [20] with different error costs algorithm [21] to push the biased separating hyperplane away from minority-class. Ertekin et al., [22] proposed an efficient active learning strategy for down-sampling. The method iteratively selects the closest example to the separating hyperplane from the unseen learning data and adds it to the learning set to retrain the SVM. With an early stopping criterion, the method can decrease significantly the learning time in the large scale imbalanced dataset. In addition, a backward pruning technique, identified as one type of down-sampling, has been employed to deal with

the imbalanced data classification [23]. In the algorithmic level, Veropoulos et al., [21] suggested a solution for cost-sensitive learning which used different penalty constants for different classes of data to make errors on minority-class samples costlier than errors on majority-class samples. The penalty regularized method deserves to be much more attention because the promising formulation is intrinsically coherent with its elementary prototype of SVM. In fact, the remedy has widely been applied and extended in many applications [19, 24]. The other type of approaches to dealing with imbalanced SVM learning is to push the learned hyperplane further away from the minority-class in the algorithms. This can also be done by cost-sensitive learning. Modifying the kernel function (Fig.2) provides one of the solutions for such cost-sensitive learning [25]. The kernel function can be conformally transformed according to the structure of the Riemannian geometry in the imbalanced data to improve the bias.

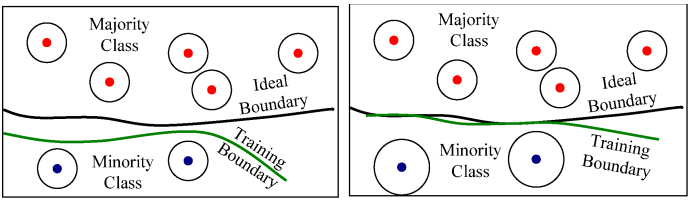


Fig. 2. Adjustment of decision boundary by a conformally transformed kernel function

The study adopts Veropoulos model [21] for the class imbalanced learning due to its coherence with the elementary prototype. The key idea of the penalty regularized model is to introduce unequal penalties to the samples in the imbalanced classes. The penalization strategy associated with error of a positive sample retains penalty higher than that with error of a negative sample in the optimization. The high penalty then translates into a bias for Lagrange multiplier because the cost of corresponding error is heavier. This drifts the separating hyperplane from the positive class towards the negative class. The model can then be started with the examination of imbalance in the learning set S. Suppose there are l^+ positive and l^- negative learning examples in S, we have:

$$S = \{(\mathbf{x}_p, y_p) \cup (\mathbf{x}_n, y_n) \mid y_p = +1,\ y_n = -1, \mathbf{x} \in \Re^d\}, \tag{16}$$

where p and n denote respectively the indices of the example in the positive and negative class:

$$\begin{aligned} &\{p \mid y_p = +1,\ p = 1, 2, ..., l^+\}, \\ &\{n \mid y_n = -1, n = 1, 2, ..., l^-\}, \text{ and} \\ &l = l^+ + l^-. \end{aligned} \tag{17}$$

With set S, the Veropoulos model based on the soft-margin SVM has been founded to learn the target discriminant function $f(\mathbf{x}) = \mathbf{w}^T \varphi(\mathbf{x}) + b$:

$$\min \frac{1}{2}\|\mathbf{w}\|^2 + C^+ \sum_{p=1}^{l^+} \varepsilon_p + C^- \sum_{n=1}^{l^-} \varepsilon_n, \tag{18}$$

subject to

$$y_p(\mathbf{w}^T \varphi(\mathbf{x}_p) + b) \geq 1 - \varepsilon_p, \text{ for } \{p \mid y_p = +1\}, \tag{19}$$

$$y_n(\mathbf{w}^T \varphi(\mathbf{x}_n) + b) \geq 1 - \varepsilon_n, \text{ for } \{n \mid y_n = -1\}, \text{ and} \tag{20}$$

$$\varepsilon_p,\ \varepsilon_n \geq 0, \tag{21}$$

where C^+ and C^- denote the penalty constants for positive and negative class, respectively. As understood, the model biases the separating hyperplane by assigning different cost for errors in the different class. In general, the error in the positive class is costlier than that in negative class. The smaller the scale of the positive class, the higher the error cost.

Following similar derivations in the elementary SVM, the dual form of the class imbalanced learning problem can also be expressed as a convex quadratic optimization problem:

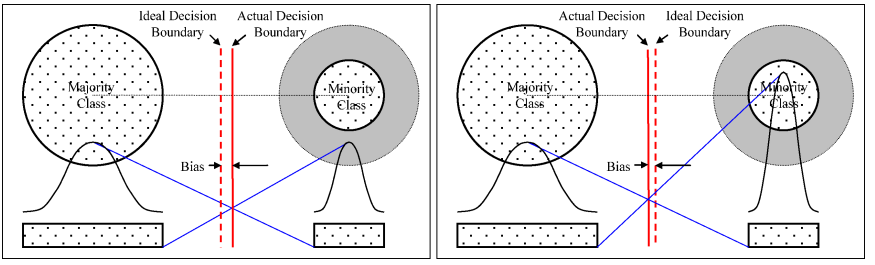
$$\begin{aligned} &\max_{\boldsymbol{\alpha}} \sum_{p=1}^{l^+} \alpha_p + \sum_{n=1}^{l^-} \alpha_n \\ &-\frac{1}{2}\Big(\sum_{p=1}^{l^+}\sum_{p=1}^{l^+} \alpha_p^2 y_p^2 \kappa(\mathbf{x}_p, \mathbf{x}_p) + 2\sum_{p=1}^{l^+}\sum_{n=1}^{l^-} \alpha_p \alpha_n y_p y_n \kappa(\mathbf{x}_p, \mathbf{x}_n) \\ &+ \sum_{n=1}^{l^-}\sum_{n=1}^{l^-} \alpha_n^2 y_n^2 \kappa(\mathbf{x}_n, \mathbf{x}_n)\Big) \end{aligned} \tag{22}$$

subject to

$$\sum_{p=1}^{l^+} \alpha_p = \sum_{n=1}^{l^-} \alpha_n, \text{ and} \tag{23}$$

$$0 \leq \alpha_p \leq C^+,\ 0 \leq \alpha_n \leq C^-. \tag{24}$$

Figure 3 illustrates the compensated effect of the Veropoulos model. Using the region of a shaded-dish to describe the size, two classes, showing as both top and lateral views, are aligned with their centers horizontally. A Gaussian distribution is assumed for data points in the shaded-dish regions. The inclined segments risen from the horizon to the top of the Gaussian curves are analogous to the penalty functions. In the beginning, the heights of the assumed Gaussian are equally normalized for an uncompensated condition. The decision boundary (separating hyperplane) drawn from the intersection of the segments vertically is actually biased from the ideal decision boundary (Fig. 3a). The Veropoulos model employing costlier penalty for misclassifications in positive class can be analogous to raise the height of the corresponding Gaussian. Due to the raised height, the decision boundary drew from the intersection shifts closer towards the ideal boundary (Fig. 3b).

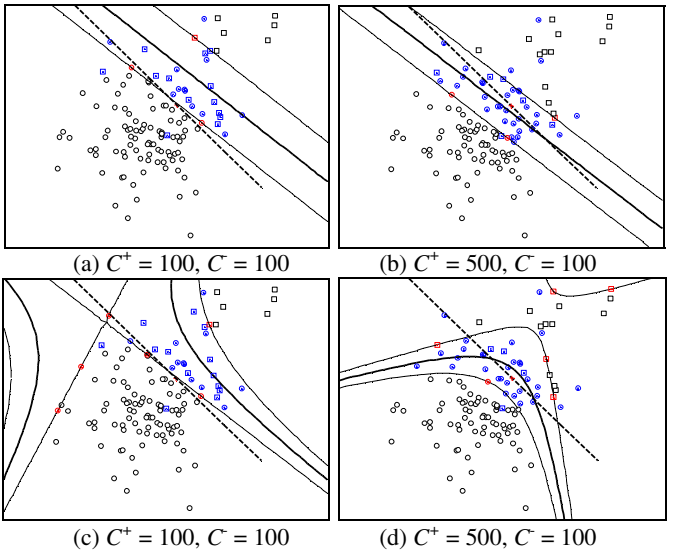


(a) Uncompensated learning (b) Compensated with costlier penalty

Fig. 3. Basic idea of the Veropoulos model

4 Computer Simulation

An issue to check the feasibility of the imbalanced learning in reducing the false alarm is performed first. A simulated imbalanced dataset consisting 20 positive examples (marked as "□") and 100 negative examples (marked as "○") is acquired with Breiman's algorithm [26] and is given in panels of Fig. 4 with an ideal decision boundary showing as a dashed line. As shown, the bias decision boundaries (solid lines in Fig. 4a and 4c) are moved closed to the ideal boundary (solid lines in Fig. 4b and 4d) by applying heavier penalties to positive examples despite the linear or second polynomial kernel is used. The decision boundary closed to the ideal boundary is



(a) $C^+ = 100$, $C^- = 100$ (b) $C^+ = 500$, $C^- = 100$

(c) $C^+ = 100$, $C^- = 100$ (d) $C^+ = 500$, $C^- = 100$

Fig. 4. Decision boundaries with ((b) and (d)) / without ((a) and (c)) the compensation. The dashed lines denote the ideal decision boundary, the solid lines denote the actual decision boundary, and dotted lines denotes respective margins.

good for the predictions of forthcoming smoke signals wherever the attributed values are unknown. Of course, it is good for generalization performance. For the sake to assess the generalization performance of the imbalanced learning, a measure of *Gmean* is employed as an index of the performance [27]:

$$Gmean = \sqrt{acc^{+} \cdot acc^{-}}, \tag{25}$$

where acc^{+} is the percentage of positive examples correctly recognized and acc^{-} is the percentage of negative examples correctly recognized in the classification. Actually, *Gmean*, the geometric mean of the values of acc^{+} and acc^{-}, will become small with either small acc^{+} or small acc^{-}. Only the condition of simultaneously large in both acc^{+} and acc^{-} makes *Gmean* a high score. The fact implies a good generalization performance is generally come with the high score of *Gmean*.

Using a 10-fold cross-validation on the same imbalanced dataset, Table 1 shows averaged *Gmean*'s with different settings of C^{+} and C^{-}. As known, the larger value of *Gmean*, the higher the generalization performance. The table concluded that a heavier penalty for larger value of C^{+} increases the performance, and reduces confidently the false alarm.

Table 1. Averaged *Gmean* in 10-fold cross-validation

C^{+}/C^{-} Ratio	$C^{-} = 0.1$		$C^{-} = 1$		$C^{-} = 10$	
	Gmean	acc^{+}/acc^{-}	*Gmean*	acc^{+}/acc^{-}	*Gmean*	acc^{+}/acc^{-}
1	0.77	0.70/0.96	0.77	0.70/0.94	0.77	0.70/0.94
1.5	0.78	0.73/0.94	0.80	0.77/0.94	0.79	0.77/0.93
2	0.81	0.80/0.93	0.79	0.77/0.93	0.81	0.80/0.93
2.5	0.83	0.83/0.93	0.88	0.87/0.92	0.85	0.83/0.90
3	0.83	0.83/0.93	0.87	0.87/0.90	0.86	0.87/0.89
3.5	0.85	0.83/0.90	0.85	0.87/0.87	0.86	0.87/0.88
4	0.81	0.85/0.81	0.84	0.89/0.83	0.84	0.87/0.84

5 Conclusion

In this paper, we proposed a real-time image processing algorithm for video-based smoke detection systems. The approach uses the parameters of background estimation, wavelet transform, and color information as features to characterize the smoke texture pattern. A class imbalanced learning algorithm is then proposed for automatic early alarm. The algorithm can be trained through SVM classification to reduce the false alarm in these state-of-the-art systems. Computer simulation shows a promising result that is worth for further realization as an application.

References

1. Xiong, Z., Caballero, R., Wang, H.-C., Finn, A.M., Leic, M.A., Peng, P.-Y.: Video-based somke detection: possibilities, techniques, and challenge. In: Suppression and Detection Research and Applications - A Technical Working Conference (SUPDET 2007), Orlando, Florida (2007)

2. Chen, T.-H., Yin, Y.-H., Huang, S.-F., Ye, Y.-T.: The smoke detection for early fire-alarming system base on video processing. In: International Conference on Intelligent Information Hiding and Multimedia Signal Processing, USA (2006)
3. Kopilovic, I., Vagvolgyi, B., Sziranyi, T.: Application of panoramic annular lens for motion analysis tasks: surveillance and smoke detection. In: 15th International Conference on Pattern Recognition, vol. 4, pp. 714–717 (2000)
4. Toreyin, B.-U., Dedeoglu, Y., Cetin, A.-E.: Wavelet based real-time smoke detection in video. In: EUSIPCO 2005 (2005)
5. Okayama, Y.: A primitive study of a fire detection method controlled by artificial neural net. Fire Safety Journal 17, 535–553 (1991)
6. Vapnik, V.N.: The nature of statistical learning theory. Springer, New York (1995)
7. Vapnik, V.N.: Statistical learning theory. John Wiley & Sons, New York (1998)
8. Schölkoph, B., Smola, A.J.: Learning with kernels. MIT Press, Cambridge (2002)
9. Chawla, N.V., Japkowicz, N., Kotcz, A.: Editorial: Special issue on learning from imbalanced data sets. SIKDD Explorations Newsletters 6(1), 1–6 (2004)
10. Visa, S., Ralescu, A.: Issues in mining imbalanced data sets - a review paper. In: Sixteen Midwest Artificial Intelligence and Cognitive Science Conference (MAICS 2005), Dayton, pp. 67–73 (2005)
11. Domingos, P.: MetaCost: a general method for making classifiers cost sensitive. In: Fifth international conference on knowledge discovery and data mining (ACM SIGKDD 1999), pp. 155–164. ACM press, New York (1999)
12. Elkan, C.: The foundations of cost-sensitive learning. In: Seventeenth international joint conference on artificial intelligence, pp. 973–978. Morgan Kaufmann, San Fransisco (2001)
13. Shi, S., Ping, Z., Wang, Y.-M., Zhou, W.-D.: Wavelet Based Real-time Smoke Detection. J. Application Research of Computers (in Chinese) 24(3) (March 2007)
14. Collins, R.T., Lipton, A.J., Kanade, T.: A System for Video Surveillance and Monitoring. In: American Nuclear Society 8th Int. Topical Meeting on Robotics and Remote Systems, Pittsburgh (1999)
15. Vapnik, V.N.: An Overview of Statistical Learning Theory. IEEE Transactions on Neural Networks 10, 988–999 (1999)
16. Cortes, C., Vapnik, V.N.: Support vector networks. Machine Learning 20, 273–297 (1995)
17. Chang, C.-C., Lin, C.-J.: LIBSVM: a library for support vector machines, http://www.csie.ntu.edu.tw/~cjlin/libsvm
18. Lessmann, S.: Solving imbalanced classification problems with support vector machines. In: International Conference on Artificial Intelligence (IC-AI), vol. 1, pp. 214–220 (2004)
19. Akbani, R., Kwek, S., Japkowicz, N.: Applying support vector machines to imbalanced datasets. In: Boulicaut, J.-F., Esposito, F., Giannotti, F., Pedreschi, D. (eds.) ECML 2004. LNCS (LNAI), vol. 3201, pp. 39–50. Springer, Heidelberg (2004)
20. Chawla, N., Bowyer, K., Hall, L., Kegelmeyer, P.: SMOTE: synthetic minority oversampling technique. Journal of Artificial Intelligence Research 16, 321–357 (2002)
21. Lin, C.-F., Wang, S.-D.: Fuzzy support vector machines. IEEE Transactions on Neural Network 13(2), 464–471 (2002)
22. Ertekin, S., Huang, J., Giles, C.L.: Active learning for class imbalance problem. In: 30th International Conference on Research and Development in Information Retrieval, ACM SIGIR, pp. 823–824 (2007)
23. Chen, X.W., Gerlach, B., Casasent, D.: Pruning support vectors for imbalanced data classification. In: International Joint Conference on Neural Networks (IJCNN 2005), pp. 1883–1888. IEEE, Montreal, Canada (2005)

24. Callut, J., Dupont, P.: F_{β} support vector machines. In: International Joint Conference on Neural Networks, pp. 1443–1448. IEEE Press, Montreal, Canada (2005)
25. Wu, S., Amari, S.: Conformal Transformation of Kernel Functions: a Data-dependent Way to Improve Support Vector Machine Classifiers. Neural Processing Letters 15, 59–67 (2002)
26. Breiman, L.: Bias, variance and arcing classifiers. Technical Report 460, Dept of Statistics, UC Berkeley, CA (1996)
27. Kubat, M., Holte, R., Matwin, S.: Learning when negative examples abound. In: van Someren, M., Widmer, G. (eds.) ECML 1997. LNCS, vol. 1224, pp. 146–153. Springer, Heidelberg (1997)

Privacy-Preserving Collaborative Social Networks

Justin Zhan, Gary Blosser, Chris Yang, and Lisa Singh

[1] Carnegie Mellon University
justinzh@andrew.cmu.edu
[2] Carnegie Mellon University
gblosser@andrew.cmu.edu
[3] Chinese University of Hong Kong
yang@se.cuhk.edu.hk
[4] Georgetown University
singh@cs.georgetown.edu

Abstract. A social network is the mapping and measuring of relationships and flows between individuals, groups, organizations, computers, web sites, and other information/knowledge processing entities. The nodes in the network are the people and groups, while the links show relationships or flows between the nodes. Social networks provide both a visual and a mathematical model for analyzing of relationships. While social network construction and analysis has taken place for a long time, social network analysis in the context of privacy-preservation is a relatively new area of research. In this paper, we focus on privately constructing a social network involving multiple independent parties. Because of privacy concerns, the parties cannot share their individual social network data directly. However, the parties could all benefit from the construction of a collaborative social network containing all the independent party network data. How multiple parties collaboratively construct a social network without breaching data privacy presents a challenge. The objective of this paper is to present a cryptographic approach for privately constructing collaborative social networks.

Keywords: Privacy, Security, Social Networks.

1 Introduction

Recent advances in data collection, data dissemination and related technologies have inaugurated a new era of research where existing social network algorithms should be reconsidered in the context of privacy preservation. The need for privacy is sometimes motivated by laws (e.g., disease transmission networks based on medical databases) or business interests (e.g., product affinity networks based on customer databases). However, there are situations where the sharing of data can lead to a mutual benefit. Despite the potential gain, this is often not possible due to the confidentiality issues which arise. It is well documented [4] that the unlimited explosion of new information through the Internet and other media

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 114–125, 2008.

has reached a point where threats against individual privacy are very common and deserve serious thought. Let us consider the following example. Suppose, several countries are involved in a multi-site terrorism study. Each country has its own data set containing terrorist social networks. These countries would like to construct a combined or *collaborative social network* using the data sets from all of the participating countries in order to more effectively identify common terrorist threats. However, due to privacy rules, one country cannot disclose its social networks to other countries. How can these countries achieve their objective? Can privacy and collaborative social network construction coexist? In other words, can the multiple, independent parties somehow construct a more comprehensive collaborative social network and obtain the desired results without compromising their data privacy? In this paper, we show that privacy and collaborative social network construction can be achieved at the same time over large data sets with reasonable efficiency.

Privacy-preservation in the context of social networks is a relatively new research area. At their seminal paper on privacy-preserving social network, Wang et. al. [9] generalized the techniques for protecting personal privacy in tabulated data and proposed some metrics of anonymity for risk analysis by disclosing social network data for public release. Singh and Zhan [8] presented a new measure, topological anonymity, that quantifies the amount of privacy preserved in different topological structures. Their measure uses a combination of known social network metrics and attempts to identify when node and edge inference breaches arise in these graphs. Zhou and Pei [11] present a type of privacy attack: neighborhood attacks. A neighborhoods attack occurs if an adversary has some knowledge about the neighbors of a target victim and the relationship among the neighbors. They present a solution against neighborhood attacks. In [2], Blosser and Zhan present a client-server model solution on how to combine social networks with privacy goals. In this paper, we will present a cryptographic approach to privacy-preserving collaborative social network construction.

The paper is organized as follows: We present our building blocks in the next section. We define our problem in section 3. Thereafter, we describe protocols for privacy-preserving collaborative social network construction in section 4. We present our conclusions in section 5.

2 Building Blocks

In this paper, we would like to follow the privacy definition proposed in [10]. The basic idea is as follows: A privacy-oriented scheme S preserves data privacy if for any private data T, the following is held:

$$|Pr(T|PPDMS) - Pr(T)| \leq \epsilon$$

where

- $PPDMS$: Privacy-preserving social network construction scheme.
- ϵ: A probability parameter.

- $Pr(T|PPDMS)$: The probability that the private data T is disclosed after a privacy-preserving social network construction scheme has been being applied.
- $Pr(T)$: The probability that the private data T is disclosed without any privacy-preserving social network scheme being applied.
- $Pr(T|PPDMS) - Pr(T)$: The probability that private data T is disclosed with and without privacy-preserving social network schemes being applied.

We call $1-\epsilon$ the privacy level that the privacy-oriented scheme S can achieve. The goal is to make ϵ as small as possible.

We have defined privacy for social network algorithms. However, often times, we need to reduce the complete privacy-preserving social network algorithm to a set of component privacy-oriented protocols. We say a privacy-preserving social network algorithm preserves privacy if each component protocol preserves privacy and the combination of the component protocols do not disclose private data. In the secure multi-party computation literature, a composition theorem [5] describes the similar idea.

Theorem 1. *Suppose that g is privately reducible to f and that there exists a protocol for privately computing f. Then there exists a protocol for privately computing g.*

Proof. Refer to [5].

We now formally define privacy for a component protocol.

Definition 1. *A privacy-oriented component protocol CP preserves data privacy if for any private data T, the following is held:*

$$|Pr(T|CP) - Pr(T)| \leq \epsilon$$

where

- *CP: Component protocol.*
- *$Pr(T|CP)$: The probability that the private data T is disclosed after a privacy-preserving component protocol being applied.*
- *$Pr(T|CP) - Pr(T)$: The probability that private data T is disclosed with and without privacy-preserving component protocol.*

We call $1-\epsilon$ the privacy level that the privacy-oriented component protocol CP can achieve. The goal is to make ϵ as small as possible.

Next, we introduce the fundamental building blocks/component protocols we consider. They are homomorphic encryption and the digital envelope technique.

2.1 Homomorphic Encryption

The concept of homomorphic encryption was originally proposed in [7]. Since then, many such systems have been proposed. In this paper, we base our privacy-oriented protocols on [6] which is semantically secure.

A cryptosystem is homomorphic with respect to some operation $*$ on the message space if there is a corresponding operation $*'$ on the ciphertext space such that $e(m)*'e(m') = e(m*m')$. In our privacy-oriented protocols, we use additive homomorphism offered by [6] in which Paillier proposed a new trapdoor mechanism based on the idea that it is hard to factor number n = pq where p and q are two large prime numbers. In the performance evaluation, Paillier compares the proposed encryption scheme with existing public-key cryptosystems. The results show that the encryption process is comparable with the encryption process of RSA in terms of the computation cost; the decryption process is faster than the decryption process of RSA.

In this paper, we utilize the following property of the homomorphic encryption functions: $e(m_1) \times e(m_2) = e(m_1 + m_2)$ where m_1 and m_2 are the data to be encrypted. Because of the property of associativity, $e(m_1 + m_2 + .. + m_n)$ can be computed as $e(m_1) \times e(m_2) \times \cdots \times e(m_n)$ where $e(m_i) \neq 0$. That is

$$d(e(m_1 + m_2 + \cdots + m_n)) = d(e(m_1) \times e(m_2) \times \cdots \times e(m_n)) \tag{1}$$

Note that a corollary of it is as follows:

$$d(e(m_1)^{m_2}) = d(e(m_1 \times m_2)), \tag{2}$$

where $\times$ denotes multiplication.

2.2 Digital Envelope

A digital envelope [3] is a random number (or a set of random numbers) only known by the owner of private data. To hide the private data in a digital envelope, we conduct a set of mathematical operations between a random number (or a set of random numbers) and the private data. The mathematical operations could be addition, subtraction, multiplication, etc. For example, assume the private data value is v. There is a random number R which is only known by the owner of v. The owner can hide v by adding this random number, e.g., $v + R$.

3 Privacy-Preserving Collaborative Social Network Framework

A social network is a graph, $G = (V, E)$, where V is a set of nodes representing persons and E is a set of edges $(V \times V)$ representing the relationships between the corresponding persons, such as friendship, values, visions, idea, financial exchange, kinship, dislike, conflict, trade, web links, sexual relations, disease transmission (epidemiology), or airline routes. The resulting structures are often very complex. Social network analysis views social relationships in terms of nodes and ties. Nodes are the individual actors within the networks, and ties are the relationships between the actors. There can be many kinds of ties between the nodes. Research in a number of academic fields has shown that social networks operate on many levels, from families up to the level of nations, and play a critical

role in determining the way problems are solved, organizations are run, and the degree to which individuals succeed in achieving their goals. In its simplest form, a social network is a map of all of the relevant ties between the nodes being studied. The network can also be used to determine the social capital of individual actors. These concepts are often displayed in a social network diagram, where nodes are the points and ties are the lines.

Problem 1. Our specific problem setup is as follows: Given two or more social networks from different sources, we need to share the information between these social networks in order to conduct more accurate social network analysis for intelligent investigation. Without information sharing, we only have only a subset of the collaborative social network that can be created by merging the individual social networks of the participants. However, due to privacy issues, parties cannot release its node and/or edge connectivity and weight information from its social network to other participants. Given this problem, how can we construct the complete social network without compromising the data privacy of each collaborator while still attaining a reasonably accurate collaborative social network.

Next, we would like to introduce some notations used in this paper.

Notation 1 – *In the component protocol involving multiple parties, we use* ADV_P *to denote* P_i*'s advantage to gain access to the private data of any other party via the component protocol.*
– $Pr(T_P\,|VIEW_P\,, Protocol\varsigma)$*: the probability that* P_j *sees* P_i*'s private data via protocol* ς*.*
– *We use* ADV_S *to denote the advantage of one party to gain the other party's private data via the component protocol by knowing the semantically secure encryptions. According to Definition,* ADV_S *is negligible.*

3.1 Email Social Network Construction Algorithm

In this paper, we would like to use emails as data sources to construct the social networks. To build up a social network based on emails, we usually need to consider three type of communications, 'To', 'CC', and 'BCC'. The following is the basic algorithm in our email social network systems.

– We ignore any messages with more than 10 recipients. This rule is used to reduce spam and ignore corporate mass mailings that could cause false ties [1].
– We assign varying weights to the edges depending on how many recipients there are and if the email is 'To', 'CC', or 'BCC' the recipient: (a) To 100 weight. (b) CC 25 weight. From our experience the CC has little to contribute to social interactions in business and is usually used to inform superiors of information. (c) BCC 50 weight. The BCC serves the same purpose as a CC usually, but other recipients cannot see the communication, this

means the recipients will not have a weight to add to the total for this tie, thus the higher value. (d) Variance The assigned weight is then divided by the total number of recipients. The recipients of most messages will also have a copy of the message in their email, so this measure prevents messages to many people from being unfairly weighted. For example, a message to 5 people that is not divided would end up giving a total weight of 500 to each person involved, divided each person only receives 100.

– We drop any ties with a total weight less than 500. This removes most spam and attempts to further limit the results to social communication between individuals [1].

Two important tasks in social network construction are (1) computing the total weighted count of communications between people, and (2) comparing the total count with the given threshold to decide whether you should keep the edge between two nodes (persons). To achieve the first step, we will design the following multi-party summation protocol. To achieve the second step, we can apply the multi-party sorting protocol with a sequence of dummy numbers.

3.2 Privacy-Preserving Solution for Collaborative Social Network Construction

Based the social network algorithm, the key issue to construct a social network is to compute the joint results from different parties, and compare with a threshold. We can then decide whether such links should exist or not in the final social network. We would like to develop two protocols to deal with this problem. One is the privacy-preserving summation protocol to securely compute the summation of the contribution from different parties. The other is to securely compare with an existing threshold.

3.3 Privacy-Preserving Multi-party Summation Protocol

Problem 2. Let us assume P_1 has a private integer number $c.count_1$, P_2 has a private integer number $c.count_2$, $\cdots$, and P_n has a private integer number $c.count_n$ where $n \geq 3$. The goal is to compute the $\sum_{i=1}^{n} c.count_i$ without compromising data privacy. One party obtains $\sum_{i=1}^{n} c.count_i$, then shares the result with other parties.

Highlight of Protocol 1: In our protocol, we randomly select a key generator, e.g., P_n who generates a cryptographic key pair (e, d) of a homomorphic encryption scheme and a large integer X which is greater than the total number of records N. P_1 sends P_2 the encryption of the private value which is masked by a digital envelope; P_2 computes the multiplication between the received term and the encryption of the masked private value by another digital envelope; This is repeated until P_n obtains $e(\sum_{i=1}^{n} c.count_i + (\sum_{i=1}^{n} R_i) \times X)$. Finally, P_n obtains $c.count$ by decrypting it, then reducing modulo X.

We present the formal protocol as follows:

Protocol 1

1. P_n generates a cryptographic key pair (e, d) of a semantically secure homomorphic encryption scheme. P_n also generates an integer X which is greater than N.
2. P_1 computes $e(c.count_1 + R_1 \times X)$ and sends it to P_2 where R_1 is a random integer generated by P_1.
3. P_2 computes $e(c.count_1 + R_1 \times X) \times e(c.count_2 + R_2 \times X) = e(c.count_1 + c.count_2 + (R_1 + R_2)X)$ and sends it to P_3. R_2 is a random integer generated by P_2.
4. Repeat until P_n computes $e(c.count_1 + R_1 \times X) \times e(c.count_2 + R_2 \times X) \times \cdots \times e(c.count_n + R_n \times X) = e(\sum_{i=1}^{n} c.count_i + \sum_{i=1}^{n} R_i \times X)$.
5. P_n computes $d(e(\sum_{i=1}^{n} c.count_i + (\sum_{i=1}^{n} R_i) \times X))\ modX = (\sum_{i=1}^{n} c.count_i + (\sum_{i=1}^{n} R_i) \times X)\ modX = \sum_{i=1}^{n} c.count_i$.

The Correctness Analysis of Protocol 1: To show the $c.count$ is correct, we need to consider:

$$d[e(c.count_1) \times e(c.count_2) \times \cdots \times e(c.count_n)]$$

$$= d[e(c.count_1 + R_1 \times X) \times e(c.count_2 + R_2 \times X) \times \cdots \times e(c.count_n + R_n \times X)] modX.$$

According to Equation 2, the left hand side

$$d[e(c.count_1) \times e(c.count_2) \times \cdots \times e(c.count_n)] = \sum_{i=1}^{n} c.count_i.$$

The right hand side

$$d[e(c.count_1 + R_1 \times X) \times e(c.count_2 + R_2 \times X) \times \cdots \times e(c.count_n + R_n \times X)] modX$$

$$= [\sum_{i=1}^{n} c.count_i + \sum_{i=1}^{n} R_i) \times X] modX.$$

Since $X > N$, $\sum_{i=1}^{n} c.count_i \leq N$, and $\sum_{i=1}^{n} R_i$ is an integer,

$$[\sum_{i=1}^{n} c.count_i + (\sum_{i=1}^{n} R_i) \times X] modX = \sum_{i=1}^{n} c.count_i.$$

Therefore, the $\sum_{i=1}^{n} c.count_i$ is correctly computed.

The Complexity Analysis of Protocol 1: The bit-wise communication cost of this protocol is α(n-1) since the cost of each step is α except for the first step.

The following contributes to the computational cost: (1)The generation of one cryptographic key pair. (2)The total number of n encryptions. (3)The total number of $2n-1$ multiplications. (4)One decryption. (5)One modular operation. (6)n additions.

Therefore, the total computation overhead is about $9n$.

Theorem 2. *Protocol 1 preserves data privacy at a level equal to* ADV_P .

Proof. We will identify the value of ϵ such that

$$|Pr(T|CP) - Pr(T)| \leq \epsilon$$

holds for $T = T_P$, $i \in [1, n]$, and CP = Protocol 1.

According to our notations,

$$ADV_P = Pr(T_P\,|VIEW_P\,, Protocol1) - Pr(T_P\,|VIEW_P\,), i \neq n,$$

and

$$ADV_P = Pr(T_P\,|VIEW_P\,, Protocol1) - Pr(T_P\,|VIEW_P\,),$$

where ADV_P is the advantage of P_n to gain access to the other parties private data by obtaining the final result $\sum_{i=1}^{n-1} c.count_i$.

Since P_1 obtains no data from other parties, $ADV_{P_1} = 0$. For $P_2, \cdots, P_{n-1}$, all the information that each of them obtains about the other parties data is encrypted, thus,

$$ADV_P = ADV_S,$$

which is negligible.

In order to show that privacy is preserved according to Definition 1, we need to know the value of the privacy level ϵ. We set

$$\epsilon = max(ADV_P\,, ADV_P\,) = max(ADV_S, ADV_P\,) = ADV_P\,.$$

Then

$$Pr(T_P\,|VIEW_P\,, Protocol1) - Pr(T_P\,|VIEW_P\,) \leq ADV_P\,, i \neq n,$$

and

$$Pr(T_P\,|VIEW_P\,, Protocol1) - Pr(T_P\,|VIEW_P\,) \leq ADV_P\,,$$

which completes the proof.

3.4 Privacy-Preserving Multi-party Sorting Protocol

Problem 3. Assume that P_1 has a private number t_1, P_2 has a private number $t_2, \cdots$, and P_n has a private number t_n. The goal is to sort t_i, $i \in [1, n]$ without disclosing t_i to P_j where $i \neq j$.

Highlight of Protocol 2: In our protocol, we randomly select a key generator, e.g., P_n, who generates a cryptographic key pair (e, d) of a homomorphic encryption scheme. Each party encrypts their number using e, then sends it to P_{n-1}. P_{n-1} computes the encryption difference of two numbers and obtains a sequence φ of n^2 elements. P_{n-1} randomly permutes this sequence and sends the permuted sequence to P_n who decrypts each element in the permuted sequence and obtains a $+1/-1$ sequence according to the decrypted results. P_n sends this $+1/-1$ sequence to P_{n-1} who determines the sorting result.

We present the formal protocol as follows:

Protocol 2

1. *P_n generates a cryptographic key pair (e, d) of a semantically secure homomorphic encryption scheme.*
2. *P_i computes $e(t_i)$, for $i = 1, 2, \cdots, n-2, n$, and sends it to P_{n-1}.*
3. *P_{n-1} computes $e(t_i) \times e(t_j)^{-1} = e(t_i - t_j)$ for all $i, j \in [1, n], i < j$, and sends the sequence denoted by φ, which is randomly permuted, to P_n.*
4. *P_n decrypts each element in the sequence φ. He assigns the element $+1$ if the result of decryption is not less than 0, and -1, otherwise. Finally, he obtains a $+1/-1$ sequence denoted by φ'.*
5. *P_n sends the $+1/-1$ sequence φ' to P_{n-1}.*
6. *P_{n-1} sorts the numbers t_i, $i \in [1, n]$.*

The Correctness Analysis of Protocol 2: P_{n-1} is able to remove permutation effects from φ' (the resultant sequence is denoted by φ'') since she has the permutation function that she used to permute φ, so that the elements in φ and φ'' have the same order. It means that if the qth position in sequence φ denotes $e(t_i - t_j)$, then the qth position in sequence φ'' denotes the result of $t_i - t_j$. We encode it as $+1$ if $t_i \geq t_j$, and as -1 otherwise. P_{n-1} has two sequences: one is φ, the sequence of $e(t_i - t_j)$, for $i, j \in [1, n](i > j)$, and the other is φ'', the sequence of $+1/-1$. The two sequences have the same number of elements. P_{n-1} knows whether or not t_i is larger than t_j by checking the corresponding value in the φ'' sequence. For example, if the first element φ'' is -1, P_{n-1} concludes $t_i < t_j$. P_{n-1} examines the two sequences and constructs the index table (Table 3.1) to sort t_i, $i \in [1, n]$.

In Table 3.1, $+1$ in entry ij indicates that the value of the row (e.g., t_i of the ith row) is not less than the value of a column (e.g., t_j of the jth column); -1, otherwise. P_{n-1} sums the index values of each row and uses this number as the weight of that row. She then sorts the sequence according the weight.

To make it clearer, let us illustrate it using an example. Assume that: (1) there are 4 elements with $t_1 < t_4 < t_2 < t_3$; (2) the sequence φ is $[e(t_1 -$

Table 1. An Index Table of Number Sorting

	t_1	t_2	t_3	$\cdots$	t_n
t_1	+1	+1	-1	$\cdots$	-1
t_2	-1	+1	-1	$\cdots$	-1
t_3	+1	+1	+1	$\cdots$	+1
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
t_n	+1	+1	-1	$\cdots$	+1

$t_2), e(t_1 - t_3), e(t_1 - t_4), e(t_2 - t_3), e(t_2 - t_4), e(t_3 - t_4)]$. The sequence φ'' will be $[-1, -1, -1, -1, +1, +1]$. According to φ and φ'', P_{n-1} builds the Table 3.2. From the table, P_{n-1} knows $t_3 > t_2 > t_4 > t_1$ since t_3 has the largest weight, t_2 has the second largest weight, t_4 has the third largest weight, t_1 has the smallest weight.

The Complexity Analysis of Protocol 2: The total communication cost is (1) The cost of $\alpha(n-1)$ from step 2. (2) The cost of $\frac{1}{2}\alpha n^2$ from step 3. (3) The cost of $\frac{1}{2}\beta n^2$ from step 4 where β denotes the number of bits for $+1$ and -1. Note that normally $\beta \ll \alpha$ (4) The cost of $\frac{1}{2}\beta n^2$ from step 5. Therefore, the total communication overhead is upper bounded by $\frac{3}{2}\alpha n^2 + \alpha(n-1)$.

The following contributes to the computational cost: (1)The generation of one cryptographic key pair. (2) The total number of n encryptions. (3)The total number of n^2 multiplications. (4) The total number of n^2 decryptions. (5)The total number of n^2 assignments. (6) $n^2 - n$ additions. (7)$g_3 nlog(n)$ for sorting n numbers.

Therefore, the total computation overhead is $g_1 + 6n + n^2 + 13n^2 + n^2 + n^2 - n + g_3 nlong(n) = 16n^2 + 5n + g_3 nlog(n) + g_1$.

Table 2. An Example of Sorting

	t_1	t_2	t_3	t_4	Weight
t_1	+1	-1	-1	-1	-2
t_2	+1	+1	-1	+1	+2
t_3	+1	+1	+1	+1	+4
t_4	+1	-1	-1	+1	0

Theorem 3. *Protocol 2 preserves data privacy at a level equal to* ADV_P .

Proof. We will identify the value of ϵ such that

$$|Pr(T|CP) - Pr(T)| \leq \epsilon$$

holds for $T = T_P$, $i \in [1, n]$, and CP = Protocol 2.

According to our notations,

$$ADV_{P_{n-1}} = Pr(T_P|View_{P_{n-1}}, Protocol 2) - Pr(T_P|View_{P_{n-1}}), i \neq n-1,$$

and

$$ADV_{P_j} = Pr(T_P|View_{P_j}, Protocol 2) - Pr(T_P|View_{P_j}), j \neq n.$$

All the information that P_{n-1} obtains from other parties is $e(t_i)$ for $1 \leq i \leq n$, $i \neq n-1$, and the sequence φ'.

Since e is semantically secure,

$$ADV_{P_{n-1}} = ADV_S,$$

which is negligible.

In order to show that privacy is preserved according to Definition 1, we need to know the value of the privacy level ϵ. We set

$$\epsilon = max(ADV_{P_i}, ADV_{P_{n-1}}) = max(ADV_{P_i}, ADV_S) = ADV_{P_i}.$$

Then

$$Pr(T_P|View_{P_{n-1}}, Protocol 2) - Pr(T_P|View_{P_{n-1}}) \leq ADV_{P_i}, i \neq n-1,$$

and

$$Pr(T_P|View_{P_j}, Protocol 2) - Pr(T_P|View_{P_j}) \leq ADV_{P_i}, j \neq n.$$

which completes the proof.

4 Discussion

Privacy-preserving social networks is an important area which is beginning to receive attention from scouter scientists. In this paper, we consider network construction and provide a solution based on homomorphic encryption. Our approach has wide potential impact in many applications. In practice, there are many environments where privacy-preserving collaborative social networks are desirable. For example, the success of homeland security aiming to counter terrorism depends on combination of social networks across different mission areas, effective international collaboration and information sharing to support a coalition in which different organizations and nations must share some, but not all, information. Information privacy, thus, becomes extremely important and our technique can be applied. We would like to mention that our solution can deal with any number of participants and the scalability issue for the privacy-preserving collaborative protocols has been addressed. For future work, we plan to show efficiency of our approach on large data sets, specifically, an email data set. Also, we want to extend this work to consider networks with varying characteristics.

References

1. Adamic, L., Adar, E.: How to search a social network. Social Networks 27(3), 187–203 (2005)
2. Blosser, G., Zhan, J.: Privacy-preserving social networks. In: Proceedings of IEEE International Conference on Information Security and Assurance (ISA 2008), Busan, Korea, April 24-26 (2008)
3. Chaum, D.: Security without identification. Communication of the ACM 28(10), 1030–1044 (1985)
4. Epic. Privacy and human rights an international survey of privacy laws and developments. Electronic Privacy Information Center (May 2003), http://www.epic.org
5. Goldreich, O.: The Foundations of Cryptography. Cambridge University Press, Cambridge (2004)
6. Paillier, P.: Public-key cryptosystems based on composite degree residuosity classes. In: Stern, J. (ed.) EUROCRYPT 1999. LNCS, vol. 1592, pp. 223–238. Springer, Heidelberg (1999)
7. Rivest, R., Adleman, L., Dertouzos, M.: On data banks and privacy homomorphisms. In: DeMillo, R.A., et al. (eds.) Foundations of Secure Computation, pp. 169–179. Academic Press, London (1978)
8. Singh, L., Zhan, J.: Measuring topological anonymity in social networks. In: Proceedings of IEEE International Conference on Granular Computing, Silicon Valley, USA, November 2-4 (2007)
9. Wang, D., Liau, C., Hsu, T.: Privacy protection in social network data disclosure based on granular computing. In: IEEE International Conference on Fuzzy Systems, Vancouver, BC, Canada, July 16-21 (2006)
10. Zhan, Z.: Privacy Preserving Collaborative Data Mining. PhD thesis, University of Ottawa (2006)
11. Zhou, B., Pei, J.: Preserving privacy in social networks against neighborhood attacks. In: Proceedings of the 24th International Conference on Data Engineering (ICDE 2008), Cancún, México, April 7-12 (2008)

Efficient Secret Authenticatable Anonymous Signcryption Scheme with Identity Privacy

Mingwu Zhang[1], Bo Yang[1], Shenglin Zhu[1], and Wenzheng Zhang[2]

[1] Department of Computer Science and Engineering, College of Informatics, South China Agricultural University, Guangzhou, 510642, P.R. China
{zhangmw,byang,zhusl}@scau.edu.cn
[2] National Laboratory for Modern Communications, Chengdu, 610041, P.R. China
wzzhang@163.com

Abstract. Three of the most important services offered by cryptography are confidential, private and authenticatability in distributed systems, such as P2P, trust negotiation and decentralized trust management. Information provider secretly encrypts a message with a hidden approach to protect message security and his privacy. Ring signcryption is an important way to realize full anonymity where the signcrypter cannot verify that this ciphertext was produced by himself. In this paper, we propose a novel construction of efficient secret authenticatable anonymous signcryption scheme in which only the actual signcrypter can authenticate that the ciphertext was produced by himself. The receiver cannot distinguish who the actual signcrypter is in the group even though he obtains all group members' private keys. Proposed scheme has the following properties: semantic security, signcrypter anonymity, signcrypter secret authenticatability, and unforgeability. We prove its security in the random oracle model under the DBDH assumption.

Keywords: signcryption, unforgeability, anonymity, ring signature.

1 Introduction

Signcryption, first proposed by Zheng [15], is a cryptographic primitive that performs signature and encryption simultaneously, at a lower computational costs and communication overheads than the traditional systems like PGP that executes signing and encryptin a message in sequential procedures. ID-based cryptography is supposed to provide a more convenient alternative to conventional public key infrastructure that was introduced by Shamir [1] in 1984. A distinguishing property of ID-based cryptography is that a user's public key can be any binary string that can identify the user's identity, while private keys can be generated by the trusted Private Key Generator(PKG).

Ring signature, first formalized by Rivest et al [11], is an anonymous approach which is useful in cases where the identity of sender must remain secret, yet the message should be verifiable. The receiver only knows that the message is produced by some member of this group, but he cannot know more information

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 126–137, 2008.

about actual signer. Ring signcryption, which is a combination of anonymous signature and encryption, provides a signcypter ambiguity, even if a member of the group who produces a ciphertext, he also cannot prove that the ciphertext was produced by himself. There exists another problem, for example, one can expose the grafter to the police using ring signcryption manner to encrypt and sign the evidence in order to avoid being retaliated. The supervisor bureau will give the exposer prize in order to advocate the exposure behaviors. When the exposer wants to give a proof that he is the exposer, the supervisor bureau cannot distinguish who is the actual exposer from the group members, and the exposer also cannot provide the enough information to prove it because of the anonymity of ring signcryption. Furthermore, the exposer don't expect the others know he is the exposer. To solve the signcrypter secret authenticatable issue, we propose a ID-based authenticatable anonymous signcryption scheme.

Naor [7] proposed a deniable ring authentication, in which a prover can confirm a verifier that the prover does or does not authenticate a message anonymously. This scheme needs only the third party (PKI). Gao et al. [13] proposed a controllable ring signature with anonymous identification protocol based on public key system. Komano et al. [16] introduced a concept of a deniable ring signature scheme(DRS) that the signer interacted with the verifier to confirm that the signer had generated the signature or not with a zero knowledge interactive proof(ZKIP). In [4], an ID-based ring signature scheme with self-authentication is constructed using Boneh's BLS scheme, which can authenciate that a member possesses a signature's ownership. All above schemes are based on signature scheme or PKI framework.

In this paper, we propose an ID-based secret authenticatable anonymous signcryption scheme. In our scheme, a message can be signcrypted secretly by ID_A that the receiver ID_B can authenticate the ciphertext generated from a member of the group where he cannot identify the actual signcrypter, but the actual signcrypter ID_A can prove that the ciphertext is generated by himself, and the others cannot authenticate it. We also prove its security in a formal model under recently studied computational assumptions in the random oracle model. Specifically, compared with recent literatures, our scheme provides higher efficiency and security than the others with the same order of ciphertext size.

The rest of this paper is organized as follows: Section 2 gives a formal ID-based verifiable anonymous signcryption and its security notions. We describe our proposed scheme in section 3 and prove its security in section 4. We give the performance and security comparison with recent literatures in section 5 and draw our conclusion in section 6.

2 Formal Model of Our Proposed Scheme

2.1 Secret Authenticatable Ring Signcryption Scheme

Proposed scheme consists of the following five algorithms.

– Setup: Take an input 1^k, where k is a security parameter, the algorithm generates a master key s and the system's public parameters $params$.

- Extract: Given an identity ID received from a user, and system master key s, this algorithm outputs the private key associated with the ID, denoted by D_{ID}.
- Signcrypt: If Alice identified by ID_A wishes to send a message m to Bob identified by ID_B ,this algorithm selects a group of n users' identities by $\bigcup ID_{i(1\leq i\leq n)}$ including the actual signcrypter ID_A , and outputs the ciphertext C.
- Unsigncrypt: When Bob receives the cipertext C, this algorithm takes the C, $\bigcup ID_i$, and Bob's private key D_B as input, and outputs plaintext m when unsigncryption is successful, otherwise it outputs $\perp$.
- Authenticate: If the message signcrypter Alice identified by ID_A wants to prove that the ciphertext is produced by herself, this algorithm outputs whether ID_A is the actual signer of ciphertext C or not.

The above algorithms have the following consistency requirements.

- *Signcrypt/Unsigncrypt*: These algorithms must satisfy the standard consistency constraint of ID-based signcryption scheme, i.e.
  $C = Signcrypt(m, \bigcup ID_i, D_A, ID_B) \Rightarrow Unsigncrypt(C, \bigcup ID_i, D_B) = m$
- *Signcrypt/Authencicate*: Only the actual signcrypter ID_A can authenticate that a ciphertext was produced by himself, i.e.
  $C = Signcrypt(m, \bigcup ID_i, D_A, ID_B) \ \Rightarrow Authenticate(C, D_A) = \top$

2.2 Security Notions

The security of our proposed the ID-based verifiable anonymous signcryption scheme should satisfy *semantics security*, $unforgeability$, *signcrypter anonymity* and *signcrypter authenticatability*.

The accepted notion of security with respect to confidentiality is *indistinguishability of signcrypters under adaptive chosen ciphtertext attack*. We define the notion via a game(IDVSC game) played by a challenger $\mathcal{C}$ and an adversary $\mathcal{A}$ as bellow:

- Initial: The challenger $\mathcal{C}$ runs the $\mathbf{Setup}(1^k)$ algorithm and gives the resulting *params* to the adversary and keeps master key s.
- Phase 1: The adversary $\mathcal{A}$ performs a polynomially bounded number of queries. These queries may be made adaptively, i.e. each query may depend on the answers to the previous queries.
  - Extraction queries: $\mathcal{A}$ produces an identity ID and receives the extracted private key $D_{ID} = \mathbf{Extract}(ID)$.
  - Signcryption queries: $\mathcal{A}$ chooses a group of n identities ID_i (i=1,2,...,n), a plaintext m and a designated message receiver ID_B. $\mathcal{C}$ randomly chooses a user $u_s \in \{ID_i\}$, computes $D_s = \mathbf{Extract}(ID_s)$ and acts as u_s to generate ciphertext $C = \mathbf{Signcrypt}(m, \bigcup ID_i, D_s, ID_B)$ and sends C to $\mathcal{A}$.
  - Unsigncryption queries: $\mathcal{A}$ chooses a group of n identities ID_i (i=1,2...,n), identity ID_r and a ciphertext C. $\mathcal{C}$ generates the privacy key $D_r = \mathbf{Extract}(ID_r)$ and sends the result of $\mathbf{Unsigncrypt}(\{ID_i\}, C, D_r)$ to $\mathcal{A}$.

  - Authenticate queries: $\mathcal{A}$ chooses a member of group $\bigcup\{\mathcal{U}_i\}$ and a ciphertext C. $\mathcal{C}$ performs the authentication algorithm and sends the result to $\mathcal{A}$.
- Challenge: $\mathcal{A}$ chooses two plaintexts $m_0, m_1 \in \mathcal{M}$, a group of identities ID_i^*(for i=1 to n), and a designated receiver ID^* on which he wishes to be challenged. The challenger $\mathcal{C}$ picks a random b from $\{0,1\}$ and computes $C^* = \mathbf{Signcrypt}(m_b, \bigcup {ID_i}^*, ID^*)$ and sends C^* to $\mathcal{A}$.
- Phase 2: The adversary $\mathcal{A}$ can ask a polynomially bounded number of queries adaptively again as in the first stage with the restriction that he cannot make the key extraction query on group member ID_i^* nor ID^*, cannot make the authentication query on ID_i^* and cannot make the unsigncryption query on C^*.
- Response: Finally, $\mathcal{A}$ returns a bit b' and wins the game if $b' = b$.

Definition 1. *(*Indistinguishability*) An ID-based verifiable signcryption scheme has the indistinguishability against adaptive chosen ciphertext attacks property (IND-IDVSC-CCA) if no polynomially bounded adversary has a non-negligible advantage in IDVSC game. $\mathcal{A}$'s advantage is defined as*

$$Adv(\mathcal{A}) = \left|Pr[b' = b] - \tfrac{1}{2}\right| = \epsilon$$

Definition 2. *(*Unforgeability*) An ID-based anonymous signcryption scheme is existentially unforgeable against adaptive chosen-message attacks and adaptive chosen-identity attacks(EUF-IDVSC-CMIA) if no polynomially bounded adversary has a non-negligible advantage in the following game:*
- *The challenger $\mathcal{C}$ runs the* setup *algorithm with a security parameter k and gives the public parameters to adversary $\mathcal{A}$.*
- *$\mathcal{A}$ performs a polynomially bounded phase 1 queries in IDVSC game.*
- *Finally, $\mathcal{A}$ outputs a cipertext C^* that was not produced by signcrypt oracle, an identity ID^* and a group of n identities $\bigcup ID_i^*$ that were not performed key extract queries, and it wins the game if the result of the $Unsigncrypt(C^*, \bigcup {ID_i}^*, ID^*)$ is not the $\perp$ symbol.*

Definition 3. *(*Anonymity*) An ID-based verifiable anonymous signcryption scheme is unconditional anonymous if for any group of n members with identities $\bigcup ID_i (1 \leq i \leq n)$, any adversary cannot identify the actual signcrypter with probability better than random guess's.*

Definition 4. *(*Secret authenticatability*) An ID-based verifiable anonymous signcryption scheme is secret authenticatable if and only if the actual signcrypter can authenticate that the ciphertext was indeed produced by himself, where the other members of the group cannot authenticate the ciphertext with non-negligible probability.*

That is, the actual signcrypter ID_s can authenticate the ciphertext secretly where the other member ID_a cannot authenticate successfully in the name of signcrypter ID_s or himself ID_a.

3 Proposed Scheme

In this section, we describe our proposed ID-based verifiable anonymous signcryption scheme. Our proposed anonymous signcryption scheme is motivated from the ID-based ring signature in [10].

1. Setup: Given a security k and l, a trusted key generator(PKG) selects a pairing $\hat{e} : \mathbb{G}_1 \times \mathbb{G}_1 \rightarrow \mathbb{G}_2$ where the order of $\mathbb{G}_1$ and $\mathbb{G}_2$ is p. Let P be a generator of G_1. Randomly chooses $s \in Z_q^*$ as the master key of PKG and computes $P_{pub} = sP$ as the corresponding public key. Next, PKG chooses some security hash functions: $H_0 : \{0,1\}^* \rightarrow \mathbb{G}_1^*$, $H_1 : \mathbb{G}_2 \rightarrow \{0,1\}^l$, $H_2 : \{0,1\}^* \rightarrow Z_q^*$. The system public parameters $params = \{G_1, G_2, \hat{e}, P, P_{pub}, H_0, H_1, H_2\}$.
2. Extract: For a user $\mathcal{U}_i$ identified by ID_i, PKG computes user public key $Q_i = H_0(ID_i)$ and corresponding secret key $D_i = sQ_i$ where s is the PKG's master key. Then PKG sends D_i to $\mathcal{U}_i$ via a secure and authenticated channel.
3. Signcrypt: Let $\bigcup\{\mathcal{U}_i\}$ (i=1,...,n) be a set of users including the actual signcrypter Alice. To signcrypt a message m on behalf of the group $\bigcup\{\mathcal{U}_i\}$ to receiver Bob(identified by ID_B, public key $Q_B = H_0(ID_B)$), the actual signcrypter Alice, indexed by s(i.e. her public/private key is (Q_s, D_s)), carries out as follows:
   - Chooses $r \in_R Z_q^*$, and computes $R = rP, R' = \hat{e}(P_{pub}, Q_B)^r, t = H_1(R')$, $c = m \oplus t$.
   - For $i = 1, ..., s-1, s+1, ..., n$, chooses $a_i \in_R Z_q^*$ to compute $U_i = a_i P$, and computes $h_i = H_2(m, \bigcup\{\mathcal{U}_i\}, t, U_i)$.
   - Chooses $a_s \in_R Z_q^*$, computes $U_s = a_s Q_s - \sum_{i \neq s}\{U_i + h_i Q_i\}$.
   - Computes $h_s = H_2(m, \bigcup\{\mathcal{U}_i\}, t, U_s)$ and $\sigma = (h_s + a_s)D_s$.
   - Finally, outputs the ciphertext of message m as $C = (\bigcup\{\mathcal{U}_i\}, c, R, h_1, h_2, ..., h_n, U_1, U_2, ..., U_n, \sigma)$.
4. Unsigncrypt: Upon receiving the cipertext $C = (\bigcup\{\mathcal{U}_i\}, c, R, h_1, h_2, ..., h_n, U_1, U_2, ..., U_n, \sigma)$, Bob uses his secret key D_B to unsigncrypt the ciphertext as follows:
   - Computes $t' = H_1(\hat{e}(R, D_B))$, and $m' = c \oplus t'$.
   - For i=1 to n, checks whether $h_i = H_2(m', \bigcup\{\mathcal{U}_i\}, t', U_i)$.
   - Checks whether $\hat{e}(P_{pub}, \sum_{i=1}^{n}(U_i + h_i Q_i)) = \hat{e}(P, \sigma)$.

   If for all $i \in \{1, 2, ..., n\}, h_i = H_2(m', \bigcup\{\mathcal{U}_i\}, t', U_i)$ and $\hat{e}(P_{pub}, \sum_{i=1}^{n}(U_i + h_i Q_i)) = \hat{e}(P, \sigma)$ hold, it is a valid message m for Bob. Otherwise it outputs $\perp$ for invalid ciphertext.
5. Authenticate: The actual signcrypter ID_s wants to give the verifier a proof that the ciphertext C was indeed produced by himself. It uses an interactive zero-knowledge proof to authenticate the ciphertext's producer as follows:
   - First, ID_s chooses $x \in_R Z_q^*$, and computes $\mu = \hat{e}(P, \sigma)^x$, and sends μ to the verifier.
   - The verifier chooses $y \in_R Z_q^*$ and sends it to ID_s.
   - ID_s computes $\nu = (x + y)(h_s + a_s)$ and returns ν to the verifier.
   - Finally, the verifier checks $\hat{e}(P_{pub}, Q_s)^\nu = \mu \cdot \hat{e}(P, \sigma)^y$. If the above equality holds, the verifier shows that the ID_s is actual signcrypter and returns $\top$, otherwise it returns $\perp$.

4 Correctness and Security Analyzes

4.1 Correctness

– Unsigncryption correctness: If the ciphertext C is generated in the way described as above algorithms, the following equations will hold.
$t' = H_1(\hat{e}(R, D_B)) = H_1(\hat{e}(rP, D_B)) = H_1(\hat{e}(sP, Q_B)^r)$
$= H_1(\hat{e}(P_{pub}, Q_B)^r) = t$
Hence, $m' = c \oplus t' = c \oplus t = m$, and for all $i \in \{1, 2, ..., n\}, h_i = H_2(m', \bigcup\{\mathcal{U}_i\}, t', U_i)$ hold.
Furthermore, $\hat{e}(P, \sigma) = \hat{e}(P, (h_s + a_s)D_s) = \hat{e}(P_{pub}, (h_s + a_s)Q_s) = e(P_{pub}, h_sQ_s + (U_s + \sum_{i \neq s}(U_i + h_iQ_i))) = \hat{e}(P_{pub}, \sum_{i=1}^{n}(U_i + h_iQ_i))$
– Authentication correctness: The verifier uses the interactive value μ, ν, y to check whether the equality $\hat{e}(P_{pub}, Q_s)^\nu = \mu \cdot \hat{e}(P, \sigma)^y$ holds.
$\hat{e}(P_{pub}, Q_s)^\nu = \hat{e}(P_{pub}, (h_s + a_s)Q_s)^{x+y} = \hat{e}(P, (h_s + a_s)D_s)^{x+y}$
$= \hat{e}(P, \sigma)^{x+y} = \mu \cdot \hat{e}(P, \sigma)^y$

4.2 Security

Theorem 1. *(*Indistinguishability*) In the random oracle model, if there is an adaptive chosen ciphertext attack adversary $\mathcal{A}(t, q_{H0}, q_{H1}, q_{H2}, q_S, q_U, q_A, \epsilon)$ who can distinguish ciphertexts from the users set $\bigcup\{\mathcal{U}_i\}$ during the IDVSC game with an advantage ϵ when running in a time t and makes q_H queries to $H_i (0 \leq i \leq 2)$, at most q_E key extract queries, q_S signcryption queries, q_U unsigncryption queries and q_A authentication queries. Then there exists another algorithm $\mathcal{B}(t', \epsilon')$ that can solve a random instance of the DBDH problem in running time $t' = O(t)$ with an advantage*

$$\epsilon' \geq \frac{1}{e^{+}} \cdot \frac{|\epsilon - q_{0}/2|}{q_{0}}$$

where k is system security parameter and n is the number of group users.

Proof. Let the distinguisher $\mathcal{B}$ receives a random instance (P, aP, bP, cP, h) of the DBDHP. His goal is to decide whether $h = \hat{e}(P, P)^{abc}$ or not. In order to solve this problem, $\mathcal{B}$ will run $\mathcal{A}$ as a subroutine and act as $\mathcal{A}$'s challenger in the *IDVSC* game. We assume that $\mathcal{A}$ will ask for $H_0(ID)$ before ID is used in any other queries. We also assume that $\mathcal{A}$ never makes an unsigncryption query on a ciphertext obtained from the signcryption oracle, and he can only make unsigncryption queries for observed or guessed ciphertext.

Setup: At first, $\mathcal{B}$ sets $P_{pub} = cP$ as system public key and sends the public *params*$= \{\mathbb{G}_1, \mathbb{G}_2, \hat{e}, P, P_{pub}, H_0, H_1, H_2\}$ to $\mathcal{A}$ after running the Setup with the parameter k. The value c is unknown to $\mathcal{B}$ and is used as the role of the PKG's master key.

Queries: For the identities extraction and the signcryption/unsigncryption on the message m, $\mathcal{B}$ simulates the hash oracles(H_0, H_1, H_2), extraction oracle, signcryption oracle, unsigncryption oracle and authentication oralce. $\mathcal{A}$ can perform

its queries adaptively in which every query may depend on the answers to the previous ones.

H_0 queries. To response these queries, $\mathcal{B}$ maintains the list L_0 of tuples (ID, Q_{ID}, b, c). When $\mathcal{A}$ queries the oracle H_0, $\mathcal{B}$ responds as follows:

- At the j^{th} query, $\mathcal{B}$ answers by $H_0(ID_j) = bP$, and let $c_j = 0$.
- For $i \neq j$, if ID_i already appears on the L_0, then $\mathcal{B}$ responds with $H_0(ID_i) = Q_i$. Otherwise, $\mathcal{B}$ generates a random $c_i \in \{0, 1\}$ that $c_i = 1$ with probability ς and $c_i = 0$ with probability $1 - \varsigma$.
- $\mathcal{B}$ picks a random $b_i \in Z_q^*$ to compute $Q_i = b_i P$.
- $\mathcal{B}$ adds the tuple (ID_i, Q_i, b_i, c_i) to the list L_0 and responds to $\mathcal{A}$ with $H_0(ID_i) = Q_i$.

H_1 queries. $\mathcal{B}$ responds as follows:

- If $(R_i, t_i) \in L_1$ for some t_i, returns t_i.
- Else randomly chooses $t_i \leftarrow \{0, 1\}^l$, and adds the pair (R_i, t_i) to L_1; returns t_i.

H_2 queries. $\mathcal{B}$ responds as follows:

- If $(m_i || \bigcup\{\mathcal{U}_i\} || t_i || U_i, h_i) \in L_2$ for some h_i, returns h_i.
- Else randomly chooses $h_i \leftarrow Z_q^*$, and adds the pair $(m_i || \bigcup\{\mathcal{U}_i\} || t_i || U_i, h_i)$ to L_2; returns h_1.

Key extraction queries. When $\mathcal{A}$ asks a query **Extract**(ID_i), $\mathcal{B}$ first finds the corresponding tuple (ID_i, Q_i, b_i, c_i) in L_0. If $c_i = 0$, $\mathcal{B}$ fails and stops. Otherwise, $\mathcal{B}$ computes the secret key $D_i = b_i P_{pub} = cQ_i$, then $\mathcal{B}$ returns D_i to $\mathcal{A}$.

Signcryption queries. $\mathcal{A}$ can perform a signcryption queries for a plaintext m, a user group $\bigcup\{\mathcal{U}_i\}$ and a designated receiver with identity ID.

- $\mathcal{B}$ randomly chooses a user $u_A \in \bigcup\{\mathcal{U}_i\}$ whose identity is $ID_A (ID_A \neq ID_j)$ where $\mathcal{B}$ can compute u_A's secret key $D_A = b_A P_{pub}$ where b_A is in the corresponding tuple (ID_A, Q_A, b_A, c_A) in the list L_0.
- Then $\mathcal{B}$ runs **Signcrypt**(m,$\bigcup\{\mathcal{U}_i\}, ID_A, ID$) to signcrypt a message m on behalf the group $\bigcup\{\mathcal{U}_i\}$ using u_A's private key.
- At last, $\mathcal{B}$ returns the result C to $\mathcal{A}$.

Unsigncryption queries. At any time, $\mathcal{A}$ can perform an unsigncryption query for a ciphertext $C = (\bigcup\{\mathcal{U}_i\}, c, R, h_1, ..., h_n, U_1, ..., U_n, \sigma)$ between the group $\bigcup\{\mathcal{U}_i\}$ and the receiver with identity ID.

- If $ID = ID_j$, $\mathcal{B}$ always notifies $\mathcal{A}$ that the ciphertext is invalid, because $\mathcal{B}$ does not know ID_j's secret key. If this ciphertext is a valid one, the probability that $\mathcal{A}$ will find is no more than 2^{-k}.
- If $ID \neq ID_j$, $\mathcal{B}$ computes $t' = H_1(\hat{e}(R, D_{ID}))$, $m' = m \oplus t'$.
- If for all $i \in \{1, 2, ..., n\}$, $h_i = H_2(m', \bigcup\{\mathcal{U}_i\}, t', U_i)$, and $\hat{e}(P_{pub}, \sum_{i=1}^{n}(U_i + h_i Q_i)) = \hat{e}(P, \sigma)$ hold, $\mathcal{B}$ notifies $\mathcal{A}$ that the ciphertext is a valid one. Otherwise, $\mathcal{B}$ notifies $\mathcal{A}$ that ciphertext C is not a valid ciphertext with symbol $\perp$.

Authentication queries. We omit the authentication queries because it cannot provide any ciphertext information according **corollary 1.**

Challenge: After a polynomially bounded number of queries, $\mathcal{A}$ chooses two message $m_0, m_1 \in \mathcal{M}$, n users whose identities are $\{ID_1, ID_2, ..., ID_n\}$ to form a ring $\mathcal{U}$ and another user whose identity is ID. If $ID \neq ID_j$, $\mathcal{B}$ fails and stops. For $\forall i \in \{1, 2, ..., n\}$, if $c_i = 1$ in the corresponding tuple (ID_i, Q_i, b_i, c_i) in L_0, $\mathcal{B}$ also fails and stops. If such $\mathcal{U}$ and the receiver are admissable, $\mathcal{B}$ chooses $b \in_R \{0, 1\}$ and let $R = aP, R' = h$, then $\mathcal{B}$ signcrypts the message m_b as described in the signcryption request and sends the ciphertext C to $\mathcal{A}$.

$\mathcal{A}$ asks a polynomially bounded number of queries just like in the first stage. In this stage, he cannot request the secret key of any user in the group $\mathcal{U}$ nor ID_j, and he cannot ask the plaintext corresponding to the ciphertext C. At the end of the simulation, he produces a bit b' for which he believes the relation C=**Signcrypt**$(m_b, \mathcal{U}, ID_j)$ holds and sends b' to $\mathcal{B}$. At this moment, if $b' = b$, $\mathcal{B}$ answers 1 as a result because his selection h allowed him to produce a ciphertext C that appeared to $\mathcal{A}$ as a valid signcrypted plaintext of m_b. If $b' \neq b$, $\mathcal{B}$ answers 0.

Success probability: Now we analyze $\mathcal{B}$'s success probability.

- The probability that $\mathcal{B}$ does not fail during the key extraction queries is ς^q where q_E is the number of key extraction queries.
- The probability that $\mathcal{B}$ does not fail during the challenge phase is $(1 - \varsigma)^n / q_{H_0}$.

Therefore, the probability that $\mathcal{B}$ does not fail during the simulation is $\varsigma^q \ (1 - \varsigma)^n / q_{H_0}$. The value is maximized at $\varsigma' = q_E/(q_E + n)$, and the probability that $\mathcal{B}$ does not abort is at least $(1/q_{H_0})(1/e)^{n+q}$ during the simulation phase. The probability that $\mathcal{B}$ gives a false answer during the unsigncryption process is no more than $q_U/2^k$. Finally,

$$\begin{aligned}\epsilon' &= |P_{a,b,c \in Z\ , h \in G_2}[1 \leftarrow \mathcal{B}(aP, bP, cP, h)] \\ &\quad - P_{a,b,c \in Z}\ [1 \leftarrow \mathcal{B}(aP, bP, cP, \hat{e}(P,P)^{abc})]| \\ &\geq \frac{|p_1 - p_0|}{q_{\ 0} e^{\ +}} = \frac{|\epsilon - q_{\ /2\ }|}{q_{\ 0} e^{\ +}} = \frac{1}{e^{\ +}} \cdot \frac{|\epsilon - q_{\ /2\ }|}{q_{\ 0}}\end{aligned}$$

Theorem 2. *(*Secret authenticatability*) The proposed scheme can realize secret authenticatability only by actual signcrypter.*

proof. Signcrypter authenticatability: If a member with ID_s produces a ciphertext C, obviously, he can authenticate the ownership of the ciphertext by authentication algorithm.

Secret authentication: This means that only actual signcrypter ID_s can authenticate the ciphertext. If an adversary ID_a can pass the authentication algorithm with interactive value μ', y', ν' where $\mu' = \hat{e}(P, \sigma)^{x'}$, then

$\hat{e}(P_{pub}, Q_a)^{\nu'} = \mu' \hat{e}(P, \sigma)^{y'}$

We can obtain

$\hat{e}(P_{pub}, Q_a)^{\overline{\ '+\ '}} = \hat{e}(P, \sigma) = \hat{e}(P_{pub}, Q_s)^{(h\ +a\ )}$

We let $z = \frac{\nu'}{(x'+y')(h\ +a\ )}$. Because a_s is randomly picked by signcrypter ID_s, it is a discrete logarithm problem for $Q_s = zQ_a$ to find z. Obviously, the adversary $ID_a(ID_a \neq ID_s)$ cannot authenticate the ciphertext and only ID_s can obtain *true* for the authentication algorithm.

Non-public authentication: If another member $ID_a(ID_a \neq ID_s)$, who uses interactive value μ', y', ν', can authenticate the ownership of a ciphertext in the name of the member ID_s. Because ID_a don't know the value a_s which is randomly picked by signcrypter ID_s, the adversary can only guess the value ν'. Let the adversary runs twice. In the second time, the same state as the first time is input, but the different challenge $y \in Z_p^*$ is returned. We let

$$\hat{e}(P_{pub}, Q_s)^{\nu} = \mu\hat{e}(P,\sigma)^{y}, \quad \hat{e}(P_{pub}, Q_s)^{\nu'} = \mu\hat{e}(P,\sigma)^{y'}$$

Thus, we can obtain

$$\hat{e}(P_{pub}, Q_s)^{\frac{-}{-}'} = \hat{e}(P, D_s)^{\frac{-}{-}'} = e(P,\sigma)$$

It means that ID_a can guess the ID_s's private key $D_s = z\sigma$ where $z = \frac{y-y'}{\nu-\nu'}$. Futhermore, ID_a can solve the discrete logarithm problem of z based on $D_s = z\sigma(D_s, \sigma \in G_1)$. According to the hardness of DLP, ID_a cannot authenticate the cihpertext using ID_s's public key.

Corollary 1. *The proposed authentication algorithm hasn't leaked any ciphertext information.*

Proof. We assume that signcrypter will not disclose his picked random number x. In authentication phase, actual signcrypter uses zero knowledge scheme to give a proof that he obtained a secret value $h_s + a_s$, which is useful to decrypt the ciphertext. Obviously, our authentication is secure and hasn't leaked any ciphertext information.

Theorem 3. *(*Anonymity*) The proposed scheme is full anonymous.*

Proof. Given a ciphertext $C = (\bigcup\{\mathcal{U}_i\}, c, R, h_1, h_2, ..., h_n, U_1, U_2, ..., U_n, \sigma)$, we know all $U_i = a_i P(i \neq s)$ and U_s are uniformly distributed in $\mathbb{G}_1$ for a_i randomly generated in Z_q^*. All h_i cannot obtain any private key information for $h_i = H_2(m, \bigcup\{\mathcal{U}_i\}, t, U_i)$ that U_i is uniformly distributed. Since r is randomly generated, $R = rP$ and $R' = \hat{e}(P_{pub}, Q_B)^r$ are uniformly distributed, and k and c are also uniformly distributed. Thus c and R can not provide any signcrypter's information.

It remains to consider whether $\sigma = (h_s + a_s)D_s$ leaks information about the actual signcrypter. We can compute the value of σ, by $\hat{e}(\sigma, P_{pub}) = \hat{e}(h_s D_s, P_{pub}) \cdot \hat{e}(a_s D_s, P_{pub}) = \hat{e}(h_s Q_s, P)\hat{e}(a_s Q_s, P) = \hat{e}(h_s Q_s, P)\hat{e}(U_s + \sum_{i \neq s}(U_i + h_i Q_i), P)$. Since $a_j D_j$ is a secret in signcryption algorithm if ID_j is signcrypter. It seems that we can check whether the equality holds to deduce whether ID_j is actual signcrypter or not: $\hat{e}(h_j Q_j, P) = \hat{e}(P_{pub}, \sigma) \cdot \hat{e}(U_j + \sum_{i \neq j}(U_i + h_i Q_i), P)^{-1}$. However, it is no use in leaking signcrypter information because the above equality not only holds when $j = s$, but also $\forall \in \{1, 2, ..., n\} \backslash \{s\}$. i.e. the signature is symmetric.

It is full anonymous that even an adversary with all the private keys corresponding to the set of $\bigcup\{\mathcal{U}_i\}$ and unbounded computing resources has no advantage in identifying any one of group member over random guessing.

Theorem 4. *(*Unforgeability*) The proposed scheme is existentially unforgeable against adaptive chosen-message and adaptive chosen-identity attacks (EUF-IDVSC-CMIA).*

proof. The unforgeability against adaptive chosen-message and chosen-identity attacks can be derived directly from the security of Chow's ID-based ring signature scheme [10] under the CDH assumption. If an adversary $\mathcal{A}$, who can forge a valid message of the proposed scheme, must be able to forge a valid Chow's ring signature. That is if $\mathcal{A}$ can forge a valid ciphertext on message m, say $C = (\bigcup\{\mathcal{U}_i\}, c, R, h_1, h_2, ..., h_n, U_1, U_2, ..., U_n, \sigma)$ of a user group $\bigcup\{\mathcal{U}_i\}$ and a designated receiver ID, then $\sigma^* = (\bigcup\{\mathcal{U}_i\}, h_1, h_2, ..., h_n, U_1, U_2, ..., U_n, \sigma)$ can be viewed as the Chow's ID-based ring signature on message $m||t$ of the ring $\bigcup\{\mathcal{U}_i\}$, where $t = H_1(\hat{e}(R, D_B))$.

5 Performance and Security Analysis

In this section, we compare our scheme with recent schemes in the literatures in Table 1 and Table 2.

In Table 1, we compare the security about ID-based ring signature/ signcryption schemes. In [4], it provides the authenticatability but no the message confidentiality, while it provides signcryption but no the signcrypter verifiability in [14]. Our proposed scheme supports confidentiality, unforgeability, signcrypter anonymity and signcrypter authenticatability.

Table 1. Comparison of security with related schemes

scheme	security			enc/sig	
	semantic security	anonymous	verifiable	Encrypt	signature
[4]	✓	✓	✓	–	✓
[14]	✓	✓	–	✓	✓
our scheme	✓	✓	✓	✓	✓

Table 2. Comparison of computing and communicating costs

scheme	sig/enc			unsig/dec			auth/ver			ciphertext size
	G_1	G_2	$\hat{e}$	G_1	G_2	$\hat{e}$	G_1	G_2	$\hat{e}$	
[4]	4n-3	1	1	$2n$	n	n	4	2	3	$n\|G_1\|+\|G_2\|+n\|Q_{ID}\|+\|m\|$
[14]	$3n+2$	1	n+2	$2n$	n	3	–	–	–	$2\|\mathbb{G}_1\| + n\|\mathbb{G}_2\| + n\|Q_{ID}\| + n\|Z_q\| + \|m\|$
our scheme	$3n+2$	0	1	$2n$	0	3	3	1	3	$(n+2)\|G_1\| + n\|Q_{ID}\| + n\|Z_q\| + \|m\|$

We consider the costly operations which include the operation in $\mathbb{G}_1, \mathbb{G}_2$ and pairing. Table 2 shows a summary of efficiency of the related schemes, where n is the number of ring members, and $|\mathbb{G}_1|, |\mathbb{G}_2|, |Q_{ID}|$ and $|Z_q|$ represent the element size of $\mathbb{G}_1, \mathbb{G}_2, Q_{ID}$ and Z_q, respectively. We can see that our scheme has higher computing efficiency than the others. Moreover, our scheme shares the same order of ciphertext size as the other schemes.

6 Conclusion

We have successfully integrated the design ideas of the ID-based verifiable anonymous signcryption scheme. In this scheme, a signcrypted message is unconditional anonymous and can be decrypted by a legal receiver but he cannot identify who is the actual signcrypter. Furthermore, the actual signcrypter can give the verifier a proof that the ciphertext is indeed produced by himself. In our scheme, the adversary, who obtains all the group members' private keys, neither does he obtain the plaintext by unsigncrypt algorithm nor the signcrypter identity. It is an interesting open problem to construct a new scheme with a constant size to the number of users.

Acknowledgment

The authors gratefully acknowledge the anonymous reviewers for their valuable comments. This work was supported by the National Natural Science Foundation of China under Grants 60573043 and 60773175, the Foundation of National Laboratory for Modern Communications under Grant 9140c1108010606, and the Foundation of the Key Lab for Guangdong Electronic Commerce Application Technology under Grant 2007gdec0f002.

References

1. Shamir, A.: Identity-based cryptosystem and signature scheme. In: Blakely, G.R., Chaum, D. (eds.) CRYPTO 1984. LNCS, vol. 196, pp. 120–126. Springer, Berlin (1985)
2. Li, C.K., Yang, G., Wong, D.S., Deng, X., Chow, S.S.M.: An efficient signcryption scheme with key privacy. In: López, J., Samarati, P., Ferrer, J.L. (eds.) EuroPKI 2007. LNCS, vol. 4582, pp. 78–93. Springer, Heidelberg (2007)
3. Bao, F., Deng, R.H.: A signcryption scheme with signature directly verifiable by public key. In: Imai, H., Zheng, Y. (eds.) PKC 1998. LNCS, vol. 1431, pp. 55–59. Springer, Berlin (1998)
4. Zhang, J.H.: An efficient identity-based ring signature scheme and its extension. In: Gervasi, O., Gavrilova, M.L. (eds.) ICCSA 2007, Part II. LNCS, vol. 4706, pp. 63–74. Springer, Heidelberg (2007)
5. Baek, J., Steinfeld, R., Zheng, Y.: Formal proofs for the security of signcryption. Journal of cryptology 20, 203–235 (2007)
6. Chen, L., Malone-Lee, J.: Improved identity-based signcryption. In: Vaudenay, S. (ed.) PKC 2005. LNCS, vol. 3386, pp. 362–379. Springer, Heidelberg (2005)

7. Naor, M.: Deniable ring authentication. In: Yung, M. (ed.) CRYPTO 2002. LNCS, vol. 2442, pp. 481–498. Springer, Heidelberg (2002)
8. Barreto, P.S.L.M., Libert, B., McCullagh, N., Quisquater, J.J.: Efficient and provably-secure identity based signatures and signcryption from bilinear maps. In: Roy, B. (ed.) ASIACRYPT 2005. LNCS, vol. 3788, pp. 515–532. Springer, Heidelberg (2005)
9. Chow, S.S.M., Yiu, S.M., Hui, L.C.K., Chow, K.P.: Efficient forward and provably secure ID-based signcryption scheme with public verifiability and public ciphertext authenticity. In: Lim, J.-I., Lee, D.-H. (eds.) ICISC 2003. LNCS, vol. 2971, pp. 352–369. Springer, Heidelberg (2004)
10. Chow, S.S.M., Yiu, S.M., Hui, L.C.K.: Efficient identity based ring signature. In: Ioannidis, J., Keromytis, A.D., Yung, M. (eds.) ACNS 2005. LNCS, vol. 3531, pp. 499–512. Springer, Heidelberg (2005)
11. Rivest, R.L., Shamir, A., Tauman, Y.: How to leak a secret. In: Boyd, C. (ed.) ASIACRYPT 2001. LNCS, vol. 2248, pp. 552–565. Springer, Berlin (2001)
12. Tan, C.H.: Analysis of improved signcryption scheme with key privacy. Information Processing Letters 99(4), 135–138 (2006)
13. Gao, W., Wang, G., Wang, X., Xie, D.: Controllable Ring Signatures. In: Lee, J.K., Yi, O., Yung, M. (eds.) WISA 2006. LNCS, vol. 4298, pp. 1–14. Springer, Heidelberg (2007)
14. Huang, X.Y., Su, W., Yi, M.: Identity-based ring signcryption scheme: cryptographic primitives for preserving privacy and authenticity in the ubiquitious world. In: 19th International conference on Advance Information Networking and Applications, pp. 649–654 (2005)
15. Zheng, Y.: Digital signcryption or how to achieve cost(signature & encryption) $\ll$ cost(signature)+cost(encryption). In: Kaliski Jr., B.S. (ed.) CRYPTO 1997. LNCS, vol. 1294, pp. 165–179. Springer, Berlin (1997)
16. Komano, Y., Ohta, K., Shimbo, A., Kawamura, S.: Toward the Fair Anonymous Signatures: Deniable Ring Signatures. In: Pointcheval, D. (ed.) CT-RSA 2006. LNCS, vol. 3860, pp. 174–191. Springer, Heidelberg (2006)

How to Publicly Verifiably Expand a Member without Changing Old Shares in a Secret Sharing Scheme

Jia Yu[1], Fanyu Kong[2], Rong Hao[1,2], and Xuliang Li[1,2]

[1] College of Information Engineering, Qingdao University,
266071 Qingdao, China
{yujia,hr,lixl}@qdu.edu.cn
[2] Institute of Network Security, Shandong University,
250100 Jinan, China
fanyukong@sdu.edu.cn

Abstract. Publicly verifiable secret sharing is a kind of special verifiable secret sharing, in which the shares of a secret can be verified by everyone, not only shareholders. Because of the property of public verifiability, it plays an important role in key-escrow, electronic voting, and so on. In this paper, we discuss a problem of how to publicly verifiably expand a member without changing old shares in a secret sharing scheme and present such a scheme. In the presented scheme, a new member can join a secret sharing scheme based on discrete logarithms to share the secret with the help of old shareholders. Furthermore, everyone besides the new member can verify the validity of the new member's share and any old shareholder doesn't need to change her old share. That means it is very convenient for key management.

Keywords: forward security, Computation Diffie-Hellman problem, bilinear maps, digital signature.

1 Introduction

Secret sharing and its variations play an important role in distributed cryptographic system. In (t, n) secret sharing [1,2], a secret is divided into n parts called shares by a dealer and these shares are distributed into n shareholders, respectively. Any authorized subset with t honest shareholders can cooperate to recover the secret, while unauthorized subset can't. Thus the secret sharing can be used to protect an important secret. However, secret sharing can't tolerate dishonest dealer and dishonest shareholders, which influences its applications greatly. An appropriate method to overcome this limitation is verifiable secret sharing (VSS) [3,4]. It can verify the validity of the shares by providing some public commitments. If a dealer sends incorrect shares during distribution phase or shareholders provide incorrect shares during secret reconstruction phase, it can be detected by the public commitments. An important extension of secret sharing system is threshold schemes [5~8]. Publicly verifiable secret sharing (PVSS) [9~12] is one kind of special VSS which achieves the goal that not only the shareholders are able to verify whether their shares are valid or not, but also anyone can verify the validity of shares that the shareholders have

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 138–148, 2008.

received. PVSS can be applied to escrow-key, electronic voting, threshold revocable electronic cash and so on.

However, in a (publicly verifiable) secret sharing the shareholders group is not always fixed. Sometimes some new member needs to expand to the system to share the secret with the original shareholders. How to publicly verifiably produce a new share for the new member to join the system in a secret sharing scheme is an important problem. If the trusted dealer is online in this time, it will be a trivial problem because the dealer can make use of distribution algorithm to publicly verifiably distribute a new share to the new member. Unfortunately, it is infeasible for the dealer to be always online because not only the resource is wasted but also the dealer is easy to become a weak point of the system to be attacked by adversaries. So we hope the share for the new member can be computed without the help of a dealer, but with the help of old shareholders. What's more, if the shares of old shareholders change, it will be inconvenient for key management. How to publicly verifiably expand a member without changing old shares in a secret sharing scheme is an interesting question. The motivation of this paper is to put forward a scheme to resolve the problem. We hope the proposed scheme can be widely applied to the secret sharing schemes based on discrete logarithms because the kind of secret sharing schemes dominates the structures of secret sharing schemes.

Previous works: Dynamic secret sharing has attracted a lot of attentions of the researchers. Desmedt and Jajodia [13] proposed a secret redistribution protocol, which can distribute the secret between disjoint groups of shareholders without a dealer. In this protocol, the threshold value can be changed and the group of new shareholders and the group of old shareholders can be fully different. Wong *et al.* [14] gave an improved version able to verify the validity of subshares and old shares using the idea of Feldman's VSS scheme [3]. An extended verifiable secret redistribution protocol for archival systems was presented in [15], which relaxed one assumption in [14]. An efficient new protocol [16] has been proposed. However, when a new member joins the secret sharing scheme, all old shareholders have to change their shares in these protocols. Refs. [17,18] proposed some new protocols to verifiably distribute a share for a new member. In these protocols, old shareholders don't change their shares after distribution. It is obviously more convenient for key management. The above protocols have no the property of public verifiability. A publicly verifiable secret sharing with enrollment ability was proposed in [19]. This scheme is based on [9] and can't be applied to the common secret sharing schemes based on discrete logarithms. Ref. [20] gave a publicly verifiable secret redistribution for threshold secret sharing scheme. However, when a new member joins the secret sharing scheme, all old shareholders still have to change their shares.

Our contribution: We discuss a problem of how to publicly verifiably expand a member without changing old shares in a secret sharing scheme and present such a scheme. In this scheme, we use the member-expansion technique in [18] and the verifiable encryption algorithm based on discrete logarithms in [9]. Different from schemes [13~18], the new scheme can publicly verifiable expand a new member in a secret sharing scheme. Different from scheme [19], the new scheme can be applied to all the secret sharing schemes based on discrete logarithms. At the same time, the scheme overcomes the deficiency of Ref. [20] that the shares of old shareholders are

variable after a new member joins a secret sharing scheme. The scheme needn't secure channels between any two participants because there is no any secret information exposed in communication channels.

Organization: In Section 2, we introduce the preliminaries of our work. A concrete description of our proposal is given in Section 3. In addition, we give the security analysis in Section 4. Finally, Section 5 concludes the paper.

2 Preliminaries

2.1 Notations and Assumptions

p and q are primes *s.t.* $q \mid p-1$. Let G denote a group with prime order p and g be a generator of group G. Let $h \in Z_p^*$ be an element of order q. The shareholders group P is composed by $P_1, P_2, ..., P_n$ and the secret s is shared (t,n) secret sharing scheme among them. The new member to join the system is P_{n+1}.

Assume that there is a dealer during initiation. Each participant is connected to a common broadcast channel. Furthermore, the system is synchronized, i.e., all members can send their information simultaneously in the same round.

2.2 Building Block

(1) Secret sharing scheme [1]
This secret sharing scheme is based on polynomial interpolation. It allows a dealer D to distribute a secret value k to n members, such that any t members can reconstruct the secret. The protocol is computational secure, i.e., any fewer than t members cannot gain any useful information about the secret if discrete logarithm is computationally difficult.

The shared secret k is in Z_p. Choose at random a polynomial

$$f(x) = k + \sum_{j=1}^{t-1} a_j x^j \pmod p \in Z_p[x], \text{ where } a_j \in_R Z_p$$

Compute the secret shares $s_i = f(i) = k + \sum_{j=1}^{t-1} a_j i^j \pmod p$ for each member $P_i \in P$.

The secret reconstruction: According to some subset B ($|B|=t$), compute

$$k = \sum_{P_i \in B} C_{Bi} s_i \pmod p,$$

where $C_{Bi} = \prod_{P_j \in B \setminus \{P_i\}} \frac{j}{(j-i)} \pmod p$.

According to some subset B ($|B|=t$), any shares for $P_j \notin B$ can be computed by the following Eq.

$$s_j = \sum_{P_i \in B} C_{Bi}(j) s_i \pmod p,$$

where $C_{Bi}(j) = \prod_{P_l \in B \setminus \{P_i\}} \frac{j-l}{i-l}$.

(2) Verifiable secret sharing scheme

The verifiable secret sharing scheme is one variation of Feldman's verifiable secret sharing scheme. All secret and shares are from Z_p. And all commitments are from the integer domain. The purpose of these modifications is to use Stadler's verifiable encryption based on discrete logarithms [9].

Like (*t*, *n*) secret sharing scheme, the dealer uses the polynomial

$$f(x) = a_0 + \sum_{j=1}^{t-1} a_j x^j \pmod p \in Z_p[x],$$

where $a_0 = k$

to generate secret shares $s_i = f(i)$ for each member $P_i \in P$. At the same time, the dealer broadcasts commits $\varepsilon_j = g^{a_j}$, $(0 \le j < t)$. Member P_i use Eq.

$$g^{s_i} \stackrel{?}{\equiv} \prod_{j=0}^{t-1} \varepsilon_j^{i^j}$$

to verify whether s_i is right or not.

3 Publicly Verifiable Member Expansion Scheme

The protocol is composed of two phases. The first phase is secret distribution phase. In this phase, a dealer publicly verifiably distributes the shares of a secret into a group of shareholders. This procedure is similar to Stadler's PVSS [10]. The second phase is member expansion phase. In this phase, a group of old shareholders that are randomly selected by the new member to join a secret sharing system help the new member publicly verifiably generate a share. The both phases are described as follows:

① The Secret Distribution Phase

Each participant $P_i (i = 1, 2, ..., n)$ randomly selects $x_i \in_R Z_p$, and then publishes $y_i = h^{x_i}$. Let $H: \{0,1\}^* \to \{0,1\}^l$ be a secure hash function.

(1) The dealer randomly selects a polynomial

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

and computes $s_i = f(i)$, $i = 1, 2, ..., n$. The dealer broadcasts g^s, $g^{a_i} (i = 1, 2, .., t-1)$.

(2) The dealer encrypts each s_i by a variation of ElGamal encryption algorithm:

She selects $l_i \in_R Z_q$, computes

$$\gamma_i = h^{l_i} \pmod p \tag{2}$$

$$\delta_i = s_i^{-1} y_i^{l_i} \pmod p \tag{3}$$

and publishes (γ_i, δ_i) as the ciphertext of the share s_i. And then selects $w_k \in Z_q$, $k = 1, 2, ..., l$, computes and broadcasts

$$T_{h,i,k} = h^{w_k} \pmod{p} \tag{4}$$

$$T_{g,i,k} = g^{y_i^{w_k}} \tag{5}$$

where $i = 1, 2, ..., n$.

She computes

$$c_i = H(g \| h \| \gamma_i \| \delta_i \| T_{h,i,1} \| T_{h,i,2} \| ... \| T_{h,i,l} \| T_{g,i,1} \| T_{g,i,2} \| ... \| T_{g,i,l}) \tag{6}$$

(3) Let $c_{i,k}$ denote the k-th bit of c_i. The dealer computes $r_{i,k} = w_k - c_{i,k} l_i$, where $k = 1, 2, ..., l$ and publishes $\Pr oof_D = (c_i, r_{i,1}, ..., r_{i,l})$.

(4) Each participant $P_i (i = 1, 2, ..., n)$ decrypts

$$s_i = \gamma_i^{x_i} \cdot \delta_i^{-1} \pmod{p} \tag{7}$$

and verifies the following equation

$$g^{s_i} = g^s \prod_{j=1}^{t-1} (g^{a_j})^{i^j} \tag{8}$$

holds or not. If it holds, P_i believes his share is correct and sets $E_i = g^{s_i}$. Otherwise, publishes s_i and broadcasts a complaint against the dealer.

(5) Each participant $P_j (j = 1, 2, ..., n)$ checks the validity of share s_i $(i \neq j)$. She computes

$$E_i = g^s \prod_{j=1}^{t-1} (g^{a_j})^{i^j} \tag{9}$$

$$T_{h,i,k} = h^{r_{i,k}} \gamma_i^{c_{i,k}} \tag{10}$$

$$T_{g,i,k} = (g^{1-c_{i,k}} E_i^{c_{i,k} \delta_i})^{y_i^{r_{i,k}}} \tag{11}$$

And then verifies whether equation (6) holds. If it holds, then believes s_i is correct. Otherwise, generates a complaint against the dealer.

② **Member Expansion Phase**

When a new member P_{n+1} will join the secret sharing system. She randomly selects $x_{n+1} \in_R Z_p$ and publishes $y_{n+1} = h^{x_{n+1}}$. And then she randomly chooses a group of t old shareholders to help her publicly verifiably generate a new share s_{n+1}. If some old shareholders chosen are dishonest, the new member will choose another group to

restart the member expansion protocol till she gets the correct share. W.l.o.g. Assume that the chosen participants are $P_1, P_2, \ldots, P_t$. Let $B = \{P_1, P_2, ..., P_t\}$.

(1) Each $P_i (i = 1, 2, ..., t)$ randomly selects $\sigma_{ij} (j = 1, 2, ..., t)$ such that

$$\sum_{j=1}^{t} \sigma_{ij} = C_{Bi}(n+1) s_i \pmod{q} \tag{12}$$

Where $C_{Bi}(n+1) = \prod_{P_l \in B \setminus \{P_i\}} \frac{n+1-l}{i-l}$.

She selects $l_{i,j} \in_R Z_q$, computes

$$\gamma_{i,j} = h^{l_{i,j}} \pmod{p} \tag{13}$$

$$\delta_{i,j} = \sigma_{i,j}^{-1} y_j^{l_{i,j}} \pmod{p} \tag{14}$$

and publishes $(\gamma_{i,j}, \delta_{i,j})$ as the ciphertext of the share $\sigma_{i,j}$. P_i computes and publishes $E_{il} = g^{\sigma_{il}}$.

(2) Each $P_i (i = 1, 2, ..., t)$ selects $w_k \in Z_q, k = 1, 2, ..., l$, computes and broadcasts

$$T_{h,i,j,k} = h^{w_k} \pmod{p} \tag{15}$$

$$T_{g,i,j,k} = g^{y_j^{w_k}} \tag{16}$$

where $j = 1, 2, ..., n$.

She computes

$$c_{i,j} = H(g \| h \| \gamma_{i,j} \| \delta_{i,j} \| T_{h,i,j,1} \| T_{h,i,j,2} \| ... \| T_{h,i,j,l} \| T_{g,i,j,1} \| T_{g,i,j,2} \| ... \| T_{g,i,j,l}) \tag{17}$$

(3) Let $c_{i,j,k}$ denote the k-th bit of $c_{i,j}$. P_i computes $r_{i,j,k} = w_{i,k} - c_{i,j,k} l_{i,j}$, where $k = 1, 2, ..., l$ and publishes $\Pr oof_D = (c_{i,j}, r_{i,j,1}, ..., r_{i,j,l})$

(4) Each $P_j (j = 1, 2, ..., t)$ decrypts

$$\sigma_{i,j} = \gamma_{i,j}^{x_j} \cdot \delta_{i,j}^{-1} \pmod{p} \tag{18}$$

and verifies the following equations

$$g^{\sigma_{i,j}} = E_{i,j} \tag{19}$$

$$\prod_{l=1}^{t} E_{il} = (E_i)^{C_{Bi}(n+1)} \tag{20}$$

hold or not. If they don't hold, abort.

(5) Member $P_l(l \neq j)$ checks the validity of value $\sigma_{i,j}$ $(j \neq i)$. She computes

$$T_{h,i,j,k} = h^{r_{i,j,k}} \gamma_i^{c_{i,j,k}} \tag{21}$$

$$T_{g,i,j,k} = (g^{1-c_{i,j,k}} E_{i,j}{}^{c_{i,j,k}\delta_{i,j}})^{y_j^{r_{i,j,k}}} \tag{22}$$

And then verifies whether equation (17) holds. If it holds, then believes $\sigma_{i,j}$ is correct. Otherwise, abort.

(6) Each $P_j(j = 1,2,...,t)$ computes

$$s_j' = \sum_{i=1}^{t} \sigma_{i,j} (\text{mod } p) \tag{23}$$

She selects $l_{j,n+1} \in_R Z_q$, computes

$$\gamma_{j,n+1} = h^{l_{j,n+1}} (\text{mod } p) \tag{24}$$

$$\delta_{j,n+1} = s_j'^{-1} y_j^{l_{j,n+1}} (\text{mod } p) \tag{25}$$

and publishes $(\gamma_{j,n+1}, \delta_{j,n+1})$ as the ciphertext of the share s_j'. P_j computes and publishes $E_j' = g^{s_j'}$.

She selects $w_k \in Z_q$, $k = 1,2,...,l$, computes and broadcasts

$$T_{h,j,n+1,k} = h^{w_k} (\text{mod } p) \tag{26}$$

$$T_{g,j,n+1,k} = g^{y_{n+1}{}^{w_k}} \tag{27}$$

She computes

$$c_{j,n+1} = H(g \| h \| \gamma_{j,n+1} \| \delta_{j,n+1} \| T_{h,j,n+1,1} \| T_{h,j,n+1,2} \| ... \| T_{h,j,n+1,l} \| T_{h,j,n+1,l}) \tag{28}$$

(7) Let $c_{j,n+1,k}$ denote the k-th bit of $c_{j,n+1}$. P_j computes $r_{j,n+1,k} = w_{j,k} - c_{j,n+1,k} l_{j,n+1}$, where $k = 1,2,...,l$ and publishes $\text{Pr}oof_D = (c_{j,n+1}, r_{j,n+1,1}, ..., r_{j,n+1,l})$.

(8) New participant P_{n+1} decrypts

$$s_j' = \gamma_{j,n+1}{}^{x_j} \cdot \delta_{j,n+1}{}^{-1} (\text{mod } p) \tag{29}$$

and verifies the following equation

$$g^{s_j'} = \prod_{i=1}^{t} E_{i,j} \tag{30}$$

holds or not. If it holds, then the new participant computes his share

$$s_{n+1} = \sum_{l=1}^{t} s_l' \pmod{p} \tag{31}$$

Otherwise, abort.

(9) Member $P_l (l \neq n+1)$ checks the validity of value s_j' $(l \neq j)$. She computes

$$T_{h,j,n+1,k} = h^{r_{j,n+1,k}} \gamma_i^{c_{j,n+1,k}} \tag{32}$$

$$T_{g,j,n+1,k} = (g^{1-c_{j,n+1,k}} (E_j')^{c_{j,n+1,k}\delta_{j,n+1}})^{y_{n+1}{}^{r_{j,n+1,k}}} \tag{33}$$

And then verifies whether equation (28) holds. If it holds, then believes s_j' is correct. Otherwise, abort.

4 Security Analysis

In security analysis, we consider the static adversary who selects the members to corrupt at the beginning of the protocol.

Theorem 1. If the shareholders to help new member P_{n+1} generate the share are honest, then member P_{n+1} can get the right new share by executing the presented protocol.

Proof. It is because:

$$\begin{aligned} s_{n+1} &= \sum_{l=1}^{t} s_l' \\ &= \sum_{l=1}^{t} \sum_{i=1}^{t} \sigma_{il} \\ &= \sum_{i=1}^{t} \sum_{l=1}^{t} \sigma_{il} \\ &= \sum_{i=1}^{t} C_{Bi}(n+1) s_i \end{aligned}$$

Theorem 2. The dishonest participating shareholders can be discovered in the presented scheme. And when $n \geq 2t-1$, even if an adversary can corrupt t-1 old shareholders at the beginning of the protocol, the new member still can get the right share.

Proof. In secret distribution phase, the dishonest participating shareholders can publish error complaint against the dealer in step (4) or (5). However, other shareholders can discover it by equation (8) or equations (9)(10)(11).

In member expansion phase, a dishonest participating shareholder can deceive other members as follows:

Case 1: She can give other shareholders error $\sigma_{i,j}$ or its ciphertext. However, other shareholders can discover it by equations (19)(20)(21)(22).

Case 2: She can give other shareholders error s_j' or its ciphertext. However, other shareholders can discover it by equations (30)(31)(32)(33).

Therefore, the dishonest participating shareholders can be discovered in the presented scheme.

When $n \geq 2t-1$, even if an adversary can corrupt t-1 old shareholders, there are still no fewer than t honest shareholders. If the selected set of participants B includes dishonest participants, the protocol will abort in steps (4)(5)(8) or (9). And then we will select another set B' of t participants to execute the protocol. Because the number of honest shareholders is no fewer than t, there must be a set B' which is composed of honest participants. So these participants can help the new member get right share. In the worst case, the member expansion protocol needs to be executed $C_n^t - C_{n-t+1}^t$ times.

Theorem 3. The presented protocol satisfies that:

(1) If an adversary corrupts t-1 members, she can't get any information about the secret and other old members' shares in the scheme.
(2) The new member P_{n+1} can't get any information about the shares of old shareholders in the scheme.
(3) Any t-1 old shareholders can't get the share of new member P_{n+1}.

Proof (Sketch)

(1) W.l.o.g, assume the adversary corrupts members $P_1, P_2, ..., P_{t-1}$. She knows the shares including $s_1, s_2, ..., s_{t-1}$. If the adversary wants to compute the secret and other old member's shares, she must compute s_t. However, she only knows values $\sigma_{t,j}\ (j = 1,2,...,t-1)$. Because $\sigma_{t,t}$ is a random value and what the adversary knows only is its ciphertext $(\gamma_{t,t}, \delta_{t,t})$, the adversary can't get s_t from equation (12). It means the adversary can't get any message about the secret and other old members' shares.

(2) What the new member P_{n+1} gets from the old shareholders are values s_j' and $\mathrm{Pr}oof_D = (c_{j,n+1}, r_{j,n+1,1}, ..., r_{j,n+1,l})$, $(j = 1,2,...,t)$. Because s_j' is random and independent to share s_j of shareholder P_j, and $\mathrm{Pr}oof_D$ has no relation to s_j, new member P_{n+1} can't get any information about the shares of old shareholders in the scheme.

(3) Any t-1 old shareholders only know no more than t-1 values s_j' because other s_j' is encrypted. So they can't compute s_{n+1} from equation (31).

5 Conclusions

We give a scheme that can publicly verifiably expand a member in a secret sharing scheme based on discrete logarithms without changing old shares in the paper. Because the shares for old shareholders don't need to change after a new member joins the secret sharing system, it is convenient for secret key management. In the near further, we will research on the efficient schemes against mobile adversary to deal with the above problem discussed.

Acknowledgments. We would like to thank anonymous referees of 2008 Pacific Asia Workshop on Intelligence and Security Informatics (PAISI 2008) for the suggestions to improve this paper. This research is supported by National Natural Science Foundation of China (60703089) and the National High-Tech R & D Program (863 Program) of China (2006AA012110).

References

1. Shamir, A.: How to Share a Secret. Communications of the ACM 22(11), 612–613 (1979)
2. Blakley, G.R.: Safeguarding cryptographic keys. In: Proc. AFIPS 1979 National Computer Conference, vol. 48, pp. 313–317. AFIPS Press, NJ (1979)
3. Feldman, P.: A Pratical Scheme for Non-Interactive Verifiable Secret Sharing. In: Proc. 28th Annual FOCS, pp. 427–437. IEEE Press, New York (1987)
4. Pedersen, T.P.: Non-Interactive and Information-Theoretic Secure Verifiable Secret Sharing. In: Feigenbaum, J. (ed.) CRYPTO 1991. LNCS, vol. 576, pp. 129–140. Springer, Heidelberg (1992)
5. Gennaro, R., Jarecki, S., Krawczyk, H., Rabin, T.: Robust Threshold DSS Signatures. In: Maurer, U.M. (ed.) EUROCRYPT 1996. LNCS, vol. 1070, pp. 354–371. Springer, Heidelberg (1996)
6. Shoup, V.: Practical threshold signature. In: Preneel, B. (ed.) EUROCRYPT 2000. LNCS, vol. 1807, pp. 207–220. Springer, Heidelberg (2000)
7. Abdalla, M., Miner, S., Namprempre, C.: Forward-secure threshold signature schemes. In: Naccache, D. (ed.) CT-RSA 2001. LNCS, vol. 2020, pp. 441–456. Springer, Heidelberg (2001)
8. Baek, J., Zheng, Y.L.: Identity-based threshold decryption. Cryptology ePrint Archive, Report 2003/164 (2003)
9. Schoenmakers, B.: A simple Publicly Verifiable Secret Sharing Scheme and its Application to Electronic Voting. In: Wiener, M. (ed.) CRYPTO 1999. LNCS, vol. 1666, pp. 148–164. Springer, Heidelberg (1999)
10. Stadler, M.: Public verifiable secret sharing. In: Maurer, U. (ed.) EUROCRYPT 1996. LNCS, vol. 1070, pp. 190–199. Springer, Heidelberg (1996)
11. Fujisaki, E., Okamoto, T.: A practical and provably secure scheme for publicly verifiable secret shaing and its applications. In: Nyberg, K. (ed.) EUROCRYPT 1998. LNCS, vol. 1403, pp. 32–47. Springer, Heidelberg (1998)
12. Young, A., Yung, M.: A PVSS as Hard as Discrete Log and Shareholder Separability. In: Kim, K.-c. (ed.) PKC 2001. LNCS, vol. 1992, pp. 287–299. Springer, Heidelberg (2001)
13. Desmedt, Y., Jajodia, S.: Redistributing secret shares to new access structures and its application. Technical Report ISSE TR-97-01, George Mason University (1997)

14. Wong, T.M., Wang, C.X., Wing, J.M.: Verifiable secret redistribution for archive systems. In: Proc. of the 1st International IEEE Security in Storage Workshop, pp. 94–106. IEEE Press, New York (2002)
15. Gupta, V., Gopinaph, K.: An Extended Verifiable Secret Redistribution Protocol for Archival Systems. In: The First International Conference on Availability, Reliability and Security 2006, pp. 8–15. IEEE Press, New York (2006)
16. Yu, J., Kong, F.Y., Li, D.X.: Verifiable Secret Redistribution for PPS Schemes. In: Proc. of the 2nd Information Security Practice and Experience Conference, Journal of Shanghai Jiaotong University Science, vol. E-11(2), pp. 71–76 (2006)
17. Li, X., He, M.X.: A protocol of member-join in a secret sharing scheme. In: Chen, K., Deng, R., Lai, X., Zhou, J. (eds.) ISPEC 2006. LNCS, vol. 3903, pp. 134–141. Springer, Heidelberg (2006)
18. Yu, J., Kong, F.Y., Hao, R., Cheng, Z.: A Practical Member Enrollment Protocol for Threshold Schemes. Journal of Beijing University of Posts and Telecommunica-tions 28(z.2), 1–3, 8 (2006) (in Chinese)
19. Yu, J., Kong, F.Y., Hao, R.: Publicly Verifiable Secret Sharing with Enrollment Ability. In: The 8th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, pp. 194–199. IEEE Computer Society, New York (2007)
20. Tan, Z.W., Liu, Z.J.: Publicly Verifiable Secret Redistribution for Threshold Secret Sharing Scheme. Journal of the Graduate School of the Chinese Academy of Sciences 21(2), 210–217 (2004)

Comparing Two Models for Terrorist Group Detection: GDM or OGDM?

Fatih Ozgul[1], Zeki Erdem[2], and Hakan Aksoy[3]

[1] School of Computing & Technology, University of Sunderland, St.Peter's Way. SR6 0DD, Sunderland, United Kingdom
[2] TUBITAK- MAM Research Centre, Information Technologies Institute, 41470 Gebze, Kocaeli, Turkey
[3] Bursa Police Department, Information Processing Unit, 16050 Bursa, Turkey
fatih.ozgul@sunderland.ac.uk, zeki.erdem@bte.mam.gov.tr, aksoy975@yahoo.com

Abstract. Since discovery of organization structure of offender groups leads the investigation to terrorist cells or organized crime groups, detecting covert networks from crime data are important to crime investigation. Two models, GDM and OGDM, which are based on another representation model - OGRM are developed and tested on eighty seven known offender groups where nine of them were terrorist cells. GDM, which is basically depending on police arrest data and "caught together" information, performed well on terrorist groups, whereas OGDM, which uses a feature matching on year-wise offender components from arrest and demographics data, performed better on non-terrorist groups. OGDM uses a terror crime modus operandi ontology which enabled matching of similar crimes.

Keywords: Social network analysis, crime analysis, visualization, criminal data mining, terrorism related analytical methodologies and software tools.

1 Introduction

Group detection refers to the discovery of underlying organizational structure that relates selected individuals with each other, in broader context; it refers to the discovery of underlying structure relating instances of any type of entity among themselves [21]. Link analysis and group detection is a newly emerging research area which is at the intersection of link analysis, hypertext – web mining, graph mining [6, 7] and social network analysis [24, 28, 31]. Graph mining and social network analysis (SNA) in particular attracted attention from a wide audience in police investigation and intelligence [11, 12]. As a result of this attention, the police and intelligence agencies realized the knowledge about offender networks and detecting covert networks are important to crime investigation [29].

Since discovery of an underlying organizational structure from crime data leads the investigation to terrorist cells or organized crime groups, detecting covert networks

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 149–160, 2008.

are important to crime investigation. Detecting an offender group, a terrorist network or even a part of group (subgroup) is also important and valuable. A subgroup can be extended with other members with the help of domain experts. An experienced investigator usually knows the friends of well-known terrorists, so he can decide which subgroups should be united to constitute the whole network. Specific software like Analyst Notebook [1] and Sentient [30] provide some visual spatio-temporal representations of offender groups in graphs, but they lack automated group detection functionality. In this paper, we make the following contributions for terrorist group detection;

- We identify and discuss converting arrest and demographics data to link table and graph format where there was no standardized way of doing this. We suggest Offender Group Representation Model (OGRM) for terrorist networks where some SNA metrics can easily be used (section 3).
- We demonstrate a terrorism modus operandi ontology system for matching similar modus operandi information (section 5). Surprisingly there has been no example of such ontology before.
- We show how two models GDM and OGDM performed for detecting terrorist and other groups (section 6).
- We discuss either police arrest data or offender demographics data is more important to detect terrorist groups (section 6-7).

2 Terrorist Group Detection from Raw Crime Data

Although there is a distinction between detecting terrorist groups and offender groups, when we focus on offender group detection in general, the most remarkable works are CrimeNet Explorer, which is developed by Xu et al. [17, 18] and Terrorist Modus Operandi Detection System (TMODS), which is developed by 21st Century Technologies [23].

2.1 CrimeNet Explorer

Xu et al. [17, 18] defined a framework for automated network analysis and visualization. Using COPLINK connect and COPLINK detect [2, 3, 4, 5] structure to obtain link data from text, CrimeNet Explorer used a Reciprocal Nearest Neighbor (RNN) based clustering algorithm to find out links between offenders, as well as discovery of previously unknown groups. CrimeNet Explorer framework includes four stages: network creation, network partition, structural analysis, and network visualization (Figure 1).

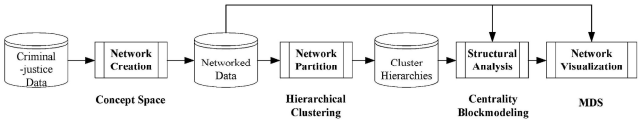


Fig. 1. CrimeNet Explorer framework

CrimeNet Explorer uses concept space approach for network creation, RNN-based hierarchical clustering algorithm for group detection; social network analysis based structural analysis and Multi Dimensional Scaling for network visualization. CrimeNet Explorer is the first model to solve offender group discovery problem and its success comes from the powerful functionality of overall COPLINK structure [2, 3, 4, 5]. On the other hand, since CrimeNet Explorer was evaluated by university students for its visualization, structural analysis capabilities, and its group detection functionality, the operationally actionable outputs of CrimeNet Explorer on terrorist groups has not been proved.

2.2 TMODS

TMODS, which is developed by 21st Century Technologies [21], automates the tasks of searching for and analyzing instances of particular threatening activity patterns (Figure 3). With TMODS, the analyst can define an attributed relational graph to represent the pattern of threatening activity he or she is looking for. TMODS then automates the search for that threat pattern through an input graph representing the large volume of observed data.

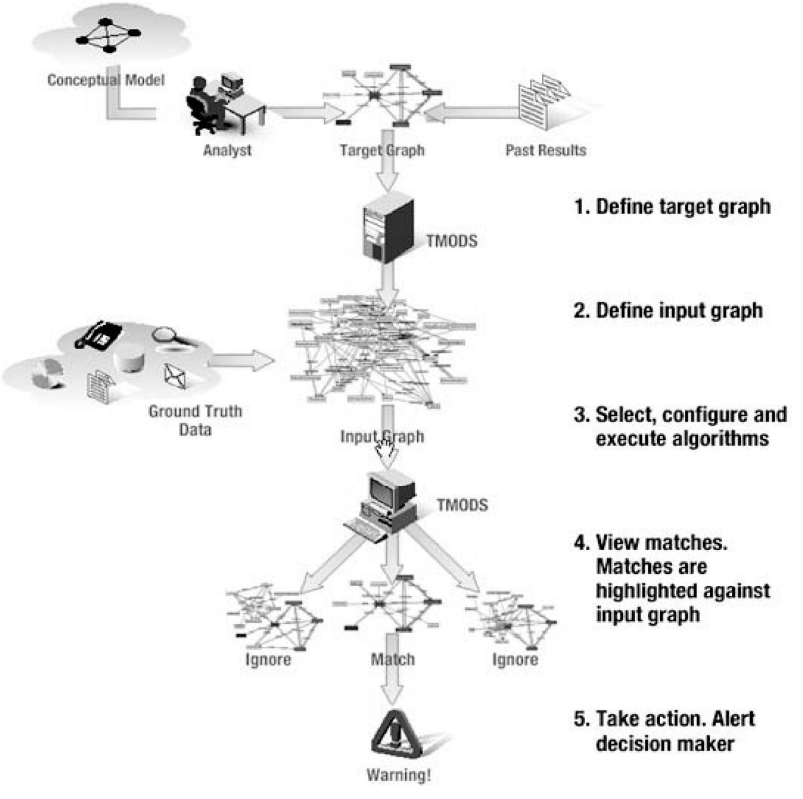


Fig. 2. TMODS framework. A possible search pattern is matched to observed activity by using a pattern in ontology.

TMODS pinpoints the subset of data that match the threat pattern defined by the analyst thereby transforming a manual search into an efficient automated graph matching tool. User defined threatening activity or pattern graph can be produced with possible terrorist network ontology and this can be matched against observed activity graph. At the end, human analyst views matches that are highlighted against the input graph. TMODS is mature and powerful distributed java software that has been under development since October 2001 [21]. But it needs a pattern graph and an analyst to run the system. Like a supervised learning algorithm, TMODS tries to tailor the results according to pre-defined threatening activity. Another possible drawback is graphs used in TMODS are multi-mode and can be disadvantageous for further analysis. Multi-mode graph means that nodes in multi-mode graphs are more than two types of entities. A person, a building, an event, a vehicle are all represented as nodes; when for instance we want to detect key players in multi-mode graph, a building can be detected as key player, not a person. This can be a cause of confusion. To overcome this confusion the definition of a one-mode (friendship) social network should be used rather than representing all entities as nodes.

3 Offender Group Representation Model (OGRM)

It is better to represent actors (offenders) as nodes and rest of the relations as edges in one-mode (friendship) social networks (Figure 3). This can produce many link types such as “co-defendant link”, “spatial link”, “same weapon link”, and “same modus operandi link”. Thereby many graph theoretical and SNA solutions can be used on one-mode (friendship) networks effectively such as friendship identification, finding key actors.

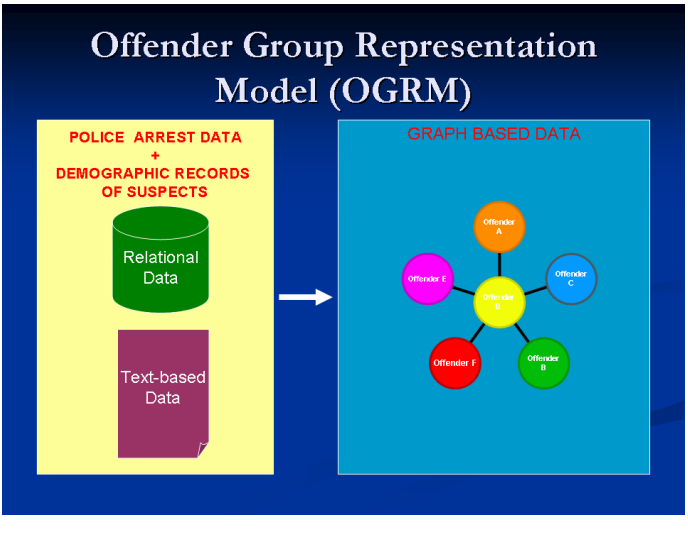


Fig. 3. To avoid confusion on representation, a one-mode network representation (vertices as offenders, links as relations) is recommended for best representation of offender groups.

4 Group Detection Model (GDM)

It is a fact that people who own the same name and surname can mislead the investigator and any data mining model. Minimum requirement to apply GDM is to have a police arrest table or text where it has to include unique crime reference

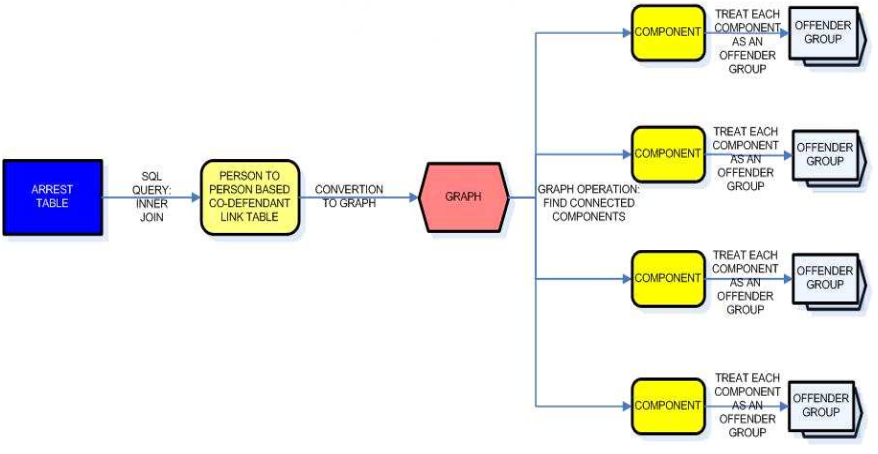


Fig. 4. Group Detection Model (GDM)

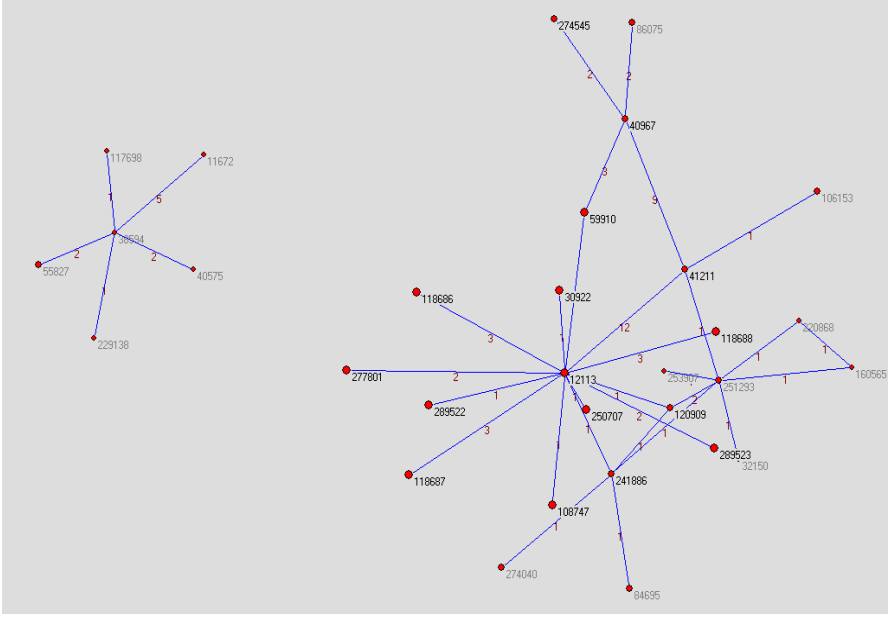


Fig. 5. This graph shows two theft groups generated from police arrest table using GDM, consisting of 31 offenders with their person_id numbers (nodes), 30 undirected "co-defendant" links (edges)

number and unique person number. This ensures that the information doesn't include duplicate arrest records and a person only has one unique number or reference. The source of link information is gathered from police arrest records using an inner join query. (Figure 4) Inner join query result, which we call co-defendant link table; consisting of From Offender, To Offender, and W (how many times this offender pair caught together by the police) is produced with inner join SQL query. Then this link table is converted to graph where nodes represent offenders, edges represent crimes using OGRM representation model.

Number of times caught together is counted to be used for edge weight (W). At this point a subgraph detection operation is needed; we used strongly connected components (SCC) algorithm because it is scalable and gives concrete results. A directed graph [8] is called strongly connected if for every pair of vertices has a path towards each other. The strongly connected components of a directed graph are its maximal strongly connected subgraphs like in figure 5. In GDM, every component represents a unique offender group because one offender can only belong to one group thereby concrete a result of group membership is obtained.

5 Offender Group Detection Model (OGDM)

Due to lack of enclosure of crime specific attributes in GDM, we needed to develop our model for better results on offender group detection. OGDM is developed to add more functionality by including time, location, modus operandi and demography similarity dimensions for offender group detection. Following steps are taken in OGDM as exhibited in figure 6. As well as the arrest table, which is also used in GDM, we also need some demographic knowledge fields about offenders; literally offender surnames and offender place of origin. The first operation to be done on arrest table is dividing all arrest records in year-wise segments. Then, year-wise arrest table is converted to co-defendant link table with an inner query operation which is consisting of fields From Offender, To Offender, and W. Then using this table a graph is obtained and its connected components are found using SCC algorithm [8]. The result of which arrested persons belong to which component also gives which persons in which year are within the same component and additional information for surname and birthplace similarity for component-to-component. For the next step, components matching operations are done; for modus operandi similar, spatially similar, time sequence similar matching in arrest table and for surname similar, birthplace similar matching in demographics table.

Ontologies are beneficial for importing the domain knowledge from the sources to reuse [14] so we decided to create a terrorists modus operandi system in OGDM. As shown in figure 7, using terrorist modus operandi ontology, we matched arrest records of terrorists in year-wise components. Top node of tree named "Terror Crimes", one below second level there are 12 terror crimes such as "being member of terrorist organisation", "delivery of terrorist progandating material". On third level from the top, there are five main types of modus operandi such as "conspiracy", "destruction", "gadget". Conspiracy type was the biggest matching modus operandi parameter on third level. The lowest level gives in 13 detailed types of modus operandi such as "on behalf of Ansar al islam", "delivering terrorist leaflets"

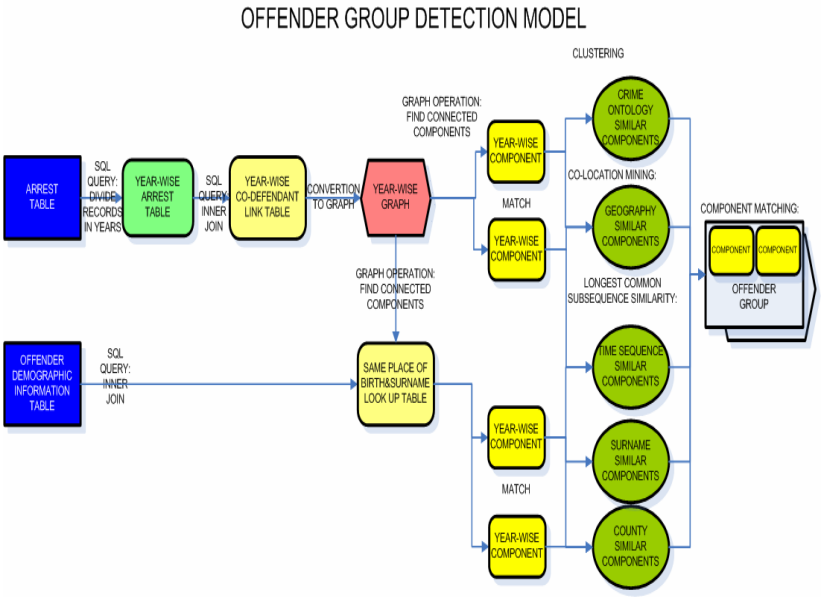


Fig. 6. Offender Group Detection Model (OGDM)

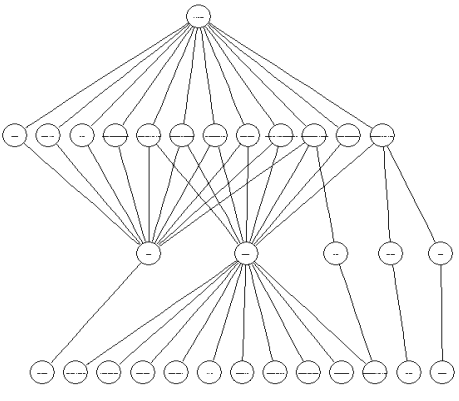


Fig. 7. Modus Operandi Ontology for Terror Crimes

As a result, we have a rank-ordered list of component pairs with a Jaccard coefficient similarity score. The Jaccard coefficient is the size of the intersection divided by the size of the union of the sets (components). So for given two components A and B, Jaccard coefficient similarity score is as;

$$J(A,B) = |A \cap B| / |A \cup B|. \quad (1)$$

As the last step, we need to help domain experts for finding out which feature in our data is the most promising; so that domain experts can decide the best threshold

can be used to reduction of unnecessary data. One possible technique for feature selection is based on comparison of means and variances. To summarize the key characteristics of the distribution of values for a given feature, it is necessary to compute the mean value and the corresponding variance. Equation 2 and 3 formalizes a scoring test (TEST), where B (for birthplace similarity score), G (for geographic similarity score), M (for modus operandi similarity score), T (for Time-series similarity score) and S (for Surname similarity score) are sets of feature values measured for different matching types (birth, geo, modus, time, and surname) and components compnum1 … to compN, are the corresponding number of samples:

$$SE\ (B, G, M, T, S) = \sqrt{\begin{array}{l} \mathrm{var}(B) / compnumB + \\ \mathrm{var}(G) / compnumG + \\ \mathrm{var}(M) / compnumM + \\ \mathrm{var}(T) / compnumT + \\ \mathrm{var}(S) / compnumS \end{array}} \tag{2}$$

$$\mathrm{TEST} = \frac{|\mathrm{Max\_mean}(B,G,M,T,S) - \mathrm{Min\_mean}(B,G,M,T,S)|}{\mathrm{SE}\ (B,G,M,T,S)} \tag{3}$$

where TEST score must be bigger then domain expert's threshold value. For k features, k pair-wise comparisons can be made, comparing each feature to another. A

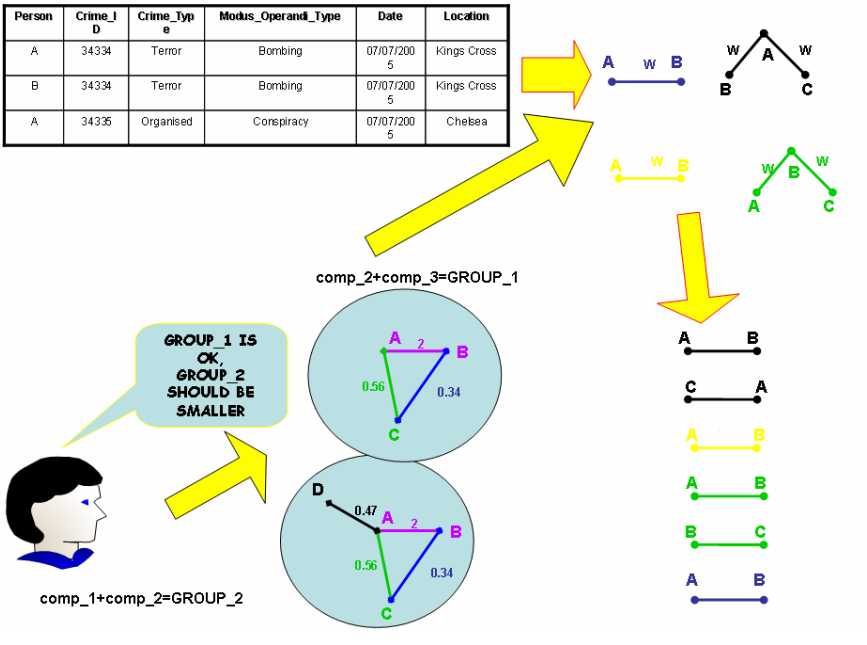


Fig. 8. Overall system with the contribution of domain expert

feature is retained if it is significant for any of the pair-wise comparisons. After deciding which features are more or less important for offender group detection, OGDM decides which components are highly likely to provide an offender group. In practise, two or more components can constitute an offender group if a domain expert thinks the size of the group should be bigger, if there is no need to extend the group size only one component is kept as the target offender group (Figure 8).

6 Terrorist Groups Detected

We conducted experiments on Bursa data set. Crime records in Bursa, Turkey crime and offender demographics data set are available from 1991 to August 2007 as 318352 crimes committed by 199428 offenders. Crime types are varying from organized crimes, narcotic gangs, theft or illegal enterprises to terrorist groups. All experimental setup and operations are done using R [10, 27] and related R libraries [9,10,19]. There were 9 terrorist groups and 78 other type offender groups available such as theft groups, drug dealing groups, mafia type groups. Domain experts decided to set threshold value for similarity as ten per cent (0,1). 87 groups, where 9 of them were terrorist groups, are selected as "golden standard" to measure the success rate of GDM and OGDM. Nine terrorist groups detected are presented in below Table 1.

Table 1.

Group Number #	Group Name	Group Type	Number of detected persons by OGDM	Number of detected persons by GDM	Number of persons in reality
2	TDKP Legal Cell	Terror	8	13	14
3	TDKP İllegal Cell	Terror	8	13	14
4	TKP/ML Cell	Terror	5	5	6
5	PKK Cell	Terror	15	16	18
6	Racist Terrorist Group	Terror	13	15	18
7	Communist DHKP/C Cell	Terror	6	19	21
8	Communist TIKB Cell	Terror	17	16	17
9	MKP Terrorist Cell	Terror	4	4	4
10	PKK Cell	Terror	15	13	15

In general terms, average success rate for terrorist groups using OGDM is 0, 7699813 whereas using GDM it is 0,853706816. Average success rate for terrorist groups are given detailed in Table 2.

Average success rate for other (non-terrorist) 78 groups such as theft groups, drug dealing groups, mafia type groups are using OGDM is 0,768451 whereas using GDM it is 0,542807. The comparison of success rates for OGDM and GDM over terrorists and non-terrorists groups are exhibited in figure 9. Domain experts' comments about that are apparently GDM does better then OGDM on terrorists groups simply because there is no need to get crime or demographic similarity matching for terrorist groups. They added that terrorist cells are isolated from the outer world, members of cells can be coming from different demographic origins or they might not have a criminal career. All information we know about them is their co-existence and we realize this fact when we see acting together. So co-defendant links or "caught-together" information is the most needed aspect we need to detect their groups. Whereas in

Table 2.

group no	Success_rate_OGDM	Success_rate_GDM
2	0,571428571	0,928571429
3	0,571428571	0,928571429
4	0,833333333	0,833333333
5	0,833333333	0,888888889
6	0,722222222	0,833333333
7	0,285714286	0,904761905
8	1	0,941176471
9	1	1
10	1	0,866666667

non-terrorist groups criminal career is a fact for the group members, and they randomly act together. They don't have a strict group discipline and many times they are peer-groups, friends or just decided to act for a couple of crimes. So they have loose tight in their relationships, whenever they want, they simply change groups, even in many cases they don't percept their counterparts as their group members.

Since "caught together" information is less valuable for non-terrorist groups, as a result OGDM performed better on non-terrorist groups to compare againist terrorist group.

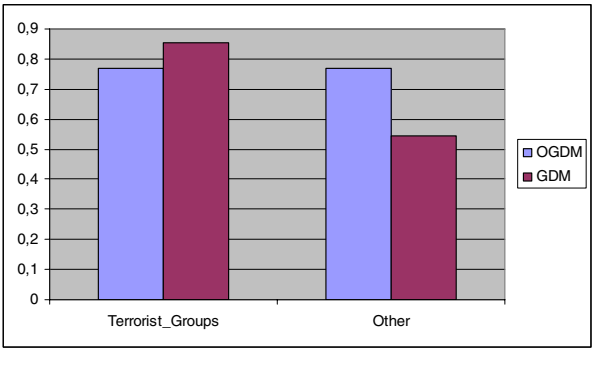


Fig. 9. Comparison of performances by OGDM and GDM over terrorist groups and non-terrorist groups

7 Conclusion

The aim for any crime data mining model must be producing *operationally actionable output* [20, 22]. Both models are proved by Bursa Police Department's positive feedback [25, 26, 32]. GDM performed well over OGDM for terrorist groups whereas OGDM performed better for other type of groups. This shows that arrest data, especially co-defendant information gathered from arrest data, is very valuable for

terrorist networks. Demographic data is less valuable then arrest data for detecting terrorist groups. It is very important for law enforcement agencies to keep their arrest data with unique crime reference numbers and person id numbers, thereby the police and other agencies can benefit from this for using this information deciding the links between terrorists.

References

1. Analyst Notebook, i2 Analyst Notebook i2 Ltd. (2007), http://www.i2.co.uk/ (accessed July 31, 2007)
2. Chen, H., Atabakhsh, H., et al.: Visualisation in Law Enforcement. In: CHI 2005, Portland, Oregon, USA, April 2-7, ACM, New York (2005)
3. Chen, H., Xu, J.J.: Fighting Organised Crimes: using shortest-path algorithms to identify associations in criminal networks. Decision Support Systems 38(3), 473–487 (2003)
4. Chen, H., Chung, W., et al.: Crime data mining: a general framework and some examples. Computer 37(4), 50–56 (2004)
5. Chen, H., Schroeder, J., et al.: COPLINK Connect: information and knowledge management for law Enforcement. Decision Support Systems 34, 271–285 (2002)
6. Cook, D.J., Holder, L.B.: Graph-Based data mining. IEEE Intelligent Systems 15(2), 32–41 (2000)
7. Cook, D.J., Holder, L.B.: Graph Mining. Wiley-Interscience, John Wiley Sons, Inc., Hoboken, New Jersey (2007)
8. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to Algorithms, 2nd edn. MIT Press and McGraw-Hill (2001)
9. Gabor Csardi igraph: IGraph class. R package version 0.1.1 (2005), http://cneurocvs.rmki.kfki.hu/igraph
10. Gentleman, R., Whalen, E., Huber, W., Falcon, S.: Graph: A package to handle graph data structures. R package version 1.10.6 (2006)
11. Getoor, L., Diehl, C.P.: Link Mining: A Survey. SIGKDD Explorations 7(2), 3–12 (2005)
12. Getoor, L., et al.: Link Mining: a new data mining challenge. SIGKDD Explorations 5(1), 84–89 (2004)
13. Guest, S.D., Moody, J., Kelly, L., Rulison, K.L.: Density or Distinction? The Roles of Data Structure and Group Detection Methods in Describing Adolescent Peer Groups. Journal of Social Structure 8(1), http://www.cmu.edu/joss/content/articles/volindex.html (viewed at July 28, 2007)
14. Gomez-Perez, A.: Ontological Engineering with examples from the areas of knowledge management, e-commerce and the semantic web. Springer, London (2005)
15. Frank, O.: Statistical estimation of co-offending youth networks. Social Networks 23, 203–214 (2001)
16. i2 Analyst‘s Notebook. Cambrige, UK, i2 Ltd. (2006)
17. Xu, J., Chen, H.C.: Fighting Organised Crimes: using shortest-path algorithms to identify associations in criminal networks. Decision Support Systems 38(3), 473–487 (2003)
18. Xu, J., Chen, H.C.: CrimeNet Explorer: A Framework for Criminal Network Knowledge Discovery. ACM Transactions on Information Systems 23(2), 201–226 (2005)
19. Lapsley, M. and from October 2002 B.D. Ripley RODBC: ODBC database access. R package version 1.1-7 (2006)
20. McCue, C.: Data Mining and Predictive Analytics Intelligence Gathering and Crime Analysis. BH Press Elsevier Oxford, England (2007)

21. Marcus, S.M., Moy, M., Coffman, T.: Social Network Analysis In Mining Graph Data. Cook, D.J., Holder, L.B. (eds.), John Wiley &Sons, Inc., Chichester (2007)
22. Mena, J.: Investigative Data Mining for security and criminal detection, Butterworth Heinemann, US (2003)
23. Moy, M.: Using TMODS to run the best friends group detection algorithm. 21st Century Technologies Internal Publication (2005)
24. Nooy, W.d., Mrvar, A., et al.: Exploratory Social Network Analysis with Pajek. Cambridge University Press, New York (2005)
25. Olay Bursa Local Newspaper, Technological tracking to criminal groups, http://www2.olay.com.tr/blocks/haberoku.php?id=5990&cins=Spot%20Bursa (viewed on December 19, 2006)
26. Police News Portal: Polis Haber Operation 'Cash' By Police (2006), http://www.polis.web.tr/article_view.php?aid=3666 (viewed on July 31, 2007)
27. R Development Core Team R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2006), http://www.R-project.org ISBN 3-900051-07-0
28. Scott, J.: Social Network Analysis. SAGE Publications, London (2005)
29. Senator, T.E.: Link Mining Applications: Progress and Challenges. SIGKDD Explorations 7(2), 76–83 (2005)
30. Sentient Data Detective, Sentient Information Systems (2007), http://www.www.sentient.nl/ (Viewed at July 31, 2007)
31. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications, p. 266 (1994)
32. Zaman National Newspaper, Police tracked down 63 crime groups with new technology help on 9th of January 2007 (2006), http://www.zaman.com.tr/webap-tr/haber.do?haberno=437444 (Viewed at July 31, 2007)

Applying Case-Based Reasoning and Expert Systems to Coastal Patrol Crime Investigation in Taiwan

Chung C. Chang and Kuo H. Hua

Department of Information Management, Chinese Culture University,
Yang Ming Shan, Taipei, Taiwan, 111
zcz@faculty.pccu.edu.tw

Abstract. Enhanced decision-making quality of crime investigation can ensure effective implementation of a nation's law-enforcement tasks across the board and contribute positively to its overall security and development. This study proposes the idea of applying case-based reasoning (CBR) and experts systems (ES) in conjunction with computer-assisted instruction (CAI) to crime investigation. It delves into illegal smuggling of immigrants in coastal patrol for case studies. Ocean territory law enforcement is the cornerstone of national security. Effective ocean territory law enforcement helps enhance overall national security and development. This study takes issues regarding illegal smuggling of immigrants found in coastal patrol as subject of case studies. It classifies and analyzes patterns of illegal smuggling for the purpose of establishing a crime investigation system (CIS) against illegal smuggling.

Keywords: case-based reasoning (CBR), experts systems (ES), computer-assisted instruction (CAI), crime investigation, coastal patrol.

1 Introduction

Case-based reasoning (CBR) is an inference method frequently used in the area of artificial intelligence (AI) and data mining. Currently CBR is extensively applied to the field of business, medicine, science, sociologies, and so forth [1, 2, 3, 4]. Researchers have proposed the basic concept of CBR arguing that new cases can be solved through the most similar case by editing it to line it up with current cases [5, 6]. So far we haven't been able to effectively apply AI tools to the field of crime investigation. This study employs CBR method in conjunction with the experts systems (ES) to structure a crime investigation system (CIS). Through effective incorporation of computer-assisted instruction (CAI), an integrated education and training system is offered for training of investigators and for effectual solution of crime investigation problems. Enhancement of quality of crime investigation decisions can ensure effective implementation of a nation's law-enforcement tasks across the board and contribute positively to its overall security and development.

2 Characteristics of Criminal Activities

According to the explanation of the dictionary of criminology, the modus operandi (MO) of a crime refers to the offender's behavioral pattern, preparation method and

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 161–170, 2008.

criminal approach. Collection, storage and classification of criminal approaches help identify criminal characteristics and behavioral patterns. Therefore, based on CBR this study categorizes and stores criminal characteristics and behavioral patterns. The same criminal approach is often present among cases that appear to be unrelated. According to the studies of criminology, three characteristics of MO are identified [7, 8]:

1. Existence: Criminal offenders will try to cover the traces of criminal activities in the crime scene. Yet criminal offenders only try to avoid tangible evidences such as fingerprints and footprints. They will always leave traces of intangible criminal approaches.
2. Repetitiveness: Also termed habitualness. Most habitual offenders tend to repeat the same criminal approach. From the perspective of psychology, criminal offenders will continue to repeat the same criminal approach as long as they can attain the goal of their offenses. This is what we call the repetitiveness of MO.
3. Consistency: Criminologists believe the personalities, characters and special habits of habitual or career offenders are all different. As they continue to repeat certain criminal activity, a consistent MO will begin to take shape.

From the discourses above we learn that the criminal activities of habitual or career offenders can be characterized by the following descriptions: (1) Not easy to change; (2) Repeated practices; (3) Presence of personal characteristics; (4) Inclined to fixed MO. Therefore, criminal offenders will always reveal their criminal approaches in intangible ways. Their criminal approaches differ because of differences in personal or group behaviors. Once we identify the criminal approach or MO of an individual or group, we can effectively narrow the scope of investigation.

CBR is derived from machine learning in AI. It is classified as a technique of learning from analogy. Through establishment of a criminal case base for a specific area, the criminal approaches and behavioral patterns of the criminal field can be collected, stored and classified. CBR is in line with many requirements of knowledge-based systems (KBS), and can be successfully used in KBS. CBR can effectively help enhance the overall criminal investigation ability and establish a comprehensive investigation concept for the purpose of maintaining societal security and eradicating crimes.

3 Research Motives and Objectives

Surrounded by oceans, Taiwan has a coastline of 1,653km. Its ocean territory includes inner waters, territorial seas, neighboring areas and exclusive economic sea areas. The area of the inner waters, territorial seas and neighboring areas has already amounted to 106,804km^2 [9]. From the data above we learn that Taiwan is an island country. How to prevent illegal entry and smuggling is an urgent issue.

Therefore, this study takes illegal smuggling issues in ocean territory law enforcement for case study and based on the method of CBR classifies and analyzes patterns of smuggling. The purpose is to establish a crime investigation system against smuggling in order to narrow down the scope of investigation. The reasons for applying CBR to crime investigation are as follows:

1. So there will be cases to follow: Crimes take place not by accidence. Analysis of criminal cases before will help with crime investigation.
2. Variations of cases are significant: Differences between two cases can be very significant, so much so that it makes identification of a common rule impossible.
3. Case integration: Through the data base we can integrate paperwork filed from the past and store it in the form of electronic file.

MO of crimes varies because of differences in time, place, individual or group behavior. Therefore, we incorporate the inference outcome of the CBR system with the judgment of ES. Due to the fact that this ES accumulates numerous expertise and research outcomes for computer system application [10, 11, 12, 13, 14], it will be able to provide effective, accurate advises for a particular case, and the investigators will be able to develop understanding of the crime scene and overall profile of criminal offenders in a timely manner for effective solution of a criminal case. CBR and ES can be employed to enhance the crime investigation ability and establish a comprehensive investigation concept.

Vastly diversified MO and criminal schemes often prevent investigators from taking best measures during the crucial hours. Therefore, organization, classification, storage, transfer, and utilization of knowledge and experiences of past cases through information technology can help enhance crime investigation ability that leads to effective solution of difficult cases. This study takes issues regarding illegal smuggling of immigrants found in coastal patrol as subject of case studies. It designs a CIS and an investigation knowledge sharing mechanism. Through the network platform, investigators can provide comprehensive recommendations for current investigations and establish an overall investigation concept. This study can help attain the following goals:

1. Through CBR to establish a case base of past cases in accordance with MO. This case base can recommend investigation methods for current cases.
2. Through the inference mechanism of ES to provide speedy, effective, and accurate advises for the case. Investigators will be able to develop understanding of the crime scene and overall profile of criminal offenders in a timely manner for effective solution of the criminal case.
3. To provide training and education. Professional knowledge and experiences of field experts, information of the legal requirements, crime investigation knowledge and investigative skills are incorporated. Through the user's interface, new investigators can learn quickly through this system.
4. To establish and develop this system via web-based approaches. Investigators have convenient access to this system through the Internet.

4 Prototype of the Crime Investigation System

To prove the feasibility of this CIS, this study designs and produces a prototype of coastal patrol CIS to simulate the actual system operation and test possible problems that the system may experience.

System users are divided into two categories: general users (the investigators) and system managers. Through the user's interface, general users input related information

and the system will in accordance with the internal decision-making rules output expert experiences of the knowledge to the investigators. Since this ES accumulates numerous experiences and research findings for system operation, it will be able to provide effective, accurate advises for the particular case. Investigators will be able to develop understanding of the crime scene and overall profile of criminal offenders in a timely manner for effective solution of the criminal case.

4.1 System Framework

The system framework of this study consists of three major modules: the case-based reasoning subsystem, the anti-smuggling duty planning subsystem and the CIS training and education subsystem. The case-based reasoning subsystem includes criminal type analysis, criminal characteristic/attribute establishment module, case input, case

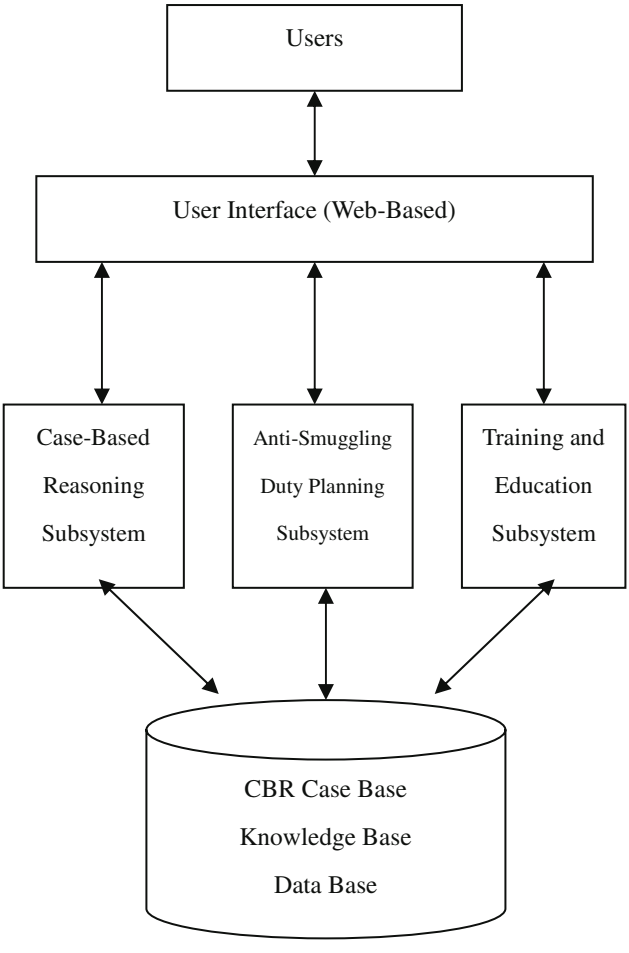


Fig. 1. The system framework of CIS

comparison, solution adaptation and outcome analysis. The anti-smuggling duty planning subsystem includes anti-smuggling duty planning and recommended measures. The CIS training and education subsystem includes case analysis training, duty planning training, related investigation theories, legal provision inquiry and online tests. The system framework is shown in Figure 1.

4.2 Case-Based Reasoning Subsystem

In this module, investigators can input smuggling crime characteristics and approaches as shown in Figure 2. The CBR case base of smuggling crime characteristics and approaches is established, which searches for and provide most similar cases through case comparison for crime investigation. Further, the similarity and weight of an attribute can be adjusted to enable the case inference mechanism of this CBR system to approximate the actual characteristics and approaches of the crime. The case output screen is shown in Figure 3. The CBR system displays the deducted criminal elements of the case to provide solution for the new problem.

In the CBR working process of the CIS, investigators organize the collected information, such as tips of informants, surveillance record, radar-identified suspicious vessels and suspicious individuals, vehicles and objects on the bank, into a comprehensive record and through the user's interface input the information into the CIS and retrieve most similar cases from the CBR case base for investigation reference. When there is a new problem for which no solution is available, following the evaluation a new case will be generated, and CIS managers will conduct rear-end CBR case base update for adjustment to possible future problems.

CBR similarity function is computed via CBR Works 4.3. Before input of various criminal attributes the data type and scope of each attribute must be first defined. Quantitative data types are set as real numbers whose scopes are defined in accordance with each indicator. Qualitative data types are set as string types. The similarity of

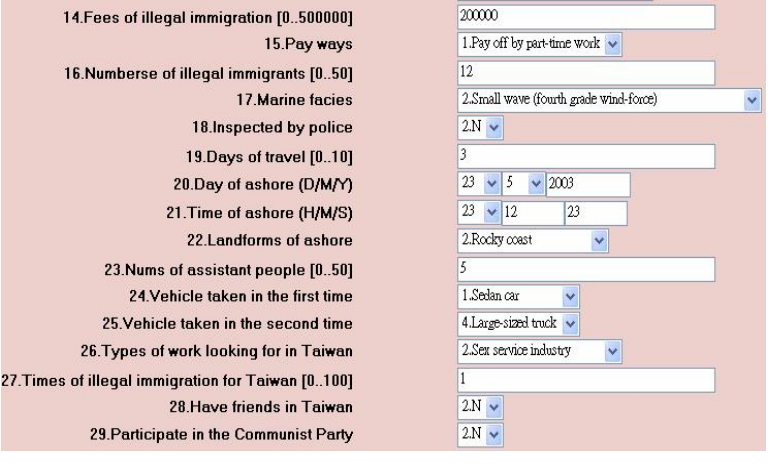


Fig. 2. The input screen of smuggling crime characteristics

	Your Question	Illegal immigration1 (63%)	Illegal immigration10 (53%)
01.Name	No Entry	'黃oo'	'林oo '
02.Sex	2.Female	2.Female	2.Female
03.Age	No Entry	24	16
04.Ancestral home	GUANGDONG	GUANGDONG	SICHUAN
05.Family condition	1.Poor	1.Poor	2.Well-fixed
06.Background	2.Junior high school	2.Junior high school	3.Secondary technical school
07.Crime record	2.N	2.N	2.N
08.Motive	3.Be-cheated to Taiwan	4.Seek part-time job in Taiwan	3.Be-cheated to Taiwan
09.Departure	Fuqingshi	Pingtan	Pingtan
10.The first time to take boat	1.Wooden boat of China	1.Wooden boat of China	2.Iron shell ship of China
11.The second time to take		4.Large-sized FRP ship of	2.Iron shell ship of

Fig. 3. The output of similar case from the case-based reasoning subsystem

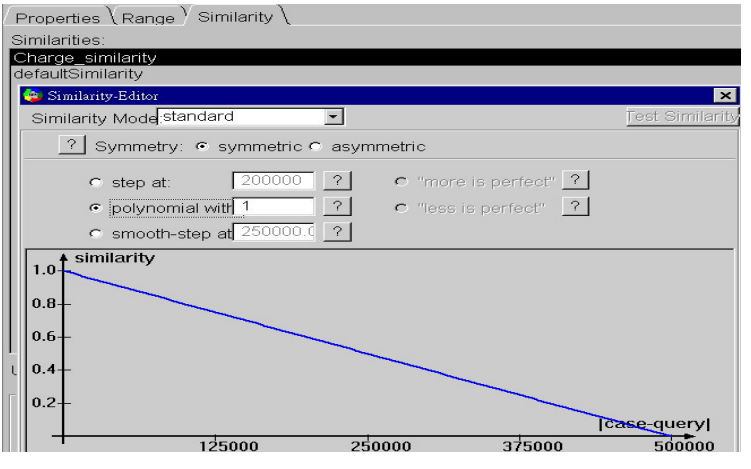


Fig. 4. The screen of default qualitative attribute similarity matrix

quantitative data is defined in accordance with each indicator, while the similarity function of qualitative data is defined in accordance with the criminal characteristic value. In the similarity function of quantitative data, the horizontal axis is the difference

between two case attribute values, and the vertical axis is the degree of similarity (the value ranging from 0-1). After the similarity values of the specific points are given, we can obtain the qualitative attribute similarity matrix as shown in Figure 4.

4.3 Anti-smuggling Duty Planning Subsystem

The knowledge acquisition process of the anti-smuggling duty planning subsystem is as follows: After interview between knowledge engineers and smuggling investigation experts, the knowledge is organized as the basis of this study's knowledge acquisition, knowledge representation and knowledge inference. There are many different kinds of knowledge. It is easier for knowledge engineers to process explicit knowledge. Tacit knowledge requires long-term accumulation for conversion of knowledge into deep knowledge. The field of criminal investigation is filled with tacit knowledge and explicit knowledge, such as the know-how in the crime scene and the ability to dissect a criminal offence.

In the anti-smuggling duty planning subsystem, the knowledge base of the rear-end ES can be established. This ES can supplement the CBR mechanism. Due to the fact that criminal characteristics and approaches will be transformed into different types as a result of time and social shifting, we need to structure an anti-smuggling duty planning subsystem that provides recommended duty arrangement for investigators.

Take smuggling for example. According to external changeable and unchangeable factors such as different cases, time, climates, and topographic features, different variables, such as security characteristics of the responsible area, hydrological data, seasonal factors, sea state changes, strategic waters and controlled vessels, must be taken into careful consideration as basis for duty assignment planning and recommendation. Investigators, as a result, will be able to carry out their investigatory missions at the best time and location in order to save human resources and increase the case-solution ratio.

In this study, the ES utilizes rule-based knowledge for expression. This method is widely applied to other ES. For instance, MYCIN utilizes backward chaining inference of rule-based knowledge to diagnose blood infectious diseases. Knowledge of the crime investigation field also requires inference. If this kind of field knowledge is expressed through semantic networks, there will be development difficulty. This study therefore recommends utilization of rule-based knowledge for design of this ES.

4.4 CIS Training and Education Subsystem

This subsystem provides the following crime investigation training and education functions:

1. Case analysis training: Utilizing the CIS framework structured in this study, this module explains how to effectively integrate and apply CBR and crime investigation. CBR Works 4.3 is employed as the inference mechanism. The software itself has the training and education function. Investigators through input of crime characteristics and approaches search for most similar cases from the case base, which in conjunction with solution adaptation, help with the crime investigation.

2. Duty planning training: The ES of this study expresses through rule-based knowledge. This method has been widely applied to other ES, because it has the training and education function within itself. Through the rule-based knowledge, users get to understand why the system makes such judgments and choices and becomes familiar with the knowledge base establishment of the rear-end ES. This system teaches users how to conduct duty assignment planning and make recommendations.
3. Related investigation theories: Related investigation theories include investigation knowledge frequently used in coastal patrol, such as secret cabins of fishing boats, position of secret compartments of fishing boats, and investigation skills for identifying secret cabins and secret compartments of fishing boats. Diagrams and motion pictures are employed to show possible position for establishment of secret cabins and secret compartments on fishing boats. Users can quickly understand the structure of a fishing boat and the spaces that can be restructured to hide smuggled substances or illegal immigrants. Through motion pictures, investigators are shown how to utilize high-tech security devices, such as metal detector, snake-pipe camera and microwave density detector, to detect secret cabins and secret compartments on fishing boats.
4. Legal provision inquiry: Training and education on legal provisions familiarizes users with all the legal provisions involved in a particular crime. The fact that users can inquire about the laws any time they need to helps them comply with the post-investigation procedures. System functions include file search, data directory and data management.
5. Online tests: The purpose of online tests is to help with curricular development in order to attain the goal of the program and enhance learning and teaching quality. Therefore, we add the coastal patrol online tests to the training and education in order to gauge users' understanding of the courses after they study the training and education materials. Users' answers to the questions will serve as the basis for system maintenance and update.

Traditionally, crime investigation experiences are passed down via apprenticeship as shown in Table 1. Through numerous duty assignments, word of mouth and body

Table 1. CIS training and education vs. traditional duty assignment instruction

CIS Training and Education	Traditional Duty Assignment Instruction
Computer	Senior officer
Visual sense, images, audio sense	Word of mouth, hand-on practice during duty assignments, body languages
Interaction with the system	Listening, watching, learning
Mouse, keyboard, screen	Listening, speaking, writing, body languages, familiarization

languages, senior officers share their experiences with their junior counterparts. The CIS training and education system can improve this situation. Through advanced information technology, related teaching materials, courses and experiences can be incorporated into multimedia interactive teaching materials that enable users to familiarize themselves with knowledge of the field more effectively and quickly.

5 Conclusions and Future Works

CBR methods in conjunction with ES are employed to enhance crime investigation ability in order to maintain societal safety and eradicate crimes. Through continuous renewal, the domain knowledge stays updated. CBR training helps users quickly learn how this CIS determines cases of the crime field and how to conduct solution adaptation. Through this training, users' judgment will become closer to the CIS outcome, and the crime-solution rate can be enhanced. Contributions of this study include:

1. This study is the first to introduce CBR to coastal patrol crime investigation in Taiwan. Through the case base established, the MO's of stowaways are classified and analyzed for establishment of a crime investigation system against smuggling in order to narrow down the scope of investigation.
2. ES is applied to duty planning for establishment of an ES prototype for coastal patrol duty planning. It provides speedy, effective and accurate duty planning to enable the investigators to develop understanding of the crime scene and overall profile of criminal offenders in a timely manner for effective solution of a criminal case.
3. Related training and education functions are integrated. Professional knowledge and experiences of domain experts, information of the legal requirements, crime investigation knowledge and investigative skills are incorporated. Through the interactive interface, new investigators can learn quickly through this system.
4. Enhances overall investigative ability. In the face of organizational criminal MO's, it provides an integrated crime investigation system and investigative knowledge sharing mechanism. Through the network platform, it provides for investigators comprehensive recommendations and helps them establish an overall investigation concept.

This study so far has produced a crime investigation system prototype. Suggestions for future studies are as follows:

1. Intensify system computation efficiency: As the cases in the case base continue to accumulate, the system will slow down in similarity computation, and its efficiency will be affected. Therefore, we need to intensify system computation efficiency in order to solve new problems quickly.
2. Add characteristics or attributes of actual crimes: During the operation of system prototype, many crime investigation characteristics or attributes were not added due to confidentiality. In actual system operations in the future, these investigative attributes can be added to enhance the usability of the system.

3. Information security issues: Due to the fact that this study uses web-based interface, information security issues must be taken into careful consideration. Criminals may break into this system to learn about criminal MO's and actual investigation activities. For online system operations, website information security issues must be addressed so the system will not become a reference to criminals.
4. System development for other field of crimes: This research studies the cases of stowaway smuggling in coastal patrol. Characteristics or attributes of other crimes such as narcotics, firearms, theft and robbery are all different. Development of investigation systems for other crimes will help enhance the overall investigative ability.

References

1. Aha, D.W.: The Omnipresence of Case-Based Reasoning in Science and Application. Knowledge-Based Systems 11(5-6), 261–273 (1998)
2. Vallespi, C., Golobardes, E., Marti, J.: Improving Reliability in Classification of Microcalcifications in Digital Mammorgrams Using Case-Based Reasoning. In: Escrig, M.T., Toledo, F.J., Golobardes, E. (eds.) CCIA 2002. LNCS (LNAI), vol. 2504, pp. 101–112. Springer, Heidelberg (2002)
3. Vilcahuaman, R., Melendez, J., de la Rosa, J.L.: FUTURA: Hybrid System for Electric Load Forecasting by Using Case-Based Reasoning and Expert System. In: Escrig, M.T., Toledo, F.J., Golobardes, E. (eds.) CCIA 2002. LNCS (LNAI), vol. 2504, pp. 125–138. Springer, Heidelberg (2002)
4. Roth-Berghofer, T.R., Cassens, J.: Mapping Goals and Kinds of Explanations to the Knowledge Containers of Case-Based Reasoning Systems. In: Muñoz-Ávila, H., Ricci, F. (eds.) ICCBR 2005. LNCS (LNAI), vol. 3620, pp. 451–464. Springer, Heidelberg (2005)
5. Schank, R.C.: Dynamic Memory: A Theory of Learning in Computers and People. Cambridge University Press, New York (1982)
6. Kolodner, J.: Case-Based Reasoning. Morgan Kaufmann Publishers, California (1993)
7. Lin, G.H.: Theory of Crime Investigation. Central Police Academy, Taoyuan, Taiwan (1998)
8. Schmalleger, F.: Criminology Today. Prentice Hall, New Jersey (1996)
9. Coast Guard Administration of Taiwan: Duties and Regulations of Coast Guard. Shan Ming Publishers, Taipei (2001)
10. Jackson, P.: Expert Systems. Addison-Wesley, Harlow (1998)
11. Durkin, J.: Expert Systems: Design and Development. Macmillan Publishing Company, New York (1998)
12. Duan, Y., Edwards, J.S., Xu, M.X.: Web-Based Expert Systems: Benefits and Challenges. Information and Management 42(6), 799–811 (2005)
13. Fuller, W.: Network Management Using Expert Diagnostics. International Journal of Network Management 9, 199–208 (1999)
14. Expert Systems at Stage IV of the Knowledge Management Technology Stage Model: the Case of Police Investigations. Expert Systems with Applications 31(3), 617–628 (2006)

Integrating Data Sources and Network Analysis Tools to Support the Fight Against Organized Crime

Luigi Ferrara[1], Christian Mårtenson[1], Pontus Svenson[1], Per Svensson[1,*], Justo Hidalgo[2], Anastasio Molano[2], and Anders L. Madsen[3]

[1] Dept. of Decision Support Systems, Swedish Defence Research Agency, SE 164 90 Stockholm, Sweden
[2] denodo technologies Europe c/ Alejandro Rodriguez, 32 Madrid 28039, Spain
[3] HUGIN Expert A/S, Gasværksvej 5, DK 9000 Aalborg, Denmark
{luigi.ferrara,christian.martenson,pontus.svenson, per.svensson}@foi.se, {jhidalgo,amolano}@denodo.com, anders@hugin.com

Abstract. We discuss how methods from social network analysis could be combined with methodologies from database mediator technology and information fusion in order to give police and other civil security decision-makers the ability to achieve predictive situation awareness. Techniques based on these ideas have been demonstrated in the EU PASR project HiTS/ISAC.

1 Introduction

The serious criminal threats facing society today require new methods for modelling and analysis. In fact, civil security decision makers, analysts and field operators fighting organized crime and terrorism across the European Union all need front-line integrated information collection and management technologies to support their cooperative work. Their adversaries are no longer organized in hierarchical structures, but instead consist of individuals and groups that are loosely organized in "dark networks" [1]. They stage attacks or set bombs against unprotected civilians, or seek to influence crowds of legitimate demonstrators so that critical riot situations occur.

In order to construct data analysis and other decision support systems that take account of these new factors, new and powerful methods and techniques from several technological domains need to be brought together and integrated.

1.1 Cross-Border and Cross-Agency Interoperability

To achieve the necessary cross-border and cross-agency interoperability, models and methods for secure sharing of information will have to be based on integrity and ownership across the information-sharing network, including dynamically modifiable role-based access rights, technology for dealing with heterogeneous data schemas and protocols, a service-oriented system architecture based on data services for sharing information, and new analysis tools to support operations during stationary as well as mobile activities.

* To whom correspondence should be addressed.

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 171–182, 2008.

1.2 Intelligence Analysis Based on Information Fusion

Fundamentally uncertain intelligence information has to be interpreted, integrated, analysed, and evaluated to provide situational awareness based on information fusion, in particular situational assessment and threat assessment methods. Relevant intelligence information originates from many sources, some of which are well-established infrastructure sources, others may be secret human intelligence information sources, some are open or public sources like mass media or the Internet [2], yet others are sensors and other physical devices of many kinds [3]. Potentially relevant data from such sources need to be stored in databases for later proactive reanalysis.

1.3 The EU PASR Project HiTS/ISAC

In the recently completed HiTS/ISAC project (EC SEC5-PR-113700) [32], financed by the EU Preparatory Action for Security Research (PASR) programme[1], environments and tools have been created for collaboratively solving a large class of social network interaction problems in law enforcement intelligence analysis.

The HiTS/ISAC problem-solving environment for interoperability and situation awareness has been demonstrated and assessed using realistic scenarios set up in cooperation with law enforcement authorities from several EU member states. The project was concluded by demonstrating a complete problem-solving environment to the project's end-user representatives using a fictitious organized-crime scenario. In that application the project showed how authorities may interoperate with information security over the network and illustrated how law enforcement authorities may cooperatively develop and share mission-critical information across national borders.

This paper deals with intelligence analysis aspects of the HiTS/ISAC demonstration system.

1.4 Structure of the Paper

A data analysis environment and toolset capable of dealing with social network analysis (Ch.2; [4]) and visualization tasks involving partly uncertain data was created by the project. Requirements on such environments are discussed in Ch. 3. The analysis system was built by combining several open-source network algorithm and visualization software packages [4][5] with a COTS (commercial-off-the-shelf) system for Bayesian belief network (BBN) modelling, embodying modern concepts and methods for management of uncertain information (Ch. 4.3; [6][7]). In addition, the use of COTS software implementing emerging database mediator technology (Ch. 4.1; [8]) made it possible to connect in a non-intrusive way, organizationally and geographically distributed and heterogeneous data sources into a single, homogeneous and secure virtual system. An architectural overview of this system is given in Ch. 4.2, below. Ch. 5 concludes the paper.

[1] Now superseded by the 7th Framework Programme for European Research 2007-13, which for the first time includes a Security section.

2 Methods for Coping with the Threats to European Security

It can be argued whether the notion of "organized crime" is appropriate for today's loosely connected networks of criminals. The new threats to European security typically come from terrorism and other large-scale criminal activities, carried out by individuals and groups that are loosely organized in "dark networks" [1]. Such networks are advantageous from a criminal's point of view since they reduce the risk of detection during planning and preparation phases. A further difficulty for law enforcement agencies is that not all actors are known in advance – the network may involve individuals without criminal records or known connections to extremist organizations.

The papers [1][9] provide examples of social network analysis in anti-terrorism applications and indicate both usefulness and some limitations of social network analysis as a basis for quantitative methods for situation awareness and decision-making in law enforcement applications. The paper [1] discusses the organizational structure of certain drug trafficking, terrorism, and arms-trafficking networks, showing how some of them have adapted to increased pressure from states and international organizations by decentralizing into smaller units linked only by function, information, and immediate need. Another interesting application of social network analysis to terrorist networks is given by [10]. In that paper, the author discusses methods for estimating the vulnerabilities of terrorist networks.

In serious-crime analysis applications, networks of relations between people, in some cases very large ones, will thus have to be set up: who knows whom, who has family relations with whom, as well as who met whom where and when, or who phoned whom when, and so on. Figuring out nested business connections across the known set of individuals or organizations is a closely related issue.

2.1 Social Network Analysis

Social network analysis (SNA) [11], is a family of methods that support statistical investigation of the patterns of communication within groups. Social scientists use these networks to analyse, *e.g.*, families, organizations, corporations, or Internet communities. The basis of the methodology is the assumption that the way that members of a group communicate with members of another group reveals important information about interesting properties of the group.

2.1.1 Structural Analysis

The emphasis in social network studies is on relations between individuals and/or groups of actors. It is sometimes referred to as *structural analysis*. In order to study the structural properties of a group, it is necessary to model it mathematically. This is most naturally done by constructing a *graph* or *network* representing the relationships within the group. Each member of the group is mapped to a node in the graph, and edges between nodes are introduced if the corresponding members communicate. Most edges link exactly two nodes; graphs where multi-edge relations are allowed are called *hypergraphs*. A hypergraph can always be embedded in an ordinary graph by introducing an extra node for each relation that involves more than two nodes.

For example, several studies, such as [12], of the citation and collaboration networks of scientists have been carried out. In these, the network of interest is the one where there is a link between all individuals who have co-authored a paper. In order to avoid having to handle hypergraphs, additional nodes are introduced for each paper, and binary relations between papers and their authors are introduced. If we are studying collaboration networks, this leads to a *bipartite* graph, where there are two different kinds of nodes, and no edge links two nodes of the same type.

An analogous example from the law enforcement domain of interest here might be that we need to model individuals who have met. In a bipartite graph of people and meetings, we can represent information about which particular meeting two specific persons attended.

2.1.2 Weights and Measures

In addition to including several nodes, edges can also be extended to include a weight or probability. This is used to model, for example, the maximum amount of information that can flow between two nodes, or to indicate the certainty with which we know that the edge is actually present in the network.

There are several important measures that can be used to characterize a network. Perhaps the simplest is to count the number of edges that different nodes have. This can be seen as a measure of the popularity of a node, and is one of the methods that are used by web search sites such as Google to rank search results [13]. Relying on the number of edges alone is not always sufficient, however. Better measures are obtained by looking at the amount of information that flows through a node. Such measures are called centrality measures. The two most important centrality measures are the *betweenness centrality* and *max-flow centrality*. The high computational complexity of the max-flow centrality problem [11][14] makes it necessary to also consider approximations to it.

Sociologists are often interested in actors that control the interaction between different groups. Such nodes are called “liaisons”, “bridges”, or “gatekeepers”, and they can also be found by calculating the centrality measures.

2.1.3 Statistical Analysis of Very Large Networks

Recently, many physicists and computer scientists have become interested in network analysis. This has led to an increased emphasis on studying the statistical properties of very large networks, such as the internet, biological food webs, and even infrastructure networks (see [15] for an overview). This influx of people to the field has also led to several new approximate algorithms with which important properties may be computed [14][16][17].

2.1.4 The Need for Management of Uncertainty

Intelligence representation languages and systems need the ability to express and reason with incomplete and uncertain information. Representation, management, and categorization of uncertainty in order to enable a machine to reason about potential relations are complex tasks. These are scientifically studied in the field of information fusion [27] which provides methods for reasoning about information arising from several different uncertain sources (see, *e.g.*, [20]).

Bayesian belief networks (BBN) [6] is one such uncertainty modelling and information fusion methodology used to represent and exploit uncertain causal relations between several variables. The BBN methodology has several potential areas of application within the intelligence domain, for instance for detecting threatening behaviours by insiders [21], for probabilistic assessment of terrorist threats [7], and for anti-terrorism risk management.

3 Requirements on Law Enforcement Problem-Solving Environments

The HiTS/ISAC project strives to contribute to a deeper awareness and understanding of modern methodological opportunities among European law enforcement authorities. These include on-demand, real-time problem solving based on scientifically sound methods of often large-scale data analysis. Organizationally, one needs to move away from "closed-room" approaches into collaborative working styles. Not only is trans-national collaboration needed between authorities in different security-related areas, such as police, coast guard, and customs services, but in order to enable effective use of modern analytical techniques and problem-solving methods there is a clear need also for cross-professional collaboration and involvement of mathematically trained analysts. A paper discussing the need for such changes in organizational culture, written from the perspective of a senior analyst is [18]. Another interesting study in the law enforcement investigative analysis area is [19].

Confidentiality, integrity and availability of information and communication systems need to be well protected [3]. Data from many different sources as well as aggregated or otherwise partially processed information, which could be sensitive and classified, must be protected from unauthorized access and modification. Although preventive countermeasures are most important, detection of misuse and intrusion must be available to deal with various types of attacks such as insider attacks and identity theft.

3.1 The HiTS/ISAC Interoperability Platform

Although database interoperability is only one of several interoperability issues that need to be addressed in a civil security intelligence system for routine operations, it is one of high technical complexity and critical importance.

A modern approach for integration of heterogeneous data sources is to make use of mediators between data sources and those consumer applications and software tools which tap these sources [22]. Mediator systems enable automatic translation between the concepts and conventions, *schemas*, of different distributed data sources, *i.e.*, names and other characteristics of their data items as well as the semantic relationships between them, offering a virtual data layer that can be queried by consumer applications using database-like query languages (*e.g.* SQL and/or XQuery).

The HiTS/ISAC project provides prototypical mechanisms for information sharing with retained integrity and confidentiality to support role-based cooperation.

Depending on the situation the users will have easy but secure access to information and services tailored to their respective needs.

In the HiTS/ISAC demonstration system no single unifying software technology is used. Instead, different technologies are contributed by different partners, making the challenge of maintaining interoperability greater than in a homogeneous single-vendor system, but at the same time providing greater benefits due to the pooled capabilities. COTS software (Denodo Virtual Data Port and HUGIN), sometimes modified open-source software (JUNG, Prefuse, PROXIMITY, Monet), and in-house middleware developments are all part of the final system.

The backbone communication network in the HiTS/ISAC demonstration system is based on services that are available today as standard products from the project partner TeliaSonera. The system satisfies the security levels required by the Swedish police, while meeting their demands for identification, tapping and integration protection, confidentiality and security. High-capacity, secure Internet solutions based on virtual private networks (VPN) and encryption techniques, as well as telephony services fixed or mobile, are examples of products on which the solution has been based.

The workstations are standardized and may be continuously updated and monitored, this way ensuring that the right software and the right versions are always used. This is of particular importance for the software used to manage the security of the system.

4 Secure Collaborative Analysis Environment (NetSCW)

The purpose of the secure collaborative problem solving environment NetSCW is to provide a common data analysis environment managing issues of interoperability, availability and reusability of data, as well as analysis processes and results. In this environment (fig. 1), subject matter experts, data analysts, database and ICT security administrators, *etc.*, can meet to share their resources and expertise, as well as collaborate in real time. The NetSCW environment uses a Service-Oriented Architecture (SOA). The SOA concept envisions an interconnected network of producers and consumers of information and supports the development of a uniform framework for description and utilization of distributed components.

The top layer comprises various data analysis-related tools, *e.g.*, tools for analysing social network data by a single analyst or collaboratively by a group of analysts.

4.1 Denodo Virtual DataPort Data Mediation Platform

The Denodo Virtual DataPort (VDP) [8][28], a state-of-the-art data mediator system, provides a solution for accessing, querying and integrating information from any kind of digital source, from structured repositories such as databases, Web Services, and applications, *via* semi-structured sources such as dynamic Web content and delimited files, to unstructured repositories (documents, emails, forums, *etc.*). The system is able to provide a single view of the heterogeneous data stores of participating law enforcement authorities, while allowing them to remain autonomous and unchanged.

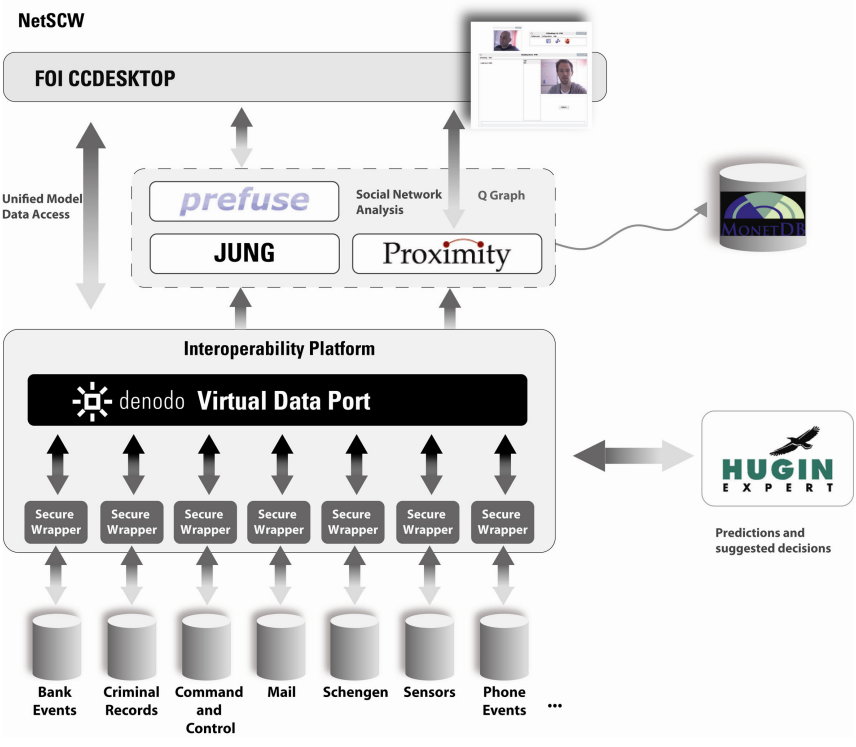


Fig. 1. Architecture of the HiTS/ISAC collaborative problem solving environment

The Denodo VDP platform combines data mediation with an advanced web automation technology that allows the exploitation of information in complex hidden web data sources. With this functionality it is possible to integrate data from such complex web sites that are being used nowadays for organized-crime related activities.

Denodo VDP provides a three-level data integration architecture, at the bottom the *data source connectivity* layer that allows integration of data views coming from different data sources (*i.e.,* employing different protocols and data formats) and isolating the complexity for accessing the information to the rest of the platform; the intermediate *transformation and enrichment* layer follows an extended relational model, able to handle both tabular and hierarchical information, and allows the combination and correlation of information by means of data views built with relational operations, such as joins, unions or selections; finally information is delivered to the consumer application through *standard interfaces* such as JDBC/ODBC, SOA/Web Services and Java APIs.

4.2 The NetSCW Collaborative Core (NetSCW/CC)

Computer Supported Collaborative Work (CSCW) [23] is an interdisciplinary research and application domain which has evolved since the mid-1980's, with contributions from social anthropology, psychology, and computer sciences. The

research focus of CSCW is Groupware, i.e., applications designed to help people involved in a common task to achieve their goals.

4.2.1 NetSCW/CCDesktop

In the HiTS/ISAC project, CCDesktop, a Java-based application, has been developed by FOI as the CSCW component of NetSCW. The CCDesktop design aspires to implement much of the desired CSCW functionality and bring it into an analytical problem solving environment. CCDesktop supplies basic functionality through a common communication interface which adds collaborative and group awareness features to existing analysis tools. In addition, CCDesktop provides a secure multimedia communication infrastructure which provides users with video, voice and text messaging facilities. The current version of CCDesktop offers the following tools: participant panel, chat board with video and voice functionality (via RTP, the Real-time Transport Protocol [31]), *simultaneous shared access* to tools for Social Network Analysis [4], as well as to middleware [5] adapting Denodo VDP output to the input requirements of the PROXIMITY graphical query language-based SNA processing tool.

A close integration of JUNG, prefuse, PROXIMITY, and the required data connection to Denodo VDP was developed by two M Sc thesis projects at FOI [4][5]. Functions for adding, merging, grouping by attribute, and removing nodes and edges did not exist in the version of JUNG that was used for the implementation, and were added. Key design requirements of these projects were:

Data collected from several different databases. A common schema is needed to reduce the analysts' mental effort.

Queries as filtering method. Working with large-scale social networks will require various filtering techniques to select appropriate subsets and reduce the volume of data.

Objects first mapped to relational data, then into graphs. Social networks consist of objects that are represented as nodes and edges, not as standard relational database tabular data.

Sets of subgraphs are combined based on given predicates. By comparing with the original graph the frequently large number of subgraphs generated by PROXIMITY/QGraph can often be greatly reduced. Using the CCDesktop system, the user may decide on a conceptual level which of these subgraphs should be merged.

Graphs easily exported. Graphs need to be readable by other programs, so they are stored either in JUNG internal format or in GraphML external format, or both.

Combining *prefuse* with JUNG. This has involved resolving several issues related to the two systems' different graph storage structures, user interaction conventions, and display functionality.

4.2.2 PROXIMITY

Social Network Analysis is a required capability of the HiTS/ISAC project. SNA can become computationally intense and the integrated database of HiTS/ISAC may potentially become quite large. Therefore, there is a need to be able to filter data and

to find interesting subgraphs, *e.g.* when one wants to find all networks connecting two criminals by either phone calls or email. The SNA query system PROXIMITY [24] provides a network filtering functionality, extended to allow integration with the other components of the HiTS/ISAC analysis system. This functionality is controlled *via* QGraph, a visual language that returns graph fragments with highly variable structure.

PROXIMITY helps human analysts discover new knowledge by analyzing complex data sets containing network-structured information, using specially developed algorithms that help manage, explore, sample, and visualize data.

PROXIMITY is a Java-based open-source software system. It uses the Monet DB, an open-source "vertical" database [25] optimized for analytical queries.

4.2.3 Prefuse

prefuse is a Java-based open-source software toolkit [26] for building interactive information visualization applications.

prefuse supports a rich set of features for data modelling, visualization, and interaction. It provides optimized data structures for tables, graphs, and trees, a host of layout and visual encoding techniques, and support for animation, dynamic queries, integrated search, and database connectivity. It has no SNA capabilities of its own but can be integrated, as demonstrated by the HiTS/ISAC project, with the JUNG framework for immediate access to SNA functionality.

4.2.4 JUNG

JUNG, the Java Universal Network/Graph Framework, [29] is an open-source software library that provides a common and extendible language for modelling, analysis, and visualization of data that can be represented as a graph or network. The JUNG library provides a variety of graph algorithms, network visualization tools, and support for dynamic graphs. It also provides a mechanism for annotating graphs, entities, and relations with metadata. This facilitates the creation of analytic tools that can examine the relations between entities in complex data sets as well as metadata attached to each entity and relation.

4.3 HUGIN

The HUGIN [6][30] software tools for Bayesian belief networks implement advanced algorithms for knowledge discovery and probabilistic reasoning. These tools consist of the HUGIN Decision Engine and the HUGIN Graphical User Interface, well-suited for developing model-based decision support systems based on Bayesian belief networks (BBNs). A BBN is an intuitive graphical knowledge representation supporting belief update in the light of observations. It consists of an acyclic, directed graph representing causal dependence relations between a set of variables representing entities of the problem domain and a set of conditional probability distributions encoding the strength of the dependence relation.

The HUGIN Decision Engine is the inference engine that takes care of the representation of models, the mathematical calculations performed as part of probabilistic inference, etc., while the HUGIN Graphical User Interface provides a graphical user interface to the functionality of the HUGIN Decision Engine.

In HiTS/ISAC a BBN model for identifying suspicious activity has been developed by knowledge engineers and domain experts in corporation. The BBN model is to be used by analysts in their everyday intelligence work as a tool to perform information fusion and analysis.

A Bayesian model for ranking suspicious bank transactions is shown in Figure 2. The model is evaluated for all transactions, resulting in a ranked list of transactions. One bank transaction, weakly suspected *a priori*, has a high rank in this list and a relatively high degree of *a posteriori* suspiciousness.

The databases used here are a border transaction database and a bank transaction database, which are assumed to reside in different countries and belong to different organizations. The BBN is fed with data from a view constructed in Denodo VDP.

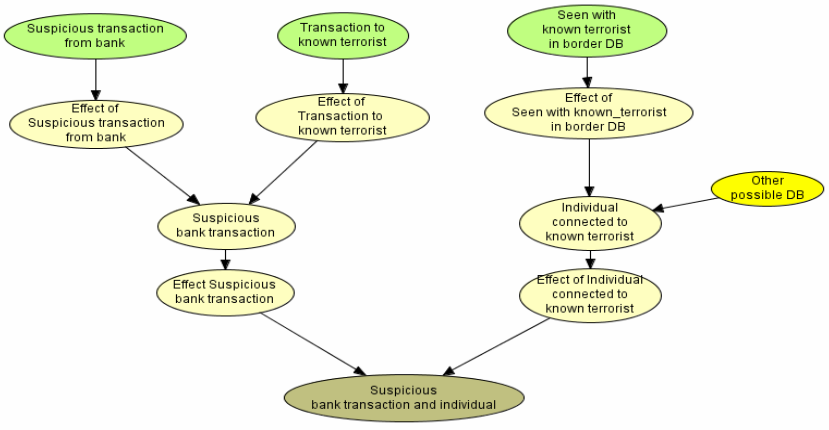


Fig. 2. Bayesian belief network used in the demonstration. Input nodes and output node are marked with extra ellipses.

For each bank transaction, the view feeds the network with information such as the degree of suspiciousness of the transaction and the names of its sender and its receiver. By comparing these names to those on the list of known terrorists, the input node "Transaction to known terrorist" is set. A view is also constructed that contains all the border transactions of the persons mentioned in any of the bank transactions. If the border database contains a border crossing of a known terrorist that occurs near-simultaneously with a crossing of a person involved in a suspect bank transaction, the input node "Seen with known terrorist in border DB" is set to reflect the time difference between the person and the known terrorist crossing the border.

5 Summary and Conclusions

HiTS/ISAC is a pre-study of interoperability and situation awareness for civil security in Europe [32]. The project has demonstrated secure network analysis environments and tools with respect to immediate applicability and user-oriented functionality.

The HiTS/ISAC project dealt with operational, methodological, and developmental issues in a demonstration scenario application. Its focus was on scientifically sound

analysis and secure management of distributed and usually uncertain information about events and relationships in organized crime and terrorist networks. The project has demonstrated how legacy databases of several authorities in different countries can be effectively integrated into a distributed problem-solving environment to provide a common user view of the combined relevant information assets of the participating authorities. Analysts working at different sites were able to work collaboratively on the same data without saving, transferring and loading files into their analysis software.

We believe that the experience gained and the lessons learned from this project can be important for determining what analysis capabilities are needed in European police intelligence work. Key technological enablers for successful use of such methodology are distributed, mediated access to large amounts of legacy data and support for real-time collaboration among analysts. Further studies are needed regarding how the operational processes used by criminal intelligence investigators and analysts need to be changed to take full advantage of the possibilities offered by new technology such as that described in this paper. In addition, a set of hard and sometimes highly controversial juridical issues need to be addressed and agreed upon. Such issues are, however, outside the scope of the HiTS/ISAC project and this paper.

Acknowledgements

The authors wish to thank their HiTS/ISAC consortium partners, in particular Saab AB, TeliaSonera, EADS, and TietoEnator Alise, for sharing their expertise in the public safety area, as well as in the design, integration, and application of the various kinds of software components discussed above.

References

1. Raab, J., Milward, H.B.: Dark networks as problems. J. Public Administration Research and Theory 13(4), 413–439 (2003)
2. Chang, K.C.-C., He, B., Li, C., Patel, M., Zhang, Z.: Structured Databases on the Web: Observations and Implications. SIGMOD Record 33(3), 61–70 (2004)
3. Popp, R., Poindexter, J.: Countering terrorism through information and privacy protection technologies. IEEE Security & Privacy, 18–27 (November/December 2006)
4. Sköld, M.: Social Network Visualization. M Sc thesis report, KTH School of Computer Science and Communication, Stockholm, Sweden (2008)
5. Asadi, H.C.: Design and Implementation of a Middleware for Data Analysis of Social Networks. M Sc thesis report, KTH School of Computer Science and Communication, Stockholm, Sweden (2007)
6. Kjaerulff, U.B., Madsen, A.L.: Bayesian Networks and Influence Diagrams. A Guide to Construction and Analysis. Springer, New York (2008)
7. Koelle, D., Pfautz, J., Farry, M., Cox, Z., Catto, G., Campolongo, J.: Applications of Bayesian Belief Networks in Social Network Analysis. In: Proc. of the 4th Bayesian Modelling Applications Workshop (2006)
8. Pan, A., Raposo, J., Álvarez, M., Montoto, P., Orjales, V., Hidalgo, J., Ardao, L., Molano y Ángel Viña, A.: The DENODO Data Integration Platform. In: Proc. 28th Int. VLDB Conf., pp. 986–989. Morgan Kaufmann, San Francisco (2002)

9. Carley, K.M., Lee, J.-S., Krackhardt, D.: Destabilizing Networks. Connections 24(3), 79–92 (2002)
10. Carley, K.M.: Estimating Vulnerabilities in Large Covert Networks. In: Proc. of the 2004 International Symposium on Command and Control Research and Technology. Evidence Based Research, Vienna, VA, USA (2004)
11. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. Cambridge University Press, Cambridge (1994)
12. Redner, S.: Citation Statistics from 110 Years of Physical Review. Physics Today 58, 49 (2005)
13. Page, L., Brin, S.: The Anatomy of a Large-Scale Hypertextual Web Search Engine. In: Proceedings of the Seventh International World Wide Web Conference, vol. 30(1-7), pp. 107–117 (1998)
14. Newman, M.E.J.: A Measure of Betweenness Centrality Based on Random Walks. Social Networks 27, 39–54 (2005)
15. Svenson, P., Mårtenson, C., Carling, C.: Complex Networks: Models and Dynamics, Swedish Defence Research Agency Technical Report FOI-R—1766—SE, Stockholm, Sweden (2005)
16. Clauset, A., Newman, M.E.J., Moore, C.: Finding Community Structure in Very Large Networks. Physical Review E70, 066111 (2004)
17. Newman, M.E.J.: Modularity and Community Structure in Networks. Proc. Natl. Acad. Sci. USA 103, 8577–8582 (2006), http://arxiv.org/abs/physics/0602124
18. Klerks, P.: The Network Paradigm Applied to Criminal Organizations: Theoretical Nitpicking or a Relevant Doctrine for Investigators? Recent Developments in the Netherlands. Connections 24(3), 53–65 (2001)
19. Gottschalk, P.: Stages of Knowledge Management Systems in Police Investigations. Knowledge-Based Systems 19, 381–387 (2006)
20. Proceedings of the International Conferences on Information Fusion 1998-2005. International Society of Information Fusion, Mountain View, CA, USA
21. Bass, T.: Intrusion Detection Systems and Multi-sensor Data Fusion. Comm. ACM 43(4), 99–105 (2000)
22. Fredriksson, J., Svensson, P., Risch, T.: Mediator-Based Evolutionary Design and Development of Image Meta-Analysis Environments. J. Intell. Information Systems 17(2/3), 301–322 (2001)
23. Ackerman, M.S.: The Intellectual Challenge of CSCW: The Gap between Social Requirements and Technical Feasibility. Human-Computer Interaction 15, 179–203 (2000)
24. Jensen, D.: PROXIMITY 4.2 Tutorial. Knowledge Discovery Laboratory, Department of Computer Science, University of Massachusetts at Amherst (2006)
25. Manegold, S., Boncz, P.A., Kersten, M.L.: Optimizing Database Architecture for the New Bottleneck: Memory Access. VLDB Journal 9(9), 231–246 (2000)
26. Heer, J., Card, S.K., Landay, J.A.: Prefuse: a Toolkit for Interactive Information Visualization. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Portland, Oregon, USA (2005)
27. International Society of Information Fusion, http://www.isif.org (accessed 2008-03-27)
28. http://www.denodo.com (accessed 2008-03-27)
29. http://jung.sourceforge.net/ (accessed 2008-03-27)
30. http://www.hugin.com (accessed 2008-03-27)
31. http://tools.ietf.org/html/rfc3550 (accessed 2008-03-27)
32. http://www.hits-isac.eu/ (accessed 2008-03-2)

Visual Analytics for Supporting Entity Relationship Discovery on Text Data

Hanbo Dai[1] , Ee-Peng Lim[1], Hady Wirawan Lauw[1], and Hweehwa Pang[2]

[1] School of Computer Engineering, Nanyang Technological University
[2] School of Information Systems, Singapore Management University

Abstract. To conduct content analysis over text data, one may look out for important named objects and entities that refer to real world instances, synthesizing them into knowledge relevant to a given information seeking task. In this paper, we introduce a visual analytics tool called **ER-Explorer** to support such an analysis task. ER-Explorer consists of a data model known as **TUBE** and a set of data manipulation operations specially designed for examining entities and relationships in text. As part of TUBE, a set of interestingness measures is defined to help exploring entities and their relationships. We illustrate the use of ER-Explorer in performing the task of finding associations between two given entities over a text data collection.

1 Introduction

1.1 Motivation

Information synthesis and analysis can be facilitated by a visual interface designed to support analytical processing and reasoning. Such an interactive visualization approach is also known as **visual analytics**[1]. In this research, we specifically focus on designing and implementing a visual analytics system to support the entity relationship discovery task that involves identifying entities and relationships from a document or a collection of documents so as to create a network of entities that are relevant to an entity relationship discovery task.

Consider the task of finding the person and organization entities that connect two terrorists from a given document collection. A domain expert will need an interactive visual tool to help in extracting entities from the documents and the relationships among these entities, judging the relevance of these entities and relationships by checking them up in documents containing them, and selecting the relevant ones to be included in the results.

For a visual analytics system to support the above retrieval task, the following system features are required.

– *Network representation of information*: Entity and relationship instances are best represented using a graph or network, especially when path and connectivity properties of these instances are to be studied and visualized along with the documents containing them.

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 183–194, 2008.

- *Interactive refinement of results*: The above retrieval task, like many others that require expert judgement, will involve much user interaction in multiple iterations. Hence the visual analytics system will have to incorporate user operations that may include or exclude entities and relationships from the retrieval results.
- *Intelligent user assistance*: Given the possibly large volume of document data and many entity and relationship instances embedded in documents, users will expect some intelligent assistance from the visual analytics system to help them gain more insight into the data. The exact form of assistance may very much depend on the task at hand. For example, entities (or relationships) may have to be ranked by their closeness to the two given terrorists so as to help user decision making.

The above are also the system features that distinguish visual analytic systems from the other visual interface systems for analyzing networks of entity and relationship instances. In social network analysis, the state-of-art visual interface systems often assume that networks of entity and relationship instances have already been identified and verified, as well as can be studied separately from the documents containing them[2,6,7]. This assumption clearly does not hold for documents which are not pre-annotated. Even if the documents are already pre-annotated, it is still challenging to determine the relevant entity and relationship instances. This often requires users to interpret text content in documents containing these instances.

1.2 Research Objectives and Contributions

In this research, we therefore aim to design a visual analytic framework for entity relationship discovery under the assumption that (a) user judgement on document content is required for identifying relevant entity and relationship instances, and (b) the discovery is an iterative process with user involvement.

Our contribution in this paper can be summarized as follows:

- We present a visual analytics framework for discovering a network of related entities found in text data. This framework consists mainly of a multi-dimensional data model and a visual interface tool for representing and manipulating entity and relationship instances.
- We design a text cube representation of the entity and relationship instances in document data. This representation, known as **TUBE**, supports semantic entity types, conceptual entity representation, inter-entity relationships and other data constructs useful for information analysis and synthesis.
- We develop a visual analytics system known as **ER-Explorer** to realize a set of user operations on a network of entities derived from a set of text documents so as to conduct entity relationship discovery.
- We illustrate our visual analytics system prototype using a case study where the entities and relationships linking two given entities can be discovered through an interactive process.

1.3 Paper Organization

We organize the rest of the paper as follows. In Section 2, we cover the related research. In Section 3, our framework for entity relationship discovery using visual analytics is presented. In Section 4, we describe the **ER-Explorer**, a visualization tool implemented based on our proposed framework. This is followed by a case study analysis in Section 5. We finally conclude the paper in Section 6.

2 Related Work

Visually analyzing social networks has been receiving growing attention and several visualization tools have been developed for this purpose. *Vister*[3] provides an environment to explore and analyze online social network, supporting automatical identification and visualization of connections and community structures. *SocialAction*[4] allows users to explore different social network analysis measures to gain insights into the network properties, to filter nodes (representing entities), and to find outliers. Users can interactively aggregate nodes to reduce complexity, find cohesive subgroups, and focus on communities of interest. However, the measures used in these systems are topological-oriented.

Xu and Chen [8] proposed a framework for automatic network analysis and visualization. Their *CrimeNet Explorer* identifies relationships between persons based on frequency of co-occurrence in crime incident summaries. Hierarchy clustering algorithm is then applied to partition the network based on relational strength.

The above systems while supporting network visualization, lack the measures for discovering associations among nodes. Their way of grouping entities is based on centrality measure or relational strength, which does not allow user judgement and may fail to group semantically identical entities.

A visual analytic system *Jigsaw*[9] represents documents and their entities visually in multiple views to illustrate connections between entities across the different documents. It takes an incremental approach to suggest relevant reports to examine next by inspecting the co-occurred entities. However, it does not use measures other than frequency of entities in documents. When the list of co-occurred entities becomes very large, it would be quite cumbersome for an analyst to find the interesting entities or documents, since considering the frequency measure alone may be restrictive. Moreover, in cases where co-occurrence relationship between entity are not semantically meaningful, the analytics ability of Jigsaw will be ineffective.

There is much research literature on path finding. *Transitive association discovery* was proposed to detect conceptual association graph in a text dataset [10]. Interestingness measures based on co-occurrence are designed. A dynamic programming algorithm was developed to compute interesting paths of various lengths from source to target entities. Document contexts of the paths are also provided. People association finding in the *ArnetMiner* project [11] also aims to detect the good associations. Since the above approaches rely on algorithms to

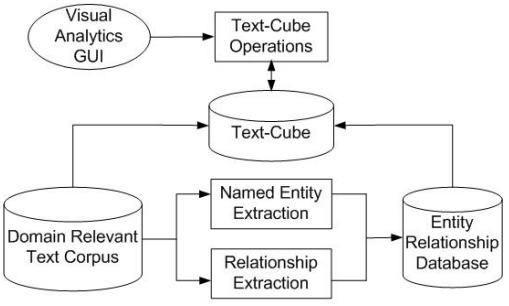


Fig. 1. System Architecture

compute paths, users have no control over the paths and entities he or she may want to explore. Moreover, semantically identical entities are also not considered.

3 Framework of Visual Entity Relationship Discovery from Text Data

3.1 Architecture of Visual Analytics Tool

As shown in Figure 1, our proposed visual analytics system architecture consists of a domain relevant text corpus on which the entity relationship discovery task is to be performed. From the text corpus, named entities and relationships will be extracted into an *entity relationship database*. A *text cube database* is then constructed from the entities and relationships. It consists of one or more text cube instances and each text cube instance is a multidimensional table with entities as dimension values and relationships as cells. Unlike a database table, a text cube also provides document evidence of the entities and relationships so as to facilitate cross checking entities and relationships with their text sources. A set of *text cube operations* are provided to manipulate the content of text cubes. These are also the data operations to be invoked by the *visual analytics GUI tool*. Users will interact with the GUI tool performing visual analysis operations on the text data without having to know the underlying text cube representations and operations.

3.2 TUBE: Text Data Cube Model and Operations

TUBE Representation. In our TUBE model, a domain relevant document collection $\mathcal{D}$ provides the raw data for analysis[12]. For each document $d \in \mathcal{D}$, the set of named objects extracted from d is denoted by $A(d)$. The entire set of named objects extracted from $\mathcal{D}$ is denoted by $A(\mathcal{D}) = \bigcup_{d \in \mathcal{D}} A(d)$. We define a mapping function $f_d : A(\mathcal{D}) \rightarrow 2^{\mathcal{D}}$ to map from a named object to its supporting document set, consisting of documents that contain the named object.

In TUBE, we introduce the notion of entity e, which is defined as a named object or a set of other entities as follows:

$$e = \begin{cases} a, & a \in A(\mathcal{D}) \\ \{e_1, e_2, \ldots, e_n\}, & e_i \textit{ is an entity.} \end{cases}$$

We say a is a component of e, $a \prec e$, if $a = e$ or $\exists e_i \in e$ s.t. $a \prec e_i$. e is said to be a *conceptual entity* if it is not a named object. The *document evidence* of e is defined as $f_d(e) = \bigcup_{a \prec e} f_d(a)$.

We now define a n-dimensional TUBE as a tuple $T = \langle S, B, M, D \rangle$. S represents the *schema* and $S = \{s_1, s_2, \ldots, s_n\}$ where s_i denotes the list of entities of dimension i. B is a *mask* with 0 or 1 values. $M = \{m_1, m_2, \ldots, m_{|M|}\}$ is a set of measures. Each m_j is associated with a measure function $mf_j()$. D represents a document collection and $D \subseteq \mathcal{D}$.

The TUBE T has $|s_1| \times |s_2| \times \ldots \times |s_n|$ cells. Each cell is denoted by c $(e_1, e_2, \ldots, e_n)$ where $e_i \in s_i$ for $1 \leq i \leq n$. Without causing any ambiguity, we may use c to denote a cell. A cell c is said to be *present* if $B(c) = 1$ or *hidden* if $B(c) = 0$. The document evidence of c is defined by $f_d(c) = \bigcap_{i=1}^{n} f_d(e_i)$. When $f_d(c)$ is not empty, we say that $e_1, \ldots, e_n$ *co-occur*. This *co-occurrence relationship* can be represented by c We also define the *named object set* of c as $A(c) = \bigcup_{i=1}^{n} \bigcup_{a \prec e_i} \{a\}$. The *support value* for a d_k in $f_d(c)$ with respect to c is defined by:

$$Sup(c, d_k) = \sum_{a \in A(c)} tf_{d_k,a} \times idf_a$$

where $tf_{d_k,a}$ is the a's frequency in d_k

$$idf_a = \frac{|\mathcal{D}|}{|f_d(a)|}$$

Given a cell c, c has a measure value $c.m_j = mf_j(c)$ derived by applying the measure function mf_j.

TUBE Operations. We have also designed a set of operations on TUBE. Given a TUBE instance $T = \langle S, B, M, D \rangle$,

- Insert operation adds an entity to a selected dimension.
- Remove operation removes an existing entity from a dimension.
- SelectCell operation assigns 0 or 1 to a specified entry in B which corresponds to a cell in T.
- Cluster operation groups a subset of entities in a specified dimension into a new conceptual entity and add this conceptual entity to that dimension.

TUBE Instances For Entity Relationship Discovery. Our entity relationship discovery uses two TUBE instances T_1 and T_2. $T_1 = \langle S^1, B^1, M^1, D \rangle$ and $T_2 = \langle S^2, B^2, M^2, D \rangle$ are 1-D and 2-D TUBE instances respectively. We initialize T_1 to have $S^1 = \{s_1^1\}$, $s_1^1 = A(\mathcal{D})$ by Insert operation. In other words, T_1 has a dimension consisting of all named objects. T_2 is initialized to have $S^2 = \{s_1^2, s_2^2\}$ where $s_1^2 = s_2^2 = A(\mathcal{D})$. In other words, T_1 is designed to maintain information about named objects and T_2 for information about the relationships of pairs of entities. Also note that any operations on one dimension of T_2 will affect the other dimension the same way.

The masks B^1 and B^2 are initialized to return 0's for all cells, making all named objects and relationships initially hidden from the network view of our visual tool.

3.3 Entity Relationships Exploration Using T_1 and T_2

Given two entities of interest known as *source entity* (s) and *target entity (t)*, a typical entity relationship discovery task would be finding interesting paths between them. Each path denoted by $e_1 \leftrightarrow \ldots \leftrightarrow e_p$ represents a chain of relationships. Each relationship denoted as $e_{i-1} \leftrightarrow e_i$, for $1 < i \leq p$, and e_1 and e_p are entities semantically equivalent to s and t respectively. Note that in this task, the relationships are non-directional. The roles of source and target are therefore exchangeable. Nevertheless, we just distinguish them for easy discussion.

Our visual tool can incrementally add named objects and relationships into the entity network presentation window as nodes and edges respectively by invoking TUBE operations on the two TUBE instances T_1 and T_2, To display an entity in the visual tool, we set the respective cell in T to have $B = 1$. To display a relationship, we set the corresponding cell in T_2 to have $B = 1$. Hiding entities and relationships can be performed in a similar way. This interactive approach to construct entity networks can be assisted by interestingness measures defined for the entity relationship discovery task.

3.4 Interestingness Measures for Entity Relationship Discovery

In this section, we define several measures to be used in T_1 and T_2 to support entity relationship discovery. For T_1, there is only one measure, i.e., $M^1 = \{m_{path\_strength}\}$. For T_2, we define $M^2 = \{m_{name\_sim}, m_{strength}, m_{d\_entity}\}$.

- $m_{path\_strength}$: the length of shortest path(s) between s and t going through an entity (a named object, since it is defined on T_1).
- m_{name_sim}: the similarity score between two entity names.
- $m_{strength}$: the relationship strength between two entities.
- m_{d_entity}: the dominance of one entity over another.

Given a cell $c(e_i, e_j)$ in T_2,

$$mf_{name\_sim}(c) = Avg_{a_u \prec e_i, a_v \prec e_j}\ NameSimilarity(a_u, a_v)$$

where *NameSimilarity* is a name comparison function that returns a value between 0 (unrelated name objects) and 1 (synonym). If e_i and e_j are conceptual entities, the measure value returned is an average of over name similarities between named objects of e_i and e_j. With this measure, we now derive a set of synonyms for an entity e_i, as denoted by

$$Synonym(e_i) = \{e_j | mf_{name\_sim}(c(e_i, e_j)) > \lambda\}.$$

The synonym entities of e_i are entities whose names are within λ edit distance from that of e_i. The function *Synonym* is helpful to detect different spellings of an entity. Grouping synonym entities together may discover new associations, since they may have different relationships with other entities.

The measure function of $m_{strength}$ for a cell $c(e_i, e_j)$ is denoted by $mf_{strength}$ $(c(e_i, e_j))$ which computes strength using *Dice Coefficient*, i.e.,

$$mf_{strength}(c(e_i, e_j)) = \log\left(1 + 2 \cdot \frac{|f_d(c(e_i, e_j))|}{|f_d(e_i)| + |f_d(e_j)|}\right)$$

The strength of a cell representing a pair of entities captures the likelihood of a relationship between them. The more documents they co-occur in, the higher the strength.

Given two entities e_i and e_j, the *d_entity* measure determines if the documents containing e_i are also those containing e_i and e_j. This happens when e_i always appears together with e_j(This implies whenever e_i appears, e_j is always there), and we say that e_j *dominates* over e_i.

$$m_{d\_entity}(c(e_i, e_j)) = \begin{cases} 1 \; if \; f_d(c(e_i)) = f_d(c(e_i, e_j)) \\ 0 \text{ otherwise} \end{cases}$$

For example, m_{d_entity}("9-11","New York") = 1 when "9-11" appears in only those documents containing both "9-11" and "New York".

For T_1, the measure $m_{path\_strength}(c(e_i))$ returns the strength of shortest path(s) between s and t going through e_i. Let $s\_path(e_i)$ denote this set of shortest paths, $m_{path\_strength}$ is defined as:

$$m_{path\_strength}(c(e_i)) = Max_{p_{ik} \in s\_path(e_i)} strength(p_{ik})$$

where

$$strength(p_{ik}) = \prod_{(c(e_x, e_y) \in p_{ik}} mf_{strength}(c(e_x, e_y))$$

When multiple shortest paths between s and t pass through entity e_i, $m_{path\_strength}(e_i)$ will take the maximum path strength among them. A large $m_{path\_strength}(e_i)$ suggests that there exists a path with edges that represent strong relationships. Hence, e_i may be a good entity to explore to establish useful linkages between s and t.

4 Visual Analytics Tool for Entity Relationship Discovery

In this section, we describe our Visual Analytics Tool, **ER-Explorer** (Entity Relationship Explorer) can be utilized. The named entity extraction in our system is performed by **BBN Identifinder** [13], which can extract entities of 24 types including *person*, *organization*, *GPE* (Geo-political entities), *date* and others. After extracting co-occurrence relationship extraction, *Lucene* is used to index all documents by their extracted named entities. The visualization part of our tool is built upon *Chisio*[1], a free Compound or Hierarchical Graph Visualization Tool based on eclipse Graphical Editing Framework.

Overview of User Interface

ER-Explorer is mainly made up of five views (see Figure 2), namely, a *Network View*, a *Document View*, a *Related Entity View*, a *Synonym Entity View* and a *Path View*.

[1] http://www.cs.bilkent.edu.tr/ ivis/chisio.html

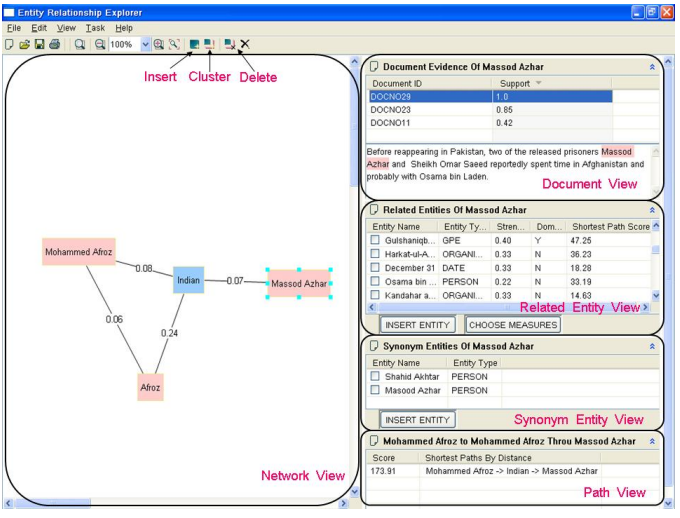


Fig. 2. ER-Explorer Interface

Network View is where the user visualizes the network and manipulates it with visual analytics operations. We visually display entities in TUBE as nodes and relationships as edges. Entities are shown as boxes in different colors associated with their entity types. Edges are weighted by the $m_{strength}$ of T_2. Each conceptual entity is visualized as a compound box drawn to enclose its component entities.

Document View shows the supporting documents of a selected entity relationship. These can be read by users to understand the context of entities and relationships. This view consists of two parts. The upper part lists the document IDs and their support values with respect to selected entity/relationship. The lower part displays the content of a document, once that document is selected. All named objects in the document semantically represented by the selected entity/relationship are highlighted.

When an entity/relationship is selected in the Network View, the **Related Entity View** displays all entities co-occurred with the entity selected or entities of the selected relationship. The *co-occurring entities* of an entity e_i is defined as $e_i.CoEntSet = \{a_j | a_j \in A(f_d(e_i)), a_j \neq e_i\}$. Co-occurring entities are shown with measures chosen by users using the "CHOOSE MEASURES" button. These measures includes $m_{strength}$, m_{d_entity} from T_2, and $m_{path\_strength}$ from T_1. A co-occurring entity can be added to the Network View by using the "INSERT ENTITY" button. When no entity/relationship is selected, this view lists all entities in the Network view with values of $m_{path\_strength}$ from T_1.

When an entity is selected in the Network View, this view displays synonym entities derived from T_2. A "INSERT ENTITY" button is also provided to add

synonym entities into the Network View. When an relationship in the Network View is selected, the Synonym Entity View will be empty.

Path View displays shortest path(s) linking the source entity and the target entity. When an entity is selected in the Network View, It lists all shortest paths through this selected entity. When no entity is selected, this view displays all shortest paths through all entities in the Network view. When a relationship is selected, this view will be empty.

4.1 Visual Analytic Operations

The visual analytics operations including *Insert*, *Delete* and *Cluster* visually implements TUBE operations. Other operations supporting visualization requirements including highlighting, zooming, dragging are also provided. These visual analytic operations can be found on the toolbar and in the Edit menu of ER-Explorer.

The visual analytics operation *Insert* corresponds to SelectCell in T_1 and T_2. Suppose a user inserts an entity e, the mask value will be changed by setting $B^1(c(e)) = 1$ in T_1. As for the mask value in T_2, we set $B^2(c(e, e_i)) = 1$, where $c(e_i) = 1$ in T_1. This reveals all relationships this entity has with all entities in the Network View.

ER-Explorer provides two ways of inserting an entity. One is using the "INSERT ENTITY" button in the Related Entity View and the Synonym Entity View. The other is utilizing the Insert button on the toolbar, which opens a window where all entities existing in the dataset can be retrieved and inserted. This is helpful when a user knows some entity of interest but does not know where to find it in any Views.

The *Delete* operation on a node representing a named object a corresponds to SelectCell operation on T_1. The mask value in T_1 will be changed by setting $B^1(c(e)) = 0$, which visually removes this node from the Network View. However, the same operation on a node representing a conceptual entity e corresponds to Remove in T_1 and T_2. T_2 will be changed by $S_1^2 = S_1^2 - \{e\}$, $B^2(c(e_i, e_j)) = 1$, where $e_i \in e$, $c(e_j) = 1$. The schema part of T_1 will also be changed by $S_1^1 = S_1^1 - \{e\}$. As a result, the conceptual entity is decomposed and its elements are displayed along with their edges connecting entities in the Network View. The *Delete* operation on a relationship $c(e_i, e_j)$ corresponds to the SelectCell operation on T_2. $B^2(c(e_i, e_j)) = 0$, which visually hides this edge.

The *Cluster* operation corresponds to Cluster in T_1 and T_2. Given a new conceptual entity e created by this operation, T_2 will be changed as $S_1^2 = S_1^2 \cup \{e\}$, $B^2(c(e, e_i)) = 1$ *and* $B^2(c(e_k, e_i)) = 0$, where $c(e_i) = 1$ in T_1, $e_k \in e$. T_1 is changed as $S_1^1 = S_1^1 \cup \{e\}$. To use *Cluster*, a user first selects the intended entities in Network View for grouping. He/she then clicks on the cluster button. Visually, all selected nodes are framed by a box representing the new conceptual entity, which can be renamed for easy reference. After this, all edges linking to the selected entities are replaced by edges linking the new conceptual entity and other entities.

5 Case Study

To demonstrate how our ER-Explorer can help to discover entities and relationships that are relevant to association between two given entities, we describe a case study where it is used to find the linkage between two terrorists, Mohammed Afroz and Massod Azhar from the **IC814** dataset. The dataset was derived from a report titled "The Hijacking of IC-814: Al Qaeda, Taliban and Pakistani Factors" which gives a detailed description and analysis about the hijacking of the Indian aircraft IC-814, a well known terrorist incident in year 1999. We extracted entities of types *person*, *organization*, *event*, *GPE*, *product* and *date* as they are more relevant to our discovery task. We then extracted relationships by identifying sentences containing at least two named entities and considered each sentence as a document.

We now describe the process a user will be involved to derive the entity network shown in Figure 3. The user may begin the entity relationship discovery task by first adding the two entities "Mohammed Afroz" and "Massod Azhar" into the Network View. With no other entity selected, the user will see a list of shortest paths between source and target nodes in the Path View. Suppose the user notices that the path $MohammedAfroz \leftrightarrow Indian \leftrightarrow MassodAzhar$ which suggests the two entities are somehow linked by "Indian". She adds "Indian" into the Network View. Next, the user may refer to Related Entity View as she selects "Mohammed Afroz" in the Network View. The Related Entity View shows a list of candidate entities sorted by interestingness measures including $m_{strength}$, m_{d_entity} and $m_{path\_strength}$. The entity "Afroz", a high $m_{strength}$ value in the view looks very similar to "Mohammed Afroz". It may then be inserted into the Network View.

As "Indian" and "Afroz" get inserted into the Network View, several new edges between them also show up in the view. In order to understand the relationships in these edges, the user refers to the Document View of each edge. She may find the only document containing both "Mohammed Afroz" and "Indian" in the sentence "After the confession of Mohammed Afroz was made public by a statement of the Indian minister" which does not imply any meaningful relationships. Hence, the corresponding edge linking the two entities is deleted. The user can also find out that "Afroz" and "Mohammed Afroz" refer to the same person. She therefore uses the *Cluster* operation to group them together and names the new conceptual entity as "C_Mohammed Afroz".

The user subsequently uses the Related Entity View and Path View to explore other entities co-occuring with "C_Mohammed Afroz" or linked to it by shortest paths. She subsequently inserted "Sandy", "Osama bin Laden", and "SIMI" into the Network View. She will also find "India" as a synonym of "Indian" and group them into a conceptual entity "C_India". By reading the document containing "Sandy", she can also find that the latter is one of the hijackers and has an alias "Shahid Akhtar Sayeed". "Shahid Akhtar Sayeed" is then inserted into the Network View. The Synonym Entity View also suggests "Shahid Akhtar Sayed" as another similar entity. Subsequent document verification concludes that they are the same and are grouped into the conceptual entity "C_Shahid

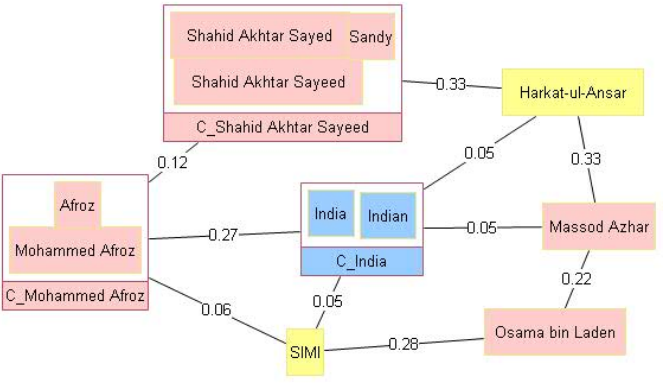


Fig. 3. The Result Network of Our Case Study

Akhtar Sayeed". After checking the supporting document of "C_Shahid Akhtar Sayeed" and "Massod Azhar", the user may find out that the two entities are indirectly linked by "Harkat-ul-Ansar", an organization.

At this point, several entities and relationships have been found while the semantics of the links among them can be summarized in three story threads between Mohammed Afroz and Massod Azhar. The first involves Mohammed Afroz's trainning sponsored by SIMI group, which has a close relation with Osama bin Laden. The latter has ever spent some time with Massod Azhar. The second conveys the information that Mohammed Afroz was active in several places in India and was also arrested there, and so was Massod Azhar. The third says that Mohammed Afroz was trained as a pilot together with Shahid Akhtar Sayeed, who is a member of Harkat-ul-Ansar organization, of which Massod Azhar was the general secretary.

6 Conclusions

In this paper, we propose an interactive visual approach to discover entity and relationships embedded in text data. We have developed a visual analytics tool called ER-Explorer which is equipped with a versatile data model known as TUBE to manipulate entity and relationship information and their supporting documents. We have demonstrated its capability on a hijacking event dataset to discover relationships between two terrorists. For our future research, we plan to extend ER-Explorer to discover associations between more than two entities and to automate some of the exploration subtasks through some tunable parameters. We are also interested to study how concise textual summary of the constructed entity network can be generated from the supporting documents for easy reading.

Acknowledgments

This work was supported in part by A*STAR Public Sector R&D, Project Number 062 101 0031.

References

1. Thomas, J., Cook, K.: A Visual Analytics Agenda. IEEE Computer Graphics and Applications 26(1), 10–13 (2006)
2. Shen, Z., Ma, K.-L., Eliassi-Rad, T.: Visual Analysis of Large Heterogeneous Social Networks by Semantic and Structural Abstraction. IEEE Transactions on Visualization and Computer Graphics 12(6), 1427–1439 (2006)
3. Jeffrey Heer, D.B.: Vizster: Visualizing Online Social Networks. In: Proceedings of the IEEE Symposium on Information Visualization (October 2005)
4. Adam Perer, B.S.: Balancing Systematic and Flexible Exploration of Social Networks. IEEE Transactions on Visualization and Computer Graphics 12(5), 693–700 (2006)
5. Krebs, V.: Mapping networks of terrorist cells. Connections: the Journal of the International Network of Social Network Analysts 24(3), 43–52 (2002)
6. Bilgic, M., Licamele, L., Getoor, L., Shneiderman, B.: D-Dupe: An Interactive Tool for Entity Resolution in Social Networks. In: Proceedings of the IEEE Symposium on Visual Analytics Science And Technology, October 2006, pp. 43–50 (2006)
7. Yang, C.C., Liu, N., Sageman, M.: Analyzing the Terrorist Social Networks with Visualization Tools. In: Proceedings of the IEEE International Conference on Intelligence and Security Informatics (May 2006)
8. Xu, J., Chen, H.: CrimeNet Explorer: A Framework for Criminal Network Knowledge Discovery. ACM Transactions on Information Systems 23(2), 201–226 (2005)
9. Stasko, J., Gorg, C., Liu, Z., Singhal, K.: Jigsaw: Supporting Investigative Analysis through Interactive Visualization. In: Proceedings of the IEEE Symposium on Visual Analytics Science And Technology, October 2007, pp. 131–138 (2007)
10. Jin, W., Srihari, R.K., Wu, X.: Mining Concept Associations for Knowledge Discovery Through Concept Chain Queries. In: Proceedings of the Pacific Asia Conference on Knowledge Discovery and Data Mining (April 2007)
11. Tang, J., Zhang, J., Zhang, D., Yao, L., Zhu, C.: ArnetMiner: An Expertise Oriented Search System for Web Community. In: Proceedings of the 6th International Conference of Semantic Web (November 2007)
12. Lauw, H.W., Lim, E.-P., Pang, H.: TUBE (TextcUBE) for Discovering Documentary Evidence of Associations among Entities. In: Proceedings of the ACM Symposium of Applied Computing (March 2007)
13. Bikel, D., Schwartz, R., Weischedel, R.M.: An Algorithm that Learns What's in a Name. Machine Learning 34(1-3), 211–231 (1999)

Feature Weighting and Selection for a Real-Time Network Intrusion Detection System Based on GA with KNN

Ming-Yang Su, Kai-Chi Chang, Hua-Fu Wei, and Chun-Yuen Lin

Department of Computer Science and Information Engineering,
Ming Chuan University, Taoyuan Campus, Taiwan
minysu@mcu.edu.tw

Abstract. A good feature selection policy which can choose significant and as less as possible features plays a key role for any successful NIDS. The paper presents a genetic algorithm combined with kNN (k-Nearest Neighbor) for feature weighting. We weight all initial 35 features in the training phase and then select tops of them to implement a NIDS for testing. Many DoS/DDoS attacks are applied to evaluate the system. For known attacks we can get the best 97.42% overall accuracy rate while only the top 19 features are considered; as for unknown attacks, we can get the best 78% overall accuracy rate by top 28 features.

Keywords: NIDS (Network Intrusion Detection System), Network Security, DoS/DDoS Attacks, Genetic Algorithm, KNN (k-Nearest Neighbor).

1 Introduction

Most NIDSs emphasize both effectiveness and efficiency at the same time. Usually effectiveness is measured by detection rate and false alarm rate, and efficiency is measured by responding time for an attack occurred. How to select less but significant features for the detection engine becomes vital. Furthermore, features should be weighted because their contributions to the classification are different each other. That is the goal of the paper. Since DoS/DDoS (Denial of Service/Distributed DoS) attack is prevalent and becoming one of the main threads to E-commerce systems, we evaluate our system by DoS/DDoS attacks.

Network intrusion detection systems (NIDSs) are traditionally divided into two broad categories, misuse detection and anomaly detection. Our system proposed in the paper belongs to the anomaly detection. For anomaly detection systems, the most difficult part is how to describe the normal profile, and it is much dependent on the feature weighting and selection. In the literature, most anomaly-based NIDSs focused on system architectures or detection engines designs, only few of them addressed on the feature weighting and selection, such as [1, 2, 3, 4, 5, 6, 7, 8, 9]. Almost all of them evaluated their approaches by KDD CUP99 TCPDUMP datasets. It means that their researches were designed for off-line and thus can't meet the requirement of real-time processing for NIDSs. This is because the announced 41 features in the KDD CUP99 were derived from connection, not packet alone. In fact, the 41 features presented in KDD CUP99 are complicated and varied [1, 6].

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 195–204, 2008.

All features used in the paper are derived from packet headers and gathered using a two-second time window. So naturally the method proposed in the paper can be implemented to be real-time, i.e. making a decision per two seconds. If necessary, we can shrink the time window to one second or half a second. Basically, we adopt genetic algorithm combined with kNN to evolve the weight vector of features. Then we drop the least weighted feature one by one to evaluate the performance of our NIDS. The rest of the paper is organized as follows. Genetic algorithm and kNN are briefly introduced in Section 2. Our proposed method is described in detail in Section 3. Experimental results are shown in Section 4. Finally, a conclusion remark is given in Section 5.

2 Background

We briefly introduce genetic algorithm and kNN in the section because our approach is a GA/kNN hybrid.

2.1 Genetic Algorithm (GA)

The GA is essentially a type of search algorithm used to solve a wide variety of problems. Its goal is to create optimal solutions to problems [10]. A potential solution is encoded as a sequence of bits, characters or numbers. This unit of encoding is called a gene, and the encoding sequence is known as a chromosome.

The GA begins with a set of chromosomes, called population, and an evaluation function that measures the fitness of each chromosome. Usually an initial population of chromosomes is created by complete randomization. During evolution, chromosomes in the population are evaluated according to the fitness function. Based on their fitness values, better chromosomes are selected as parents by selection procedure, and then the parents perform crossover and mutation to form new children chromosomes. Finally some chromosomes in the current generation are replaced by the new ones if necessary, to form the next generation. The evolution is going on until some predefined situation is met, such as the number of iteration reached or acceptable fitness value appeared.

The design of fitness function is the most important issue in GA because a good one can significantly improve the outcome of GA. In the paper, we apply the classification result of kNN to design our fitness function.

Headings. Headings should be capitalized (i.e., nouns, verbs, and all other words except articles, prepositions, and conjunctions should be set with an initial capital) and should, with the exception of the title, be aligned to the left. Words joined by a hyphen are subject to a special rule. If the first word can stand alone, the second word should be capitalized. The font sizes are given in Table 1.

Here are some examples of headings: "Criteria to Disprove Context-Freeness of Collage Languages", "On Correcting the Intrusion of Tracing Non-deterministic Programs by Software", "A User-Friendly and Extendable Data Distribution System", "Multi-flip Networks: Parallelizing GenSAT", "Self-determinations of Man".

2.2 kNN (k-Nearest Neighbor)

One common classification scheme based on the use of distance measures is that of the k-Nearest Neighbor. When a classification is to be made for a new item, its distance to each item in the sampling set must be computed. Only the k closet entries in the sampling set are considered further. The new item is then classified to the class that contains the most items from this set of k closet items.

The distance between two instances represents their similarity; hence ingredients of an instance denote corresponding features. Euclidean distance is usually adopted in the kNN. For any two n-feature instances, say $X = (x_1, x_2, \ldots, x_n)$ and $Y = (y_1, y_2, \ldots, y_n)$, their Euclidean distance is computed as:

$$dist(X, Y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2} \quad (1)$$

A major drawback of such distance measure, i.e., similarity measure used in the kNN is that it regards all features equally. Such phenomenon is especially severe when only a small subset of the features is really contributed to classification.

3 Our Approach

Our proposed NIDS is described by three parts in the section. We present all features that are considered in the paper, state the encoding of a chromosome and the fitness function, and finally describe the details about selection, crossover and mutation.

3.1 Features for NIDS Design

The paper proposes a fast mechanism to detect DoS/DDoS attack from network traffic, so all of the features come from the headers, including IP, TCP, UDP, ICMP, ARP, and IGMP. Table A1 lists all of the 35 features considered in the paper. Every experimental instance contains the 35 feature values observed from one time unit (two seconds in the paper) of network traffic. Since some feature values may be extremely large and dominate the distance calculation in Equation (1), we have to normalize each feature value to the interval [0, 1] as follows.

$$\text{Normalization of } f_i = \frac{1 - e^{-kf_i}}{1 + e^{-kf_i}} \quad (2)$$

In Equation (2), f_i denotes the observed value of feature i, e represents the Euler number (e ≈ 2.7182818284), and k is a constant depend on feature and determined by domain expert in the training phase. Equation (2) makes one instance independent to other instances during the normalization. This is why our proposed method can be easily implemented to a real-time network detection system.

3.2 Genes, Chromosomes and Fitness Function

The goal of GA is to find an optimal weight vector, say $W = [w_1, w_2, \ldots, w_n]$, in which w_i represents the weight of feature i, $1 \le i \le n$. The W will influence the distance

computation and finally promote the kNN classification. In the paper, for any two n-feature instances, say $X = (x_1, x_2, \ldots, x_n)$ and $Y = (y_1, y_2, \ldots, y_n)$, we compute their distance by the following equation.

$$dist(X,Y) = \sqrt{w_1(x_1 - y_1)^2 + w_2(x_2 - y_2)^2 + ... + w_n(x_n - y_n)^2} \quad (3)$$

Each w_i, a real number in the interval [0, 1], is a gene, and a feasible weight vector is a chromosome. Chromosomes in the initial population are randomly generated. After evolution of GA in the training phase, we can get an optimal weight vector, which leads to the best result of kNN classification.

The fitness function we applied to evaluate each chromosome during the evolution is shown in Equation (4), which denotes the overall accuracy rate of kNN classification on the training dataset. In the equation, *Total* represents the total number of instances in the labeled training dataset, and *FP* and *FN* represent the numbers of false positive instances and false negative instances, respectively.

$$\text{fitness} = \frac{Total - FP - FN}{Total} \quad (4)$$

3.3 Selection, Crossover, and Mutation

Roulette Wheel selection is used to choose two chromosomes from parent population to produce children chromosomes. By Roulette Wheel, the larger fitness value a chromosome has, the more chance it gets to be chosen as parent.

We apply two-point crossover to exchange genes between parent chromosomes. The genes between two crossover points of parents are exchanged to produce children. In mutation, we set every chromosome has 10% probability to mutate. For a chromosome that gets the chance to mutate, we set each gene also has 10% probability to be randomly changed. If a mutated chromosome's fitness is larger than that of its original form, it then replaces the original one. Otherwise, the mutated chromosome is ignored.

4 Experimental Results and Analyses

A commercial application named IP Traffic [11] was applied to produce background traffic, which can generate any amount of TCP/UDP/ICMP packets. Two hosts running IP Traffic played sender and receiver, respectively, and we deployed the receiver in the LAN and the sender transmitting packets through the Internet. Using IP Traffic, we can choose protocol and data source with mathematical laws (Pareto, Uniform, and Exponential), file or packet generator with configurable contents. Inter-packet delay and packet size can be selected. Normally we kept the network traffic amount in the range of 10 to 80 Mbps. One laptop in the LAN launched DoS attacks against the victim. Our system was coded by Microsoft Visual C++, and run on a laptop with Windows XP.

All of the 65 attack programs used for experiments were downloaded from the VX Heavens website (http://vx.netlux.org/), which is maintained by the well-known

Kaspersky. In the website, there are 207 programs for Win32 belonging to the DoS category. However, maybe missing some specific DLLs only 65 can be executed. Thus, totally 65 DoS attcaks were considered in the experiments. All experimental datasets can be found in the website on http://163.25.149.41/isi2008 (or http://163.25.148.92/isi2008/). Three kinds of datasets were used: sampling dataset, training dataset, and testing dataset. Every instance in the datasets is a 35-feature vector obtained from sniffing network packets for one unit of 2 seconds. Certainly, every instance has been normalized by Equation (2). The sampling dataset contains 100 normal instances and 100 attack instances, which is used for kNN classification. The training dataset contains 600 normal instances and 600 attack instances, using to calculate fitness value of chromosome in GA evolution.

Initially, we randomly generated 30 chromosomes for the initial population. Using the training set, each chromosome, say $W = [w_1, w_2, \ldots, w_{35}]$, was evaluated by computing its fitness value by Equation (4), while the distance calculation followed Equation (3). So, in order to obtain one chromosome's fitness value, Equation (3) should be computed $1{,}200 \times 200 = 180{,}000$ times because there are totally 1,200 instances in the training set and 200 instances in the sampling set. Figure 1 shows the fitness values during the evolution for the case of 30 chromosomes and 100 generations. For each generation, only the best chromosome's fitness was cared. The experiment was repeatedly run 5 times. In the figure, the bold line depicts the maximal fitness value among all runs, and the thin line depicts the average of all runs.

In Figure 1, we have got the maximal fitness among all runs to be 0.965, which was obtained by the chromosome given below. Table A2 lists the sorted 35 features by their weights.

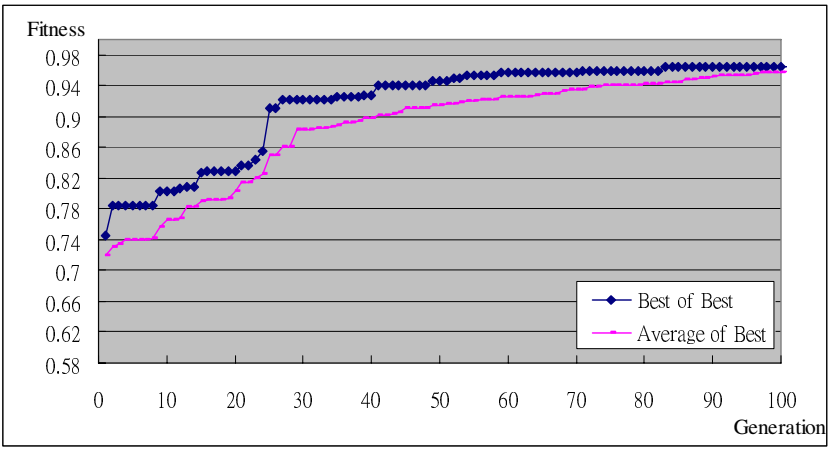


Fig. 1. Fitness vs. generation

[**1:**0.437, **2:**0.0453, **3:**0.0537, **4:**0.3091, **5:**0.8914, **6:**0.1056, **7:**0.8767, **8:**0.4591, **9:**0.3627, **10:**0.9682, **11:**0.7047, **12:**0.6193, **13:**0.0317, **14:**0.0489, **15:**0.026, **16:**0.0217, **17:**0.7894, **18:**0.0189, **19:**0.5274, **20:**0.1518, **21:**0.6561, **22:**0.064, **23:**0.4444, **24:**0.1203, **25:**0.0544, **26:**0.2404, **27:**0.9812, **28:**0.1592, **29:**0.0646, **30:**0.6866, **31:**0.7586, **32:**0.9327, **33:**0.8225, **34:**0.5816, **35:**0.0128]

All attack instances in the sampling and training datasets were derived from 20 out of the total 65 DoS/DDoS programs. Using the same 20 attack programs to generate TestA dataset for known attacks, the other 45 programs were applied to generate TestB dataset for unknown attacks. Both TestA and TestB contained 600 attack instances and 600 normal instances.

While all of the 35 features were considered, performances for TestA and TestB are shown in Table 1. The true positive rate and false positive rate for TestA are 93.83% and 0.33% respectively, and for that of TestB are 51.17% and 1%, respectively. The overall accuracy for TestA is 96.75% and for TestB is 75.08%. The term of overall accuracy is computed as (TP+TN) / (TP+FP+TN+FN).

Table 1. Performances for TestA and TestB as all of the 35 features being considered

	TestA	TestB
TP	93.83%	51.17%
FP	0.33%	1%
TN	99.67%	99%
FN	6.17%	48.83%
Overall Accuracy	96.75%	75.08%

Next we removed feature one by one from the least one; as one feature being removed, we retrained for the best chromosome and re-evaluated by TestA and TestB. The results are shown in Figure 2 for TestA and Figure 3 for TestB. For comparison, we also depict performances of untrained weight vectors in both figures by white lines. An untrained weight vector has all members to be 1, i.e., all features being treated equally. In Figure 2, the best performance in term of overall accuracy occurred as top 19 features were applied, and it was 97.42%. Contrast to all 35 features was applied in which the overall accuracy was 96.75%. In Figure 3, the best accuracy was 78% as top 28 features were applied. However, as all 35 features were applied, the result was slightly lowered to be 75.08%. It is interesting to conclude that as the number of applied features is larger than (or less than) a specific number the performance of weighted vector is significantly superior to that of untrained vector. This may due to some features are not necessary to be considered. As the features are refined and the size is reduced, the performance of weighted vector is close to that of untrained weighted vector in term of overall accuracy.

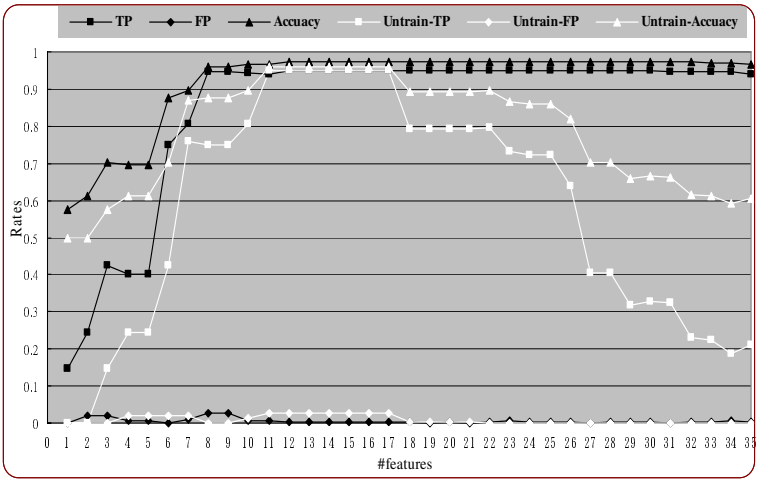


Fig. 2. Performances with different number of features for TestA

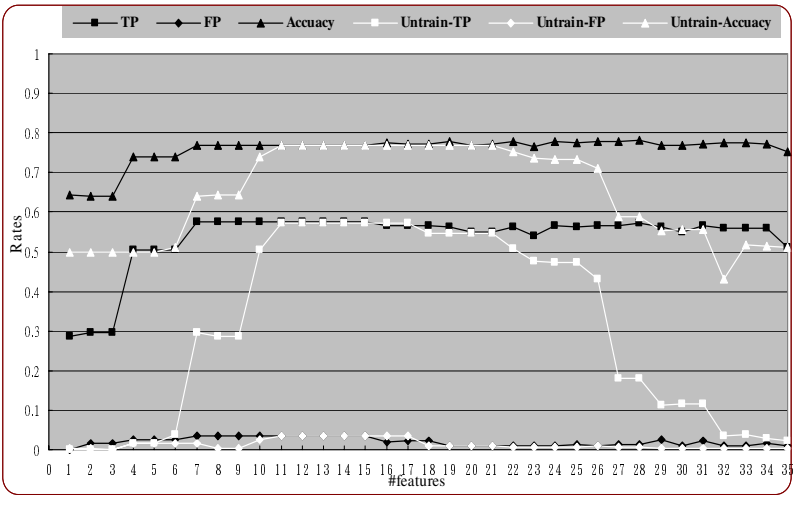


Fig. 3. Performances with different number of features for TestB

5 Conclusion

We have proposed a method to weight features of DoS/DDoS attacks, and analyzed the relationship between detection performance and number of features. Different to previous works, our method can be and really have been implemented to a real-time NIDS. This is because all features applied in the paper are directly collected from packet headers. All 35 features were evaluated and weighted. According to our experiments on DoS/DDoS attacks, (1) for known attack detection the best overall accuracy was 97.42% for which only top 19 features were considered; (2) for

unknown attack detection the best overall accuracy was 78% for which top 28 features were considered.

Acknowledgments. This work was partially supported by the National Science Council under contracts NSC 95-2221-E-130-003 and 96-2221-E-130-009.

References

1. Middlemiss, M.J., Dick, G.: Weighted Feature Extraction using a Genetic Algorithm for Intrusion Detection. In: Proceedings of the Evolutionary Computation, vol. 3, pp. 1669–1675 (2003)
2. Hofman, A., Horeis, T., Sick, B.: Feature Selection for Intrusion Detection: An Evolutionary Wrapper Approach. In: Proceedings of the IEEE Neural Networks, vol. 2, pp. 1563–1568 (2004)
3. Sung, A.H., Mukkamala, S.: Identifying Important Features for Intrusion Detection Using Support Vector Machines and Neural Networks. In: Proceedings of the IEEE Symposium on Applications and the Internet, pp. 209–216 (2003)
4. Abbes, T., Bouhoula, A., Rusinowitch, M.: Protocol Analysis in Intrusion Detection Using Decision Tree. In: Proceedings of the IEEE Conference on Information Technology: Coding and Computing, pp. 404–409 (2004)
5. Lee, C.H., Chung, J.W., Shin, S.W.: Network Intrusion Detection Through Genetic Feature Selection. In: Proceedings of the IEEE Conference on Software Engineering, Artifical Intelligence, Networking, and Parallel/Distributed Computind (SNPD), pp. 109–114 (2006)
6. The UCI KDD Archive, http://kdd.ics.uci.edu/databases/kddcup99/kddcup.names
7. Stein, G., Chen, B., Wu, A.S., Hua, K.A.: Decision Tree Classifier f or Network Intrusion Detection With GA-based Feature Selection. In: Proceedings of the ACM Southeast Regional Conference, vol. 2, pp. 136–141 (2005)
8. Mukkamala, S., Sung, A.H.: Feature ranking and Selection for Intrusion Detection Using Support Vector Machines. In: Proceedings of the Conference on Information and Knowledge Engineering, pp. 503–509 (2002)
9. DARPA 1999 Intrusion Detection Evaluation, http://www.ll.mit.edu/IST/ideval/data/data_index.html
10. Holland, J.H.: Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence. MIT Press, Cambridge (1992)
11. IP Traffic, http://www.omnicor.com/netest.htm

Appendix

Table A1. List of all features

No	Prot.	Feature	No	Prot.	Feature
1	IP	S.IP slots hit	19	TCP	URG_Flag+URG_data count
2	TCP	S.IP+S.port slots hit	20	TCP	ACK_Flag+ACK count
3	TCP	S.IP+D.port slots hit	21	TCP	Checksum_error count
4	TCP	S.IP+SYN count	22	TCP	Same_length_interval count
5	TCP	S.IP+URG_Flag+URG_data count	23	TCP	Port (20)+length(>1400) count
6	TCP	S.IP+ACK_Flag+ACK count	24	UDP	S.port count
7	ARP	S.IP+ARP count	25	UDP	D.port count
8	IP	D.IP slots hit	26	UDP	checksum_error count
9	IP	Header length!=20 count	27	UDP	Same_length_interval count
10	IP	MF_Flag count	28	ICMP	Type error count
11	IP	(Total length>1400 ‖ <40) &&TTL == 64 count	29	ICMP	checksum_error count
12	IP	checksum_error count	30	ICMP	Length>1000 count
13	TCP	S.port slots hit	31	IGMP	Type error count
14	TCP	D.port slots hit	32	IGMP	checksum_error count
15	TCP	S.port count	33	IGMP	Length>1000 count
16	TCP	D.port count	34	ARP	Size error count
17	TCP	Sequence_number==0 count	35		Total packet number
18	TCP	SYN count			

Table A2. List of sorted features by their weights

rank	Prot.	feature	weight	rank	Prot.	Feature	weight
1	UDP	Same_length_interval count	0.9812	19	TCP	S.IP+SYN count	0.3091
2	IP	MF_Flag count	0.9682	20	UDP	checksum_error count	0.2404
3	IGMP	checksum_error count	0.9327	21	ICMP	Type error count	0.1592
4	TCP	S.IP+URG_Flag+URG_data count	0.8914	22	TCP	ACK_Flag+ACK count	0.1518
5	ARP	S.IP+ARP count	0.8767	23	UDP	S.port count	0.1203
6	IGMP	Length>1000 count	0.8225	24	TCP	S.IP+ACK_Flag+ACK count	0.1056
7	TCP	Sequence_number==0 count	0.7894	25	ICMP	checksum_error count	0.0646

Table A2. (*continued*)

8	IGMP	Type error count	0.7586	26	TCP	Same_length_interval count	0.064
9	IP	(Total length>1400 ‖ <40)&&TTL == 64 count	0.7047	27	UDP	D.port count	0.0544
10	ICMP	Length>1000 count	0.6866	28	TCP	S.IP+D.port slots hit	0.0537
11	TCP	checksum_error count	0.6561	29	TCP	D.port slots hit	0.0489
12	IP	checksum_error count	0.6193	30	TCP	S.IP+S.port slots hit	0.0453
13	ARP	Size error count	0.5816	31	TCP	S.port slots hit	0.0317
14	TCP	URG_Flag+URG_data count	0.5274	32	TCP	S.port count	0.026
15	IP	D.IP slots hit	0.4591	33	TCP	D.port count	0.0217
16	TCP	Port (20)+length(>1400) count	0.4444	34	TCP	SYN count	0.0189
17	IP	S.IP slots hit	0.437	35		Total packet number	0.0128
18	IP	Header length!=20 count	0.3627				

Locality-Based Server Profiling for Intrusion Detection

Robert Lee and Sheau-Dong Lang

School of Electrical Engineering and Computer Science
University of Central Florida
4000 Central Florida Blvd.
Orlando, FL, 32816 U.S.A.
{rlee,lang}@cs.ucf.edu

Abstract. Detection of intrusion on network servers plays an ever more important role in network security. This paper investigates whether analysis of incoming connection behavior for properties of locality can be used to create a normal profile for network servers. Intrusions can then be detected due to their abnormal behavior. Experiments show that connections to a typical network server do in fact exhibit locality, and attacks can be detected through their violation of locality.

Keywords: Computer network security, local area networks, network servers, intrusion detection.

1 Introduction

The number of computers connected to the Internet continues to increase at a rapid pace. These computers are ready targets for malicious attacks. While network firewalls protect most computers, firewalls cannot stop all intrusions. Furthermore, if an intrusion is not detected, a trusted computer can be hijacked for use as a drone in a botnet or for other undesirable purposes. In particular, network servers are vulnerable to intrusions because they often sit at the boundary between the inside and outside of a firewall and receive connections from both sides. If such a server is hijacked, the attacker then has access to computers inside the firewall. Typically, a large number of users rely on the continuous availability of a server. Thus, it is important to detect server intrusions when they do occur.

The approaches to detecting malicious intrusions can be divided into two major categories: signature-based detection and behavior-based detection. Signature-based detection relies on past observation of the characteristics of a network connection that was known to be an intrusion. Those characteristics make up a signature for that type of intrusion. If a new connection matches an existing signature, it is probable that the connection represents an intrusion. One advantage of this approach is a high confidence that the new connection is indeed an intrusion. One disadvantage is the inability to recognize an intrusion that does not match any of the existing signatures.

Behavior-based detection does not look for the characteristics of a specific type of intrusion. Instead, a profile of normal network traffic is created. A network connection that does not fit the profile is then suspect. Note that behavior-based detection will easily coexist with signature-based detection. One need not be used in lieu of the other; in fact, the two types of detection complement each other.

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 205–216, 2008.

The particular type of behavior we wish to investigate involves the principle of locality. A practical application of locality arises in the memory cache of a modern computer. A piece of data that has been retrieved recently from memory is likely to be accessed again in the near future, so it is worthwhile to store it temporarily in the cache, where it can be accessed with less delay. Periodically, a piece of data that has not been accessed for a long time is removed from the cache to make room for newer data. For intrusion detection, we use two types of working sets: one type contains IP addresses of incoming network connections and the other type contains the ports used by the connections. Both types function as a kind of cache, with the expectation that connections are likely to be made from IP addresses that have connected recently, and that these connections will likely reuse ports that have been used recently.

In this paper, we describe a locality-based method for detecting anomalous incoming connections to a network server. Analysis of the normal incoming connections to the server yields a profile. We demonstrate that connections to a production server do in fact fit a profile; hence, we can use the profile to detect intrusions.

The rest of the paper is organized as follows. Section 2 describes related work. Section 3 details the connection analysis algorithm. Section 4 explains the experimental setup and results. Finally, section 5 concludes the paper.

2 Related Work

In [1], locality-based analysis was used to build a profile of outgoing network traffic from a personal computer on a local area network. The profile included two important parameters for the working set of recently contacted IP addresses: the removal interval, or how often an IP address was removed from the set, and the size threshold. If the working set grew above the size threshold, this violation of locality indicated that suspicious behavior was occurring. The analysis used the fact that the pattern of outgoing network traffic from a personal computer was closely correlated with human behavior, including common tasks such as email and web surfing. A behavior-based approach has been used to detect the spread of worms [3] and viruses [5]. The authors of [3] noted that outgoing connections created by worms fail to connect more often than those from legitimate users or applications, so worms could be detected by counting these failed connection attempts. The virus detection method in [5] relied on the observation that, under normal conditions, there is a limit to the rate at which a computer makes new outgoing connections. In [6], malicious network behavior was detected by correlating network traffic with user activity. [4] found that a human user is likely to make network connections to recently visited machines, in accordance with the principle of locality. [7] and [8] incorporated the principle of locality to develop a worm detection system using a sliding time window of different sizes. The window included the connections made during the time interval, and the different sizes were designed to capture attacks with varying rates of connection.

These past studies focused on the behavior of outgoing traffic. In contrast, we investigate whether incoming connections to a network server, which include both human-initiated connections and automated network traffic, also exhibit evidence of locality. We propose that the incoming connections to a network server can be subjected to locality-based analysis to create a normal profile.

3 Connection Analysis

Our connection analysis algorithm watches each packet as it arrives at the server and determines to what connection and session it belongs. A connection is defined as a stream of packets between two IP addresses. A session is a pair of (source, destination) ports used by a connection. Each connection may have more than one associated session. The algorithm keeps a working set of incoming connections. This working set has variable size with removal of the least recently used connection at regular intervals of time, as described in [2]. The connection working set stores the most recent access time for each connection.

The algorithm also associates each connection with a working set containing its sessions. Like the overarching connection working set, the session working sets use variable size with removal of the least recently used session at fixed removal intervals. See Figure 1 below for a diagram of the working set structure.

Over time, the connection working set will grow or shrink based on the behavior of the incoming connections. If new connections are always made from IP addresses already in the working set, the size will decrease as connections are automatically removed. If new connections are always made from IP addresses not in the working set, the size will increase if the rate of new connections exceeds the rate at which connections are removed from the working set. Each session working set will exhibit similar behavior based on the ports used by its associated connection.

If the incoming packet belongs to a connection that already exists in the working set, the algorithm updates the most recent usage time of the connection. If no connection exists in the working set for the packet, a new connection is created. Likewise, the algorithm updates the most recent usage time in the session working set if the packet belongs to an existing session; otherwise, it adds a new session for the packet. At regular intervals of time, the algorithm scans through the connection working set

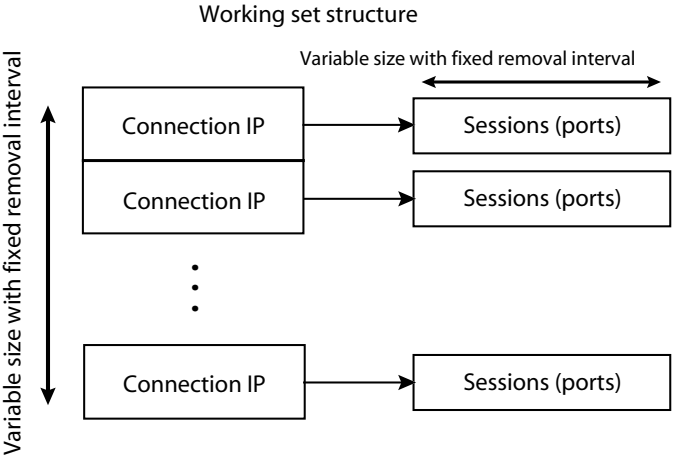


Fig. 1. The structure of the working set for incoming internal connections

and removes the least recently used connection. The interval is a constant parameter called the connection removal interval. The algorithm also scans through each session working set at regular intervals and removes the least recently used session from each. The session removal interval can be adjusted independently from the connection removal interval. Over time, the working sets will vary in size depending on the behavior of the incoming connections and sessions.

A working set exhibits locality if its size does not grow past a certain level, or threshold, over time. The locality arises in the idea that connections are likely to be made from IP addresses already in the connection working set and incoming packets are likely to belong to a session already in a session working set. Once a normal size threshold for a given server has been determined, the growth of the working set past that size threshold would indicate violation of locality, and hence would detect anomalous network behavior that could correspond to a malicious attack. The two types of working sets function concurrently to detect different kinds of attacks. Violation of locality in the connection working set would indicate an attack originating from multiple IP addresses. Violation of locality in a session working set would indicate an attack that connects to multiple ports. In our experiments, we will use working set analysis to show that connections to a typical network server and their associated sessions do in fact exhibit locality.

4 Experimental Results

We used the Wireshark network protocol analyzer [9] to capture the headers of all network packets incoming to and outgoing from a production network server for a class at our university over a period of four workdays, Tuesday through Friday, from December 5 through 8, 2006. This server, which ran the Microsoft Windows Server 2003 operating system, was used by students to access class-related information. During the four-day period, 29666 connections were made to the server, with 2520223 incoming packets.

4.1 Connection Working Set Behavior

We chose a one-hour window from 13:00 to 14:00 during each day to focus our study. See Table 1 below for the breakdown of incoming connections and packets during the one-hour window each day. Note that there is strong consistency in the connection counts among the four days; in fact, the connection count never varies by more than 2.1% from the mean.

Table 1. Number of incoming connections and packets by day during the 13:00 to 14:00 window

Day	Incoming Connections	Incoming Packets
Tuesday	336	31698
Wednesday	338	27436
Thursday	349	30041
Friday	344	37416

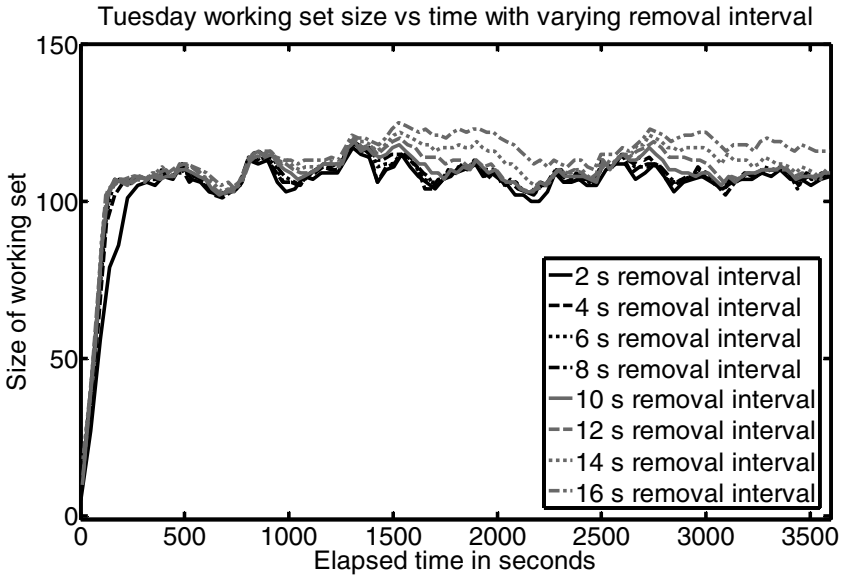


Fig. 2. The size of the Tuesday connection working set over time, given a fixed removal interval. The size does not exceed 125.

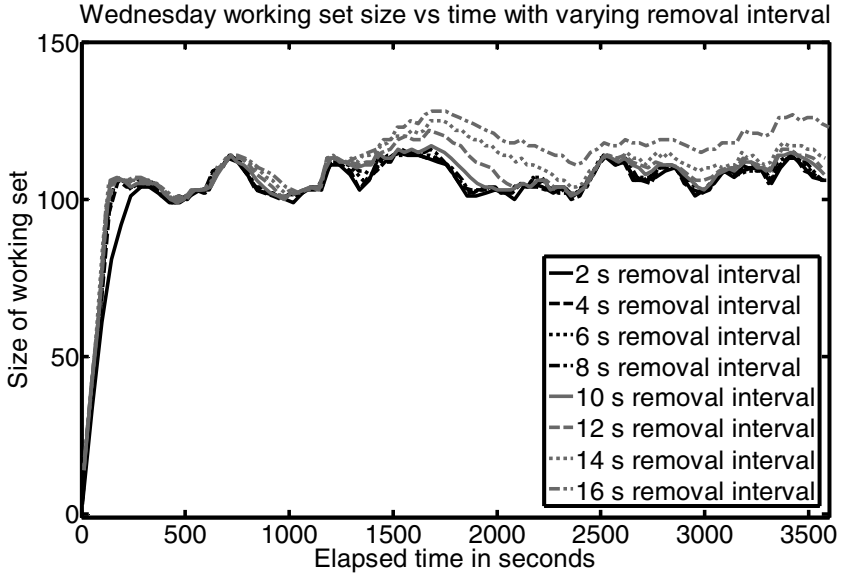


Fig. 3. The size of the Wednesday connection working set over time, given a fixed removal interval. The size does not exceed 129.

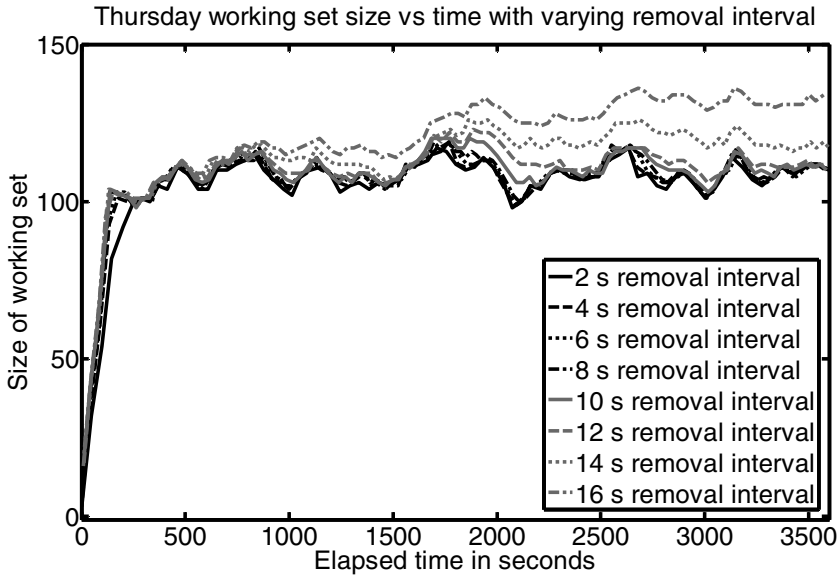


Fig. 4. The size of the Thursday connection working set over time, given a fixed removal interval. The size does not exceed 136.

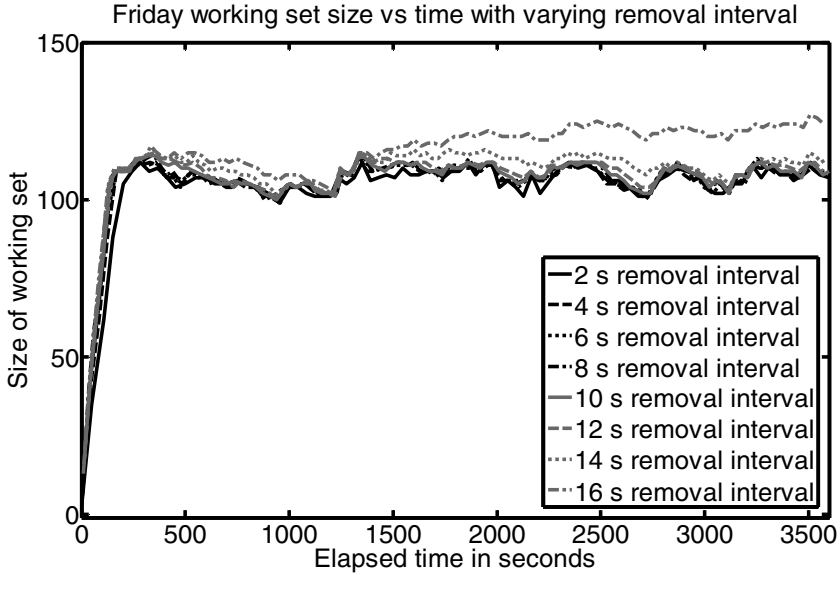


Fig. 5. The size of the Friday connection working set over time, given a fixed removal interval. The size does not exceed 127.

These four sets of data, one for each day from Tuesday through Friday, serve as four separate observations of connection behavior. We analyzed each of the four sets individually. Figures 2, 3, 4, and 5 below show connection working set behavior for

Tuesday, Wednesday, Thursday, and Friday, respectively. Each figure shows the size of the working set during the one-hour period, given a certain fixed removal interval. The eight lines in each figure represent removal intervals of 2, 4, 6, 8, 10, 12, 14, and 16 seconds.

Several observations stand out from these data. On each of the four days, the working set size reaches a near-constant level within about 300 seconds. There is clear consistency in the behavior of the working sets across the four days. The working set sizes do not exceed 125, 129, 136, and 127 on Tuesday, Wednesday, Thursday, and Friday, respectively. As the removal interval increases, the maximum size of the working set also increases. This is expected, because connections are being removed less often from the working set. The data support the idea that incoming connections to a network server behave in accordance with the principle of locality. For this network server, we can choose a connection removal interval of between 2 and 16 seconds and a threshold of about 150 for violation of locality. The experimental data also indicate that the working set size remains constant for a wide range of removal intervals (between 2 and 10 seconds), but starts to deviate when the removal interval is greater than 10 seconds. Thus, the working set need not be updated very frequently during real-time monitoring.

These working sets include only incoming connections to a network server. The working sets in [1] that consisted of outgoing connections from a personal computer differed in the size threshold for violation of locality. Under normal behavior, which included mostly human-initiated activity, the working set did not grow past a size of 6 connections when using a removal interval of 2 seconds. Here, our server working sets do not grow past a size of about 120 connections when using a removal interval of 2 seconds. Note that we do not yet have experimental data for attacks on the server from multiple computers to test whether violation of locality would occur, although we plan to carry out such experiments in the future. However, we expect that an attack originating from multiple IP addresses, such as a distributed denial-of-service attack, would cause the size of the working set to exceed the threshold. The consistency of the working set size over time assures us that normal incoming connections to the server follow a predictable pattern.

4.2 Session Working Set Behavior

We carried out a TCP/IP fingerprinting attack against the server using nmap [10], a network security scanner. This attack connects to more than 1000 ports on the target machine in an attempt to determine the operating system running on the machine, based on the characteristic pattern of open and closed ports exhibited by various operating systems. The attacker can then design an attack against the particular operating system that was detected. Figures 6, 7, and 8 below are semilog graphs of session working set behavior under the fingerprinting attack, using a session removal interval of 2, 4, and 6 seconds, respectively. In addition, we also conducted a port scan attack using nmap. This attack connects to more than 1700 ports on the target machine to determine which ports are open and possibly vulnerable to attack. Figures 9, 10, and 11 below are semilog graphs of session working set behavior under the port scan attack, using a session removal interval of 2, 4, and 6 seconds, respectively.

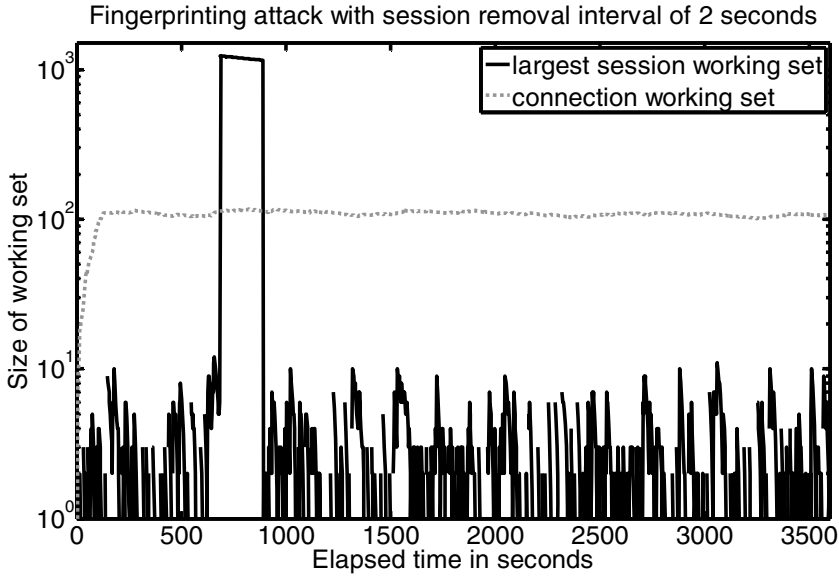


Fig. 6. The size of the largest session working set over time, given a removal interval of 2 seconds, under a fingerprinting attack. Maximum size of the largest session working set is 12 when not under attack.

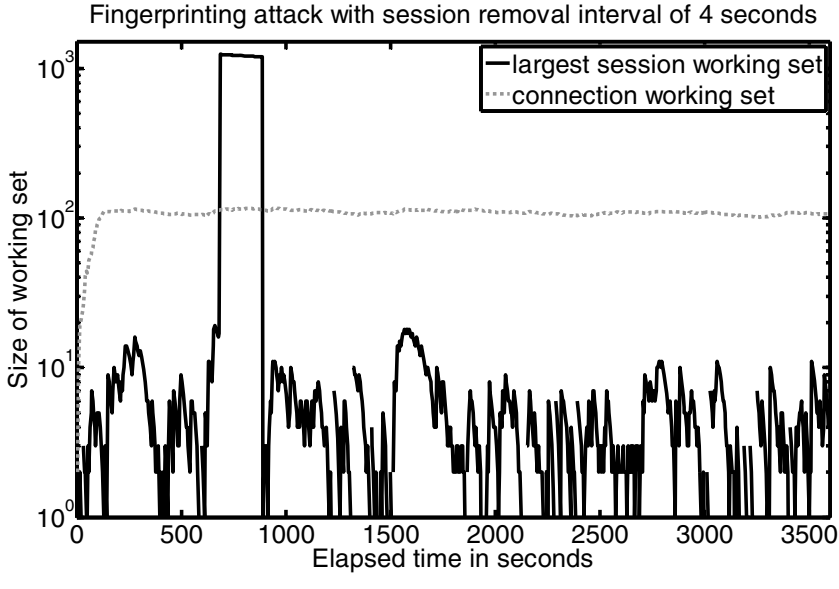


Fig. 7. The size of the largest session working set over time, given a removal interval of 4 seconds, under a fingerprinting attack. Maximum size of the largest session working set is 19 when not under attack.

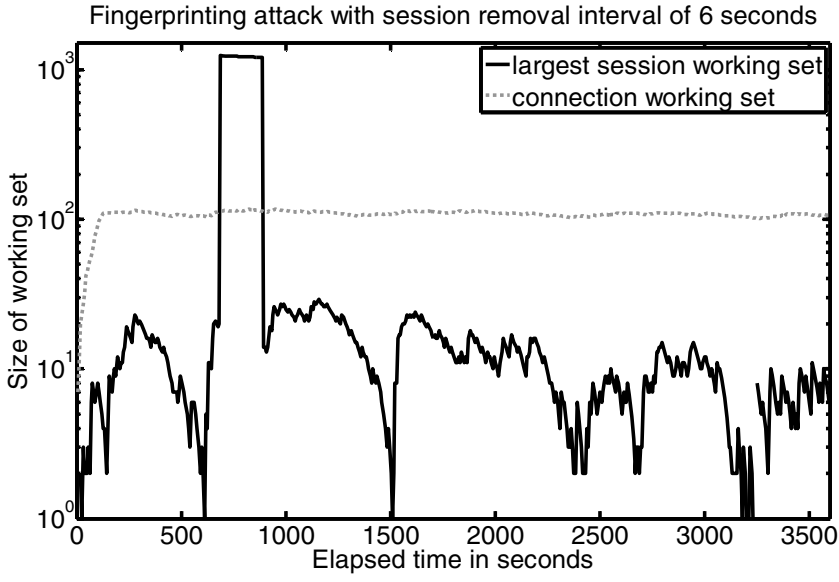


Fig. 8. The size of the largest session working set over time, given a removal interval of 6 seconds, under a fingerprinting attack. Maximum size of the largest session working set is 29 when not under attack.

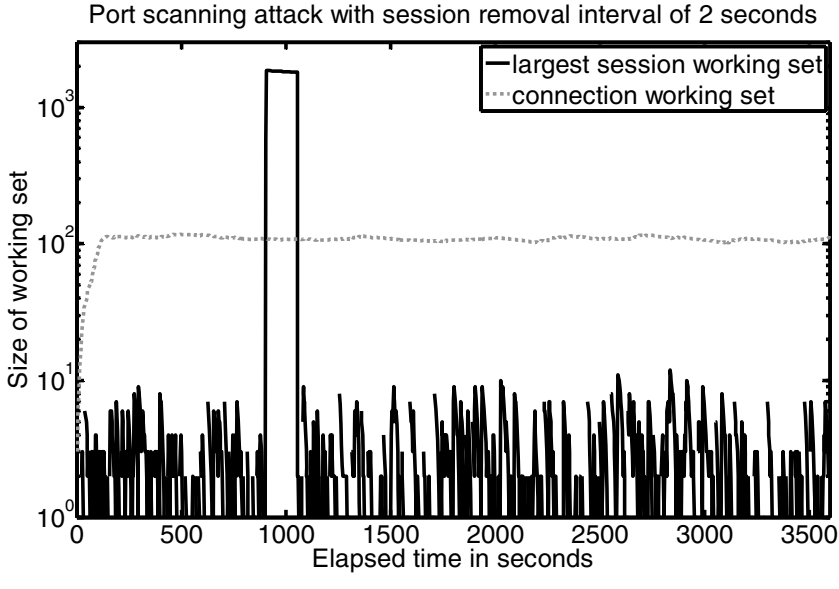


Fig. 9. The size of the largest session working set over time, given a removal interval of 2 seconds, under a fingerprinting attack. Maximum size of the largest session working set is 12 when not under attack.

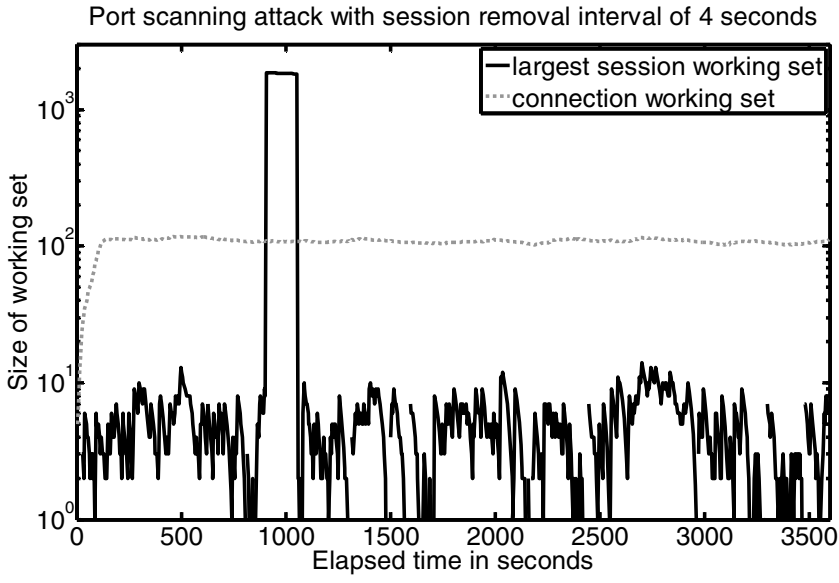


Fig. 10. The size of the largest session working set over time, given a removal interval of 2 seconds, under a fingerprinting attack. Maximum size of the largest session working set is 14 when not under attack.

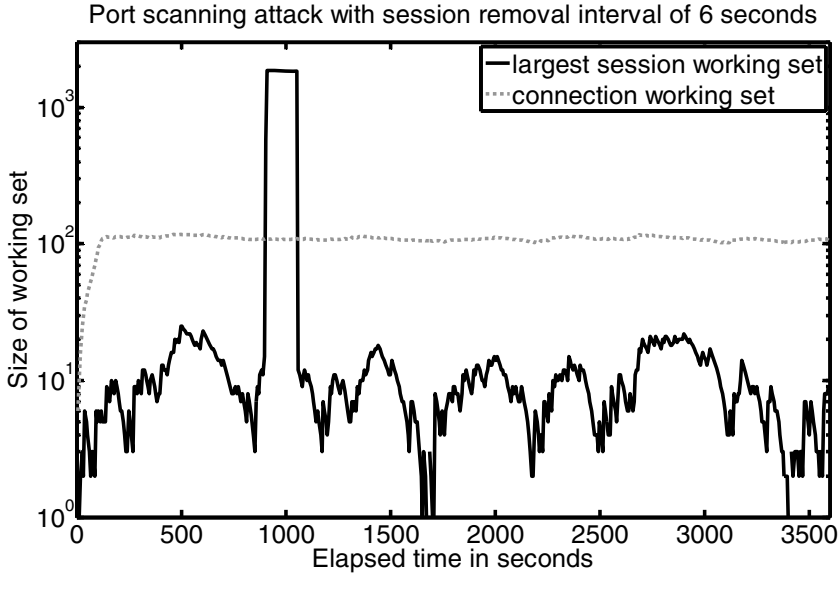


Fig. 11. The size of the largest session working set over time, given a removal interval of 2 seconds, under a fingerprinting attack. Maximum size of the largest session working set is 25 when not under attack.

The figures also include the size of the connection working set as a separate line. Recall that each connection in the connection working set has an associated session working set. Each figure shows the size of the largest session working set among all the connections. The figures include a one-hour period surrounding the attack.

In Figures 6 through 8 above, the fingerprinting attack appears as a steep rise in the size of the largest session working set to over 1000. In Figures 9 through 11 above, the port scan attack appears as a steep rise in the size of the largest session working set to nearly 2000. As sessions are removed from that largest working set, the size gradually decreases. The sudden drop in the size of that set occurs when the connection that performed the fingerprinting attack is removed from the connection working set. When there is no attack present, the size of the largest session working set never exceeds a certain level. For the fingerprinting attack, with a removal interval of 2, 4, and 6 seconds, the maximum size is 12, 19 and 29, respectively. For the port scan attack, with a removal interval of 2, 4, and 6 seconds, the maximum size is 12, 14 and 25, respectively. The data support the thesis that incoming sessions behave in accordance with the principle of locality. For this network server, we can choose a session removal interval of between 2 and 6 seconds and a threshold of about 35 for violation of locality.

Note that the connection working set size, represented by a dotted line in the six attack data figures above, remains nearly constant at about 100. This is consistent with our earlier observations regarding connection working set behavior. We would not expect an attack from a single IP address to affect the connection working set.

5 Conclusion and Future Work

Our experimental results showed that connections to a typical network server and their associated sessions exhibit locality. We demonstrated that this allows us to generate a normal profile for the network server. By conducting attacks on the server, we confirmed that violation of locality occurred in a session working set, allowing us to detect the attacks.

A strength of our locality-based intrusion detection method is its simplicity of implementation in real-time. One type of attack for which signature-based detection complements our method is “sneaky” scans, that is, scans that connect at a rate too slow to affect the working set in a detectable manner.

In the future, we wish to gather data over several weeks to confirm that the network connection and session behavior remains consistent over longer periods. The session working set successfully detected attacks from a single computer on multiple ports. We also plan to mount coordinated attacks on a server from multiple computers to show that the connection working set is able to detect such attacks.

References

1. Zhou, M., Lee, R., Lang, S.-D.: Locality-Based Profile Analysis for Secondary Intrusion Detection. In: Proc. 8th International Symposium on Parallel Architectures, Algorithms and Networks, pp. 166–173. IEEE Computer Society Press, Washington (2005)

2. McHugh, J., Gates, C.: Locality: a New Paradigm for Thinking About Normal Behavior and Outsider Threat. In: ACM New Security Paradigms Workshop. ACM Press, New York (2004)
3. Berk, V., Bakos, G.: Designing a Framework for Active Worm Detection on Global Networks. In: Proc. 1st IEEE International Workshop on Information Assurance. IEEE Computer Society Press, Washington (2003)
4. Hofmeyr, S.: An Immunological Model of Distributed Detection and Its Application to Computer Science Security. Ph.D. dissertation, Dept. of Computer Science, Univ. of New Mexico (1999)
5. Williamson, M.: Throttling Viruses: Restricting Propagation to Defeat Malicious Mobile Code. In: 18th Annual Computer Security Applications Conference. IEEE Computer Society Press, Washington (2002)
6. Cui, W., Katz, R.H., Tan, W.: BINDER: An extrusion-based break-in detector for personal computers. Technical report, Hewlett-Packard Laboratories, Palo Alto, CA (2004)
7. Sekar, V., Xie, Y., Reiter, M.K., Zhang, H.: A multi-resolution approach for worm detection and containment. In: Proc. International Conference on Dependable Systems and Networks, pp. 189–198. IEEE Computer Society Press, Washington (2006)
8. Sekar, V., Xie, Y., Reiter, M.K., Zhang, H.: Is host-based anomaly detection + temporal correlation = worm causality? Technical report, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA (2007)
9. Wireshark Network Protocol Analyzer, http://www.wireshark.org/
10. Nmap Network Security Scanner, http://nmap.org/

A Simple WordNet-Ontology Based Email Retrieval System for Digital Forensics*

Phan Thien Son[1,2], Lan Du[1,2], Huidong Jin[1,2],
Olivier de Vel[3], Nianjun Liu[1,2], and Terry Caelli[1,2]

[1] NICTA Canberra Lab, Locked Bag 8001, Canberra ACT 2601, Australia
Huidong.Jin@nicta.com.au
[2] RSISE, the Australian National University, Canberra ACT, 0200, Australia
[3] Command, Control, Communications and Intelligence Division, DSTO, PO Box 1500, Edinburgh SA 5111, Australia

Abstract. Because of the high impact of high-tech digital crime upon our society, it is necessary to develop effective Information Retrieval (IR) tools to support digital forensic investigations. In this paper, we propose an IR system for digital forensics that targets emails. Our system incorporates WordNet (i.e. a domain independent ontology for the vocabulary) into an Extended Boolean Model (EBM) by applying query expansion techniques. Structured Boolean queries in Backus-Naur Form (BNF) are utilized to assist investigators in effectively expressing their information requirements. We compare the performance of our system on several email datasets with a traditional Boolean IR system built upon the Lucene keyword-only model. Experimental results show that our system yields a promising improvement in retrieval performance without the requirement of very accurate query keywords to retrieve the most relevant emails.

1 Introduction

As our dependency on information and communications technology increases, so does our exposure to computer-related vulnerabilities and threats. The UK National Hi-Tech Crime Unit (NHTCU) estimated that the financial impact of hi-tech crime on UK business rose from £190 million in 2003[1] to £2.4 billion in 2005[2]. In Australia, it has also been observed that the financial impact of hi-tech crime increased from 2003 to 2005 by more than AU$10 million[3]. As a consequence, the development of an effective IR system for digital forensics is necessary to enhance information security.

* The authors thank the reviewers for suggestive comments. NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

[1] http://www.forensicfocus.com/index.php?name=Downl oads& d_op=getit&lid=9
[2] http://acl.ldc.upenn.edu/W/W98/W98-0704.pdf
[3] http://www.aic.gov.au/publications/htcb/htcb004.html

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 217–228, 2008.

Digital forensics aims to recreate the temporal causal sequence of events arising from the unauthorized intrusion of digital systems [1] and to present it as evidence in a court of law. It deploys computer techniques for the recovery, authentication, and analysis of digital evidence, which can be collected from various sources, such as storage devices like hard drivers, removable disks, networks, or embedded digital systems [2]. It is clear that digital forensics plays an importance role in cyber-crime investigations.

Nowadays email is one of the most important sources of digital evidence. The analysis of emails can reveal a large amount of important information that is of interest to the investigator, such as intellectual property theft. Considering the importance of emails, we develop a novel email retrieval system, which can index, summarize and analyze textual data semantically, to retrieve suspicious information for investigators using more flexible query keywords. In general, the system uses WordNet to expand and refine Boolean queries with semantically related terms. Specifically, for negation queries, the retrieved emails are limited to those that are judged to be relevant to positive query keywords but irrelevant or negatively relevant to negated query keywords.

The rest of the paper is organized as follows: an overview of related work is given in Section 2. Section 3 discusses our query processing technique. Following this, the mathematical IR model is described in Section 4. Section 5 explains the implementation of our system. Some experiments with this system are reported and compared in Section 6, followed by concluding comments in Section 7.

2 Related Work

Since our system can be seen as the combination of the WordNet ontology and the Extended Boolean Model (EBM), we first trace the development of EBM and WordNet based query expansion techniques.

Salton et al. [3] noticed that it is impossible for the pure Vector Space Model (VSM) to incorporate phrase-like constructs or a set of synonyms to facilitate users to explicitly express structured queries by using “AND” and “OR” Boolean operators, like that in the traditional Boolean Model (BM). By taking into account this concern, they proposed an extended Boolean IR strategy based on P-normal model under the premise of VSM discussed in [4]. We agree with Salton et al. that for disjunction queries (i.e. “OR” queries), a document containing all query terms should be conceptually more relevant than that containing fewer query terms; but for conjunction queries (i.e. AND), a document containing none of the query terms should be less relevant than that containing some but not all of query terms. However, the choice of P value limits the performance of this model, which means the final optimal P value varies for different query types and its value still has to be experimentally determined. Based on the Generalized VSM (GVSM) proposed in [5], Wong et al. presented another approach to map the Boolean retrieval environment to the vector space environment so that Boolean queries can be handled by explicitly expressing term vectors. In fact, our system can be taken as the extension of VSM with Boolean query processing

in terms of query expansion based on WordNet and the generalized Boolean query formulation.

For query expansion based on WordNet, there have been quite a few attempts in the literature of IR with the purpose of handling term-mismatching problems. These approaches attempt to use the semantic information extracted from various lexical sources to refine user's queries in order to improve the accuracy of IR. How to decide the semantically related terms to be added to the original queries is still a challenging issue. In 1994, Voorhees [6] conducted a series of experiments on WordNet based query expansion. She used three different query expansion strategies. It was found that the performance improvement can only be observed on short queries; there was no significant difference in retrieval performance for long queries. However, she did not take into consideration the Boolean queries in a logical model [7], which query expansion can benefit from. In 1998, Mandala et al. [8] presented an approach that used co-occurrence-based and predicated-argument based thesaurus to enrich WordNet. It yielded a substantial improvement. Similarly, in the work of [9], a local thesaurus was generated by projecting the initial query result onto a global thesaurus. A hybrid approach that combined the local and global thesaurus was used to refine the original query. Analogously, other query expansion techniques considering such as syntactic, statistical, word sense, semantic network, have been evaluated in [10], [11], [12], [13] and [14] respectively.

3 Query Processing

The technique we present here to process user queries is built upon the adoption of query term weighting and query re-formulation. A primitive approach of handling semantic negation query is also discussed.

3.1 Introduction to WordNet

WordNet is a large general-purpose lexical dictionary manually-built by Goerge Miller and his colleagues at Princeton University based on the phycholinguistic principles [15]. Due to the capability of searching dictionaries conceptually, it has been pervasively taken as a corpus independent ontology for vocabulary in Natural Language Processing (NLP). The vocabulary in WordNet is divided into four syntactic taxonomies, i.e. nouns, verbs, adjectives, and adverbs, which are grouped into synonym sets (i.e. synset). Synsets are the basic objects of WordNet representing the underlying lexicalized concepts. They are generated according to the different word senses and linked by the semantic relations, which include synonymy, antonymy, hyponymy/hypernymy, meronym/holonym, troponymy and entailment [15].

In this paper, we consider only two relationships: (1) Synonymy: a relationship in which two terms involved are expected to be interchangeable; (2) ISA: a relationship is also referred as a hyponym/hypernym relationship in the same context; instead of the features inherited from the hypernym, a hyponym has

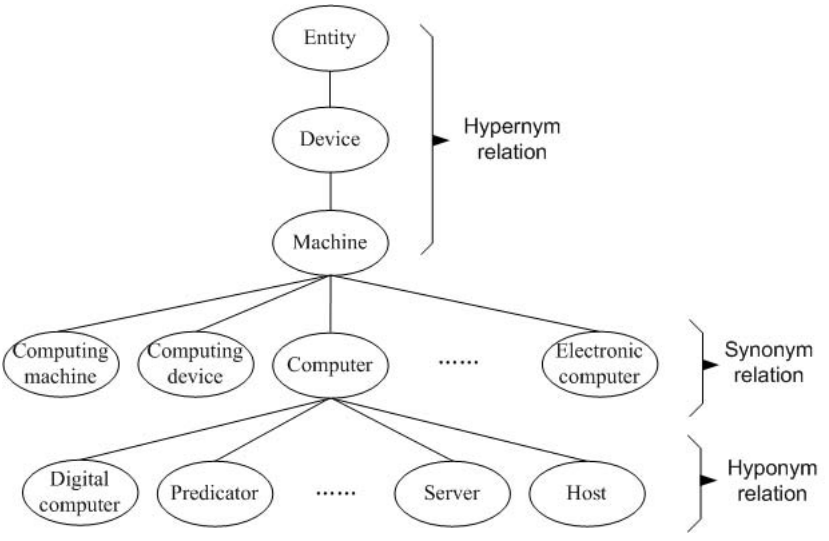


Fig. 1. Hierarchical relation of the word "computer" in WordNet with the sense of a "machine for performing calculations automatically"

at least one feature that discriminates itself from its hypernyms and from other hyponyms of that hypernym [15]. For example, Fig. 1 shows the hierarchical tree structure of WordNet which is constituted with synonym, hypernym and hyponym by adopting an example term "computer" in the sense of "machine". We call two terms semantically similar if there is a short direct/indirect link between two terms in WordNet. The semantic distance is the minimum number of edge(s) from one term to the other, and the distance between two terms which are not semantically similar is considered to be infinite. This definition will be used in next section to develop a term weighting function based on edge-counting measure.

3.2 Query Specification

We use the Extended BNF (EBNF) notation [16] to specify the three kinds of Boolean queries, e.g. disjunction query, conjunction query and negation query. The BNF for those queries is defined as bellow:

$$Q = 1^* \langle [+| - |\text{white-space}] \ \text{keyword} \rangle \tag{1}$$

where "+" for conjunction (i.e. required terms which have to be contained in all retrieved documents); "−" for negation (i.e. negated terms which cannot be contained in any of the retrieved documents); "white-space" for disjunction indicates the terms are optional.

As an example of boolean queries in EBNF, consider the query "+trackball +compaq keyboard −celeron". This query requires: 1) that the terms "trackball"

and "compaq", or their semantically related expansion terms must appear in the retrieved documents; 2) that the term "keyboard" and its semantically related expansion terms may or may not appear in the retrieved documents, which means they are optional; 3) and documents that contain the term "celeron" and its semantically related expansion terms must not be retrieved.

3.3 Query Expansion

In the field of IR, the most common way of undertaking query expansion with semantically related terms is to use linguistic information extracted from various lexical sources, such as WordNet. In order to tackle the term-mismatching problem caused by user query keywords that are not always the same as the ones by which the document has been indexed in describing the same semantics, our system utilizes a more flexible query expansion method to refine user queries before conducting a search. We use WordNet, as an underlying reference ontology for the vocabulary, to enlarge the query such that terms semantically related to the concepts indicated in the original query are included in the expanded query.

In particular, synonymy and hypernymy/hyponymy relationships in the WordNet are used. The reasons include: firstly, synonyms are interchangeable, which means the substitution of one term for another in a linguistic context does not alter the true value of the content [15]; secondly, the exploration of hypernymy/hyponymy relations can find more generic and specific concepts respectively. It is observed that there is a large probability that users could also be interested in documents which contain the hyponyms of a query keyword "A", if they have specified "A" in their queries. For purpose of illustration of our query expansion method, we here consider a query keyword "conference" with a one depth expansion and its first sense in WordNet. Now, traversing the hierarchical tree structure of WordNet upwards gives us the hypernyms that are "meeting" and "group meeting"; Similarly, we can have its hyponyms, which are "seminar", "colloquium" and "symposium", by traversing the tree downwards; but there is no synonym synset under the chosen sense for "conference". Finally, all semantic expansion terms collected from our query expansion process based on the two relationships are put together with the original keywords and each assigned a weight calculated according to Eq. 14 (see Section 4.2) to distinguish their importance.

3.4 Query Restriction

In order to remove the unwanted terms along with the corresponding unwanted meanings [17] and increase the accuracy of IR for digital forensics, we employ negation queries to restrict the retrieved documents. While normal users may be interested in expanding results based on semantic similarity, forensics investigators may be more interested in confining the meaning of keywords to some particular ones. For example, it is possible that a user may consider "notebook" in the meaning of "a book with blank pages for recording notes or memoranda" (e.g. jotter), but not in the meaning of "a small compact portable computer"

(e.g. laptop). Then the query can be expressed as "notebook −computer" to remove the second meaning. Our basic idea of query restriction is preventing emails that are positively relevant to negated query keywords from appearing in retrieved email collection, and from affecting the overall retrieval performance.

4 Mathematical Model of Our System

In this section, we elaborate the mathematical model which is used in our digital forensics email retrieval system. It consists of two parts, namely the email retrieval model and the email ranking algorithm. The BM and VSM are integrated into our model in an innovative way.

4.1 Email Retrieval Model

Our email retrieval model discussed below can be regarded as the extension of a BM by combining Boolean logic and set theory. Elaborately, let K indicate the set of user query keywords, $K_{required}$ indicate the set of keywords the user requires to exist in the emails, $K_{optional}$ indicate the set of optional query keywords that may or may not appear in the emails, and $K_{prohibited}$ indicate the set of keywords the user wants to remove from the original query. Intuitively, we have

$$K = K_{required} \cup K_{optional} \cup K_{prohibited} \tag{2}$$

where $K_{required} \cap K_{optional} = \emptyset$, $K_{required} \cap K_{prohibited} = \emptyset$ and $K_{optional} \cap K_{prohibited} = \emptyset$. Then, given the set T of tokens indexed in a collection of emails E, a function $token: K \rightarrow T$ is defined as

$$token(k) = \begin{cases} t & \text{if } t \text{ is an appropriate token for query keyword } k \\ null & \text{otherwise} \end{cases} \tag{3}$$

then, we have

$$T_{required} = \{t \mid t \in T \wedge \exists k \in K_{required} : \ token(k) = t\} \tag{4}$$

$$T_{prohibited} = \{t \mid t \in T \wedge \exists k \in K_{prohibited} : \ token(k) = t\} \tag{5}$$

$$T_{optional} = \{t \mid t \in T \wedge \exists k \in K_{optional} : \ token(k) = t\} \tag{6}$$

Now, let N be a collection of pre-defined thresholds dis, we define the function $related: T \times T \times N \rightarrow \{true, false\}$ to determine if two tokens t_i and t_j are semantically related with distance dis as shown in Eq. 7. Similary, the function $extend: T \times N \rightarrow T$ is given by Eq. 8

$$related(t_i, t_j, dis) = \begin{cases} true & \text{if semantic distance between } t_i \text{ and } t_j \text{ is } dis \\ false & \text{otherwise} \end{cases} \tag{7}$$

$$extend(t_i, dis) = \begin{cases} \{t_j \mid related(t_i, t_j, dis))\} & \text{if } dis \neq 0 \\ \{t_i\} & \text{else if } dis = 0 \end{cases} \tag{8}$$

$$included(t_i, e_j) = \begin{cases} true & \text{if } e_j \text{ has token } t_i \\ false & \text{otherwise} \end{cases} \tag{9}$$

Basically, dis determines the number of edges in WordNet tree structure with which a query is expanded. Given a collection of emails E, Eq. 9 gives us another function $include : T \times E \rightarrow \{true, false\}$. Consequently, the set of emails retrieved with respect to $T_{required}$, $T_{optional}$, and $T_{prohibited}$ are

$$E^{dis}_{required} = \bigcap_{t \in T_{required}} \left\{ e \;\middle|\; \exists t_0 \in \bigcup_{i=0}^{dis} extend(t,i) : \; included(e, t_0) \right\}, \quad (10)$$

$$E^{dis}_{optional} = \bigcap_{t \in T_{optional}} \left\{ e \;\middle|\; \exists t_0 \in \bigcup_{i=0}^{dis} extend(t,i) : \; included(e, t_0) \right\}, \quad (11)$$

$$E^{dis}_{prohibited} = \bigcap_{t \in T_{prohibited}} \left\{ e \;\middle|\; \exists t_0 \in \bigcup_{i=0}^{dis} extend(t,i) : \; included(e, t_0) \right\}. \quad (12)$$

Finally, the collection E^{dis} of retrieved emails in response to the refined query with dis depth expansion is modeled as

$$E^{dis} = \begin{cases} E^{dis}_{required} - E^{0}_{prohibited} & \text{if } \left| E^{dis}_{required} \right| \neq 0 \\ E^{dis}_{optional} - E^{0}_{prohibited} & \text{if } \left| E^{dis}_{required} \right| = 0. \end{cases} \quad (13)$$

4.2 Email Ranking Algorithm

Once a list of emails is retrieved according to the user's information needs expressed by his/her query, our system computes a semantic similarity between the query and emails. It is important to note that, as discussed in Section 3, the expansion terms are used to retrieve more emails that would satisfy the user's information needs. Clearly, no matter how sophisticated the algorithm is, there is a probability that users are not interested in the extended keywords, or our system gives inappropriate senses for these original keywords. Moreover, preliminary experiments show that, the greater distance from an extended keyword to its original keyword, the higher probability that it does not comply with the sense of the original keyword. Therefore, it is necessary to reduce the weight of expansion terms according to the number of edges between them and their corresponding original query keywords, so as to reflect their different semantic importance. Based on the above observations, our term weighting function $w : T \rightarrow R$ is defined as

$$w(t) = \begin{cases} 1.0 & \text{if } \exists k \in K : token(k) = t \\ \frac{1.0}{2^{\rho}} & \text{else } \rho = \min\left(\{\tau \mid \exists k \in K : \; related(token(k), t, \tau)\}\right). \end{cases} \quad (14)$$

Thereafter, the final email ranking algorithm (i.e. the query-to-email similarity function $sim(e, q)$) utilizes the traditional TF-IDF algorithm from VSM as:

$$sim(e,q) = \sum_{t_p \in P} w(t_p) \times tf(t_p, e) \times idf(t_p) - \sum_{t_q \in Q} w(t_q) \times tf(t_q, e) \times idf(t_q), \quad (15)$$

where $P = \left(\bigcup_{t_r \in T_{required}} \bigcup_{i=0}^{dis} extend(t_r, i)\right) \bigcup \left(\bigcup_{t_o \in T_{optional}} \bigcup_{j=0}^{dis} extend(t_o, j)\right)$ and $Q = \bigcup_{t_p \in T_{prohibited}} \bigcup_{\mu=0}^{dis} extend(t_p, \mu)$.

5 System Implementation

In our system, we use the user-friendly query formulation as described in Section 3 and the mathematical model discussed in Section 4. Since this system is implemented using the Java programming language, it can run well on both Windows and Linux platforms. Fig. 2 illustrates its main user interface. Since our system was developed to work in an investigative environment involving suspicious email retrieval, it should be able to deal with several popular mail agents as well as future mail agents. Currently, this system can import emails from several popular email agents including Microsoft Outlook Express, perform retrieving functions (i.e. both keyword-based and ontology-based search) and show users the ranked documents in descending order of estimated relevance to users' queries.

In order to deal with a variety of mail agents, to communicate with WordNet, and to have acceptable usability, the system is implemented mainly based on JGoodies (http://www.jgoodies.com), Jmbox (http://sourceforge.net/projects/jmbox), Flexdock (https://flexdock.dev.java.net), Log4j (http://logging.apache.org/log4j), JWNL (http://sourceforge.net/projects/jwordnet), etc. As shown in Fig. 2, a novice investigator can simply pose his/her queries and run without modifying any configuration, while our system still provides several options to refine results for advanced users. For example, a user can choose the part-of-speech of a word.

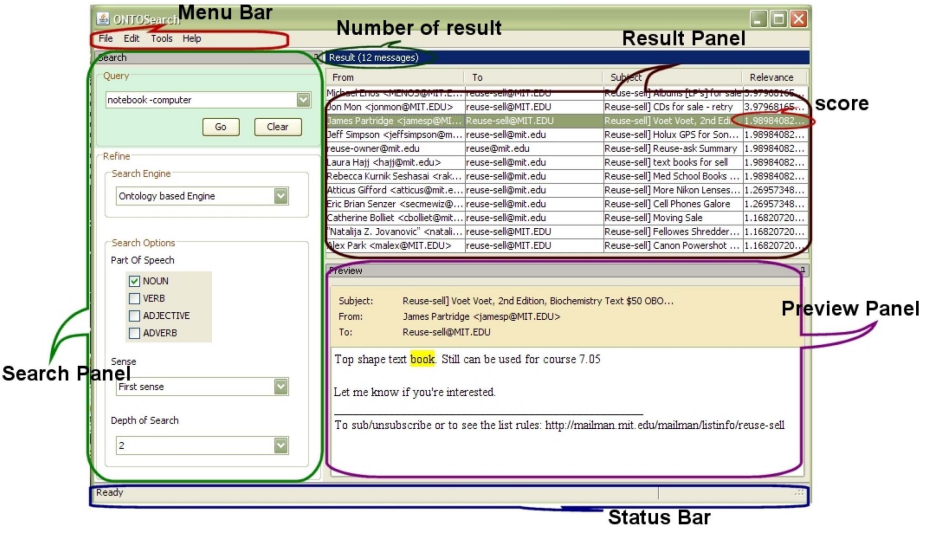


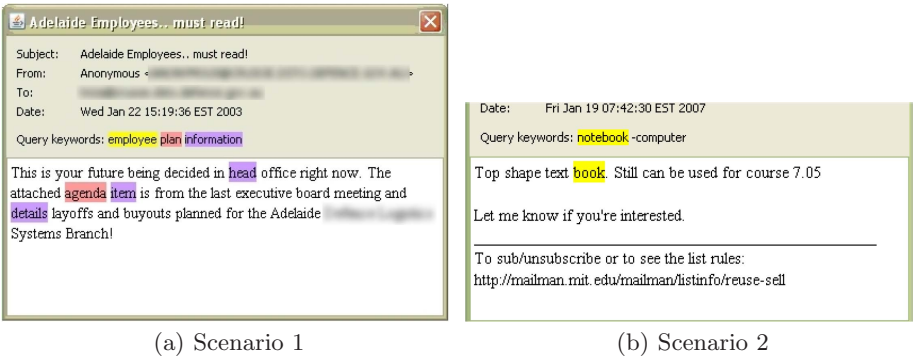
Fig. 2. The main user interface of the implemented system

6 Experiments and Results

Our email retrieval system has been tested on several email datasets, each of which has a number of emails ranging from 300 to 3000. Those email datasets include email sets from the MIT Reuse Ask/Sell Meeting Web Gateway and the DBWorld newsletter respectively. Specifically, we have used the following default settings, as shown in Fig. 2, in the reported experiments. Firstly, we set the semantic distance threshold *dis* to 2. The reason is that a value equal to 1 is too shallow to retrieve all relevant emails, whilst a value of 3 could be so deep that too many irrelevant terms are added to the query, which reduces the retrieval performance; Secondly, the first sense is chosen to extend the keywords, as WordNet arranges word senses in order of term usage frequency. Although each word may have several senses, Krovetz et al [18] have found that retrieval effectiveness could not be significantly affected by word sense ambiguity if retrieved documents have many words in common with the query. Clearly, a word sense disambiguation technique can still provide a better result. Therefore, users can choose different senses in our user interface to further disambiguate different word meanings.

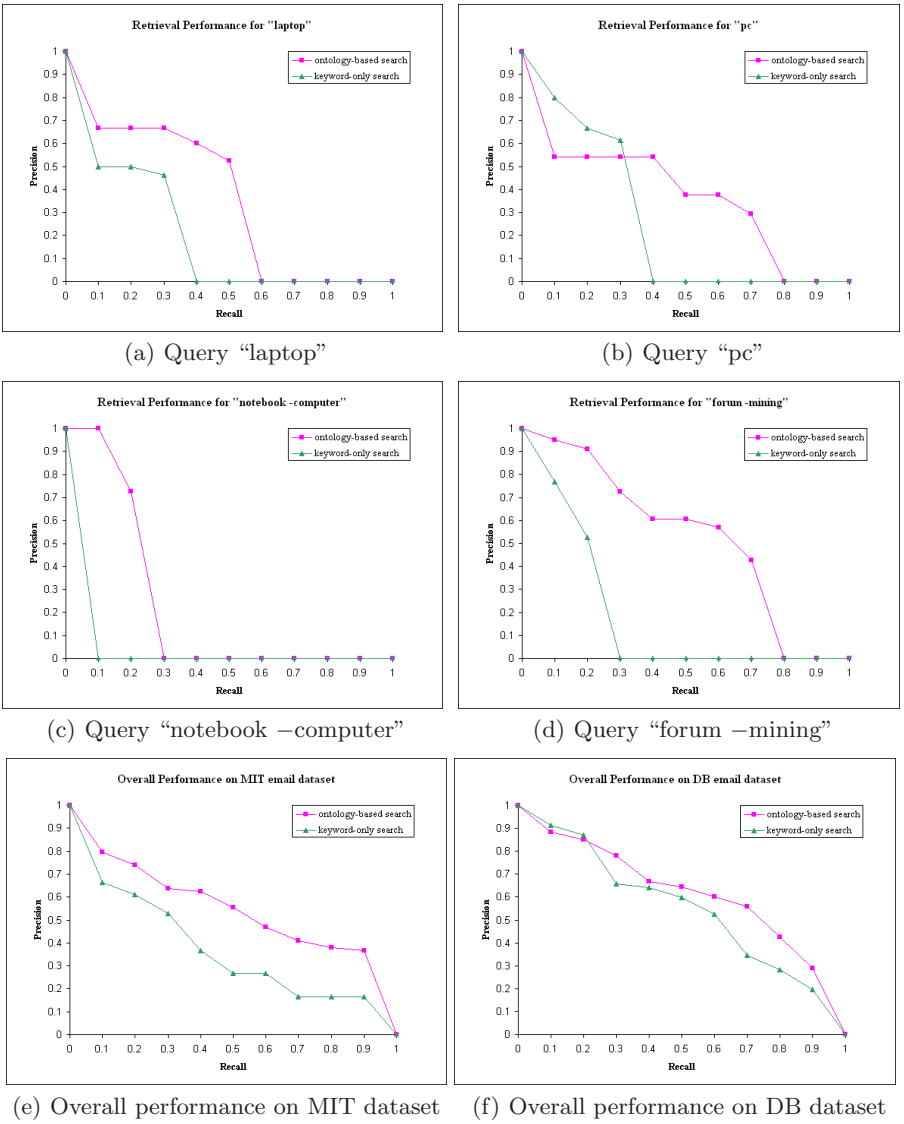
This system has been evaluated with a number of Boolean queries with around three keywords on average. In the following, we report and discuss the experimental results in terms of typical examples, and present their precision-recall curves.

We first use two typical scenarios to evaluate the effectiveness of our system, i.e. one adopted from a synthetic digital forensics scenario [2], the other using a manual ranking of all emails. The first scenario involves a confidential restructuring proposal of a large company that was disclosed by an anonymous email. A forensics team was called in to find out who emailed this document. Suppose that an investigator is notified about this leaking event, but he/she does not know what exactly the content of the email was, and who sent it. With our system, he/she can simply pose such a query, e.g. "employee plan information",



(a) Scenario 1 (b) Scenario 2

Fig. 3. Two examples of retrieved emails in the scenarios. Some contents in scenario 1 have been deliberately obfuscated for data confidentiality reasons.



(a) Query "laptop" (b) Query "pc"

(c) Query "notebook −computer" (d) Query "forum −mining"

(e) Overall performance on MIT dataset (f) Overall performance on DB dataset

Fig. 4. Comparative precision-recall performance results for various types of queries ((a)-(d)), as well as overall performance results for the two email datasets ((e)-(f))

and our system can retrieve the most suspicious emails. Fig. 3(a) shows one of the suspicious emails which has been assigned a quite high rank (in the top five). All query terms and their semantically related expansion terms are highlighted with different colors. Similarly, in the second scenario we assume that there is a person who receives a newsletter that includes a variety of topics. This email dataset has approximately 3000 emails. The investigator wishes to find emails

relating to "notebook" but not to "computer". He does this by using the query "notebook −computer". The first retrieved email is shown in Fig. 3(b). Both scenarios clearly shows the effectiveness of our system.

We have carried out a series of experiments to compare our system (i.e. ontology-based search) with a keyword-only search-based system which we implemented using the Lucene IR library (http://lucene.apache.org). Fig. 4 illustrates the retrieval performance of these two systems on our two email datasets, i.e. one with 321 emails from the MIT Reuse Ask/Sell Meeting Web Gateway, the other with 419 emails from the DBWorld newsletter, and using 10 random boolean queries for each. All emails were ranked manually by a research assistant. For the query "laptop", as shown in Fig. 4(a), the precision of the ontology-based system sits well above that of the keyword-only system till the recall reaches the value 0.6. However, in this example, WordNet does not include all proper nouns for the term "computer", e.g. HP Pavilion, which causes the precision of both systems to fall to the value zero for higher recall. Similar results were observed for the query "pc" (see Fig.4(b)). In this example our system is able to achieve a high precision with a corresponding high recall. In contrast, the precision value of the keyword-only search quickly falls to zero when the recall reaches a value equal to 0.4. Figures 4(c) and 4(d) illustrate the performance on negation queries. As can be seen, the ontology-based search system performs better than the keyword-based system in both negation query examples. For example, consider the query "notebook −computer". In WordNet, "notebook" has two senses, i.e. "a book with blank pages for recording notes or memoranda" and "a small compact portable computer". It is observed that the ontology-based system can successfully remove the second sense from "notebook" by the negated term "computer".

The overall precision-recall performance curves for the two email datasets are shown in Figures 4(e) and 4(f) respectively. On the MIT data set, ontology-based search significantly outperforms keyword-only search by sometimes up to twice the performance. When the recall value is 0.50, e.g., the precision of ontology-based search is 0.56, compared with 0.27 for keyword-only search. In Fig. 4(f), it is clear that the precision of ontology-based search exceeds that of keyword-only search when the recall is greater than 0.20. Ontology-based search yields an average improvement of 13.30%, though keyword-only search does slightly better when the recall value is quite low.

7 Conclusion and Discussion

In this paper, we have proposed an interesting email retrieval system for digital forensics based on an innovative WordNet ontology-driven mathematical model. The system we have implemented: 1) integrates query expansion by WordNet; 2) adopts Boolean logic and set theory to determine relevant emails; and 3) combines the TF-IDF algorithm and an edge-counting based term weight measure to calculate query-to-document similarity. Results of our experiments are promising, especially comparing with a keyword-only search system, and clearly illustrate the effectiveness of our system. The system also provides a user friendly

interface for forensic investigators. We plan to adopt term sense disambiguation [19,20] along with automatic phrase indexing/recognizing techniques to further enhance our system. The system will be evaluated in real-world applications.

References

1. Casey, E.: Digital Evidence and Computer Crime: Forensic Science, Computers, and the Internet with CDROM. Academic Press, Inc., London (2000)
2. de Vel, O.Y., Liu, N., Caelli, T., Caetano, T.S.: An embedded bayesian network hidden markov model for digital forensics. In: Mehrotra, S., Zeng, D.D., Chen, H., Thuraisingham, B., Wang, F.-Y. (eds.) ISI 2006. LNCS, vol. 3975, pp. 459–465. Springer, Heidelberg (2006)
3. Salton, et al.: Extended boolean information retrieval. Commun. ACM 26(11), 1022–1036 (1983)
4. Salton, G., McGill, M.: Introduction to modern information retrieval. McGraw-hill, New York (1983)
5. Wong, et al.: Generalized vector spaces model in information retrieval. In: SIGIR 1985, pp. 18–25. ACM Press, New York (1985)
6. Voorhees, E.M.: Query expansion using lexical-semantic relations. In: SIGIR 1994, pp. 61–69 (1994)
7. Parapar, et al.: Query expansion using WordNet with a logical model of information retrieval. In: IADIS AC, pp. 487–494 (2005)
8. Mandala, et al.: The use of WordNet in information retrieval. In: Proceedings of Use of WordNet in Natural Language Processing Systems, pp. 31–37 (1998)
9. Grootjen, F.A., van der Weide, T.P.: Conceptual query expansion. Data Knowl. Eng. 56(2), 174–193 (2006)
10. Moldovan, D.I., Mihalcea, R.: Using WordNet and lexical operators to improve internet searches. IEEE Internet Computing 4(1), 34–43 (2000)
11. Finkelstein, et al.: Placing search in context: the concept revisited. ACM Trans. Inf. Syst. 20(1), 116–131 (2002)
12. Zukerman, et al.: Query expansion and query reduction in document retrieval. In: ICTAI 2003 (2003)
13. Liu, et al.: An effective approach to document retrieval via utilizing WordNet and recognizing phrases. In: SIGIR 2004, pp. 266–272 (2004)
14. Gong, et al.: Web query expansion by WordNet. In: Andersen, K.V., Debenham, J., Wagner, R. (eds.) DEXA 2005. LNCS, vol. 3588, pp. 166–175. Springer, Heidelberg (2005)
15. Miller, G.A.: WordNet: a lexical database for English. Commun. ACM 38(11), 39–41 (1995)
16. Wirth, N.: What can we do about the unnecessary diversity of notation for syntactic definitions? Commun. ACM 20(11), 822–823 (1977)
17. Widdows, D.: Orthogonal negation in vector spaces for modelling word-meanings and document retrieval. In: Dignum, F.P.M. (ed.) ACL 2003. LNCS (LNAI), vol. 2922, pp. 136–143. Springer, Heidelberg (2004)
18. Krovetz, R., Croft, W.B.: Lexical ambiguity and information retrieval. ACM Trans. Inf. Syst. 10(2), 115–141 (1992)
19. Liu, et al.: Word sense disambiguation in queries. In: CIKM 2005, pp. 525–532 (2005)
20. Budanitsky, A., Hirst, G.: Evaluating WordNet-based measures of lexical semantic relatedness. Comput. Linguist. 32(1), 13–47 (2006)

Preservation of Evidence in Case of Online Gaming Crime

Patrick S. Chen[1], Cheng-Yu Hung[2], Chiao-Hsin Ko[3], and Ying-Chieh Chen[4]

[1,2] Dept of Information Management, Tatung University, Taipei, Taiwan
No. 40, Sec. 3, Jhongshan N. Rd., Jhongshan District, Taipei City, 104, Taiwan
chenps@ttu.edu.tw,
james340652@hotmail.com
[3] Dept of Information Management, Central Police University, Taoyuan, Taiwan
No. 56, Shujen Rd., Takang Village, Kueishan Hsiang, Taoyuan County, 333, Taiwan
heartlet@hotmail.com
[4] Dept of Information Management Chiao Tung University, Hsinchu, Taiwan
No. 1001, Ta Hsueh Rd., Hsinchu, 300, Taiwan
bomy@npa.gov.tw

Abstract. Along with the rapid growth of network applications and digital contents, online gaming has become a very successful industry in recent years. It not only brings tremendous business opportunities but also attracts marvelous customers to be online players. However, due to the lack of virtual property protection and related information security schemes, more and more players have violated the law because of its real world profit. Unfortunately, online gaming crime has turned out one of the most serious cybercrimes in many countries, such as in Taiwan, South Korea, China, and so on. In order to solve this judicial problem and effectively protect virtual property, we propose a Virtual Property Description Language (VPDL) as the syntax to record and express virtual property. There are six models, including Identity Model, Ownership Model, Trading Model, Content Model, Revocation Model, and Security Model. It can effectively protect virtual property and strengthen the competence of digital evidence by recording the legal source, ownership, trading track, and handling of virtual properties so as to lower the cases of online gaming crime. The online gaming companies may use it in the aspects of virtual property management, tracing historic records and collecting legal digital evidence, while the existing system mechanism of online game doesn't need to be changed.

Keywords: digital evidence, online gaming, virtual property, Virtual Property Description Language.

1 Introduction

Along with the fast-pace development of the Internet, various network applications have been developed, including online gaming, electronic commerce, online auction, online banking, and so on. Among them, online gaming is most successful. According to the statistical data of the MIC [1], in 2004, the online game market was NTD 7.22 billion. Compared with the year 2003, the online game market grew by 5%. In 2006,

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 229–244, 2008.

the whole output of online game in Taiwan was NTD 8.78 billion, gaining 16.6% compared with the year 2005. In 2007, the whole output of online game in Taiwan was NTD 9.5 billion, gaining 9.2% compared with the year 2006 [2]. According to a report from analyst firm DFC Intelligence [3], the worldwide online game market is forecasted to grow from $3.4 billion in 2005 to over $13 billion in 2011.

Online gaming is indeed prosperous, for instance, the online game "Lineage", which was developed by NCSoft.com of South Korea, has reached US $200 million revenue in 2001 and US $500 million in 2005 [4]. In 2004, more than 2.5 million players in South Korea and 2.6 million players in Taiwan are their online members. The number of Lineage players occupied almost one-fourth of all network users in both countries.

Virtual properties in online gaming have a very high value in the present market. The trading of virtual property, such as players in the game, virtual currency, virtual equipments and related virtual items, has become common practice. Since the number of virtual properties is limited and some virtual equipment cost time and energy for developing, many players who need these assets would like to trade for them. Out of the imbalance of supply and demand, some virtual properties have very high values in the marketplace. For instance, one UserID valued at US $2,000 dollars in an auction website[1], and one virtual space station valued at US $100,000 dollars[2]. Even the virtual currency in an online game can be converted into cash through exchange with other players. The value of these virtual properties might exceed real assets and virtual property trading is indeed prosperous and flourishing. When virtual properties become valuable in the real world, online game is no longer just entertainment. The involvement of money can easily lead to conflicts of profit, resulting criminal behaviors [5].

According to the statistics from Consumers' Foundation of Taiwan, the number of online gaming complaints has turned out the most serious problem. The foundation shows that there are 909 complaints related to online gaming complaints within 5,947 cases in 2005 [6]. Furthermore, when complaints or criminal cases needed to go to arbitration, these sometimes caused null or insufficient digital evidence so as to unable to clamp down offenders. More serious problems binding multinational gangsters have evolved international hi-tech criminal problems [7].

Most of the criminal cases are related to virtual properties since real markets have developed for the virtual properties giving them real world values. Take example from Taiwan, the number of thefts, frauds, robberies cases from online gaming has increased to 1300 cases from 55 only 2 years earlier and the amount of cases have turned out the most serious cybercrime. Another analysis paper of online gaming characteristics shows that the majority of online gaming crime is theft (73.7%) and fraud (20.2%) [8]. The age of offenders is quite low (63% in the age range of 15-20), and 8.3% of offenders are under 15 years old.

---

[1] http://cgi.ebay.com/ws/eBayISAPI.dll?ViewItem&item=8113549502&indexURL=0&photoDisplayType=2#ebayphotohosting, 2007.

[2] http://news.bbc.co.uk/1/hi/technology/4104731.stm, http://news.bbc.co.uk/1/hi/technology/4374610.stm, 2007.

Researchers in [9] indicate that the online game has security problems given below:

1. Practice fraud by a conspiracy;
2. Practice fraud by taking advantage of the vulnerabilities of game rule;
3. Defraud of the virtual properties with using other's accounts and passwords;
4. Conduct Distributed Denial of Service (DDoS) attacks against other players to make them attain unfair results;
5. Practice fraud by taking advantage of the game itself;
6. Attack by taking advantage of the player's authentication mechanism;
7. Practice fraud by the Internal staves themselves;
8. Get the virtual properties of the game by cheating;
9. Practice fraud by taking advantage of the vulnerabilities of game flow design;
10. Revise the game's programs to gain unfair benefits;

In order to prevent the possible security problems above, we propose a Virtual Property Description Language (VPDL) to effectively and efficiently protect virtual property. The original idea of VPDL is from the Open Digital Rights Language (ODRL), which is well applied to digital rights management[3]. In ODRL, the main protective objective is aimed at digital rights. Nevertheless, the application of ODRL is infeasible to deploy protection schemes on virtual property in diverse online gaming systems. Those protection schemes on virtual property includes providing legal source prove, ownership attribute, trading footprint, event content, and the history of tractions to trace the life of virtual properties. Therefore, we first proposed a draft on VPDL framework[4] [10]. Based on this draft, we will elaborate the content of VPDL in details with six models and implement the application so as to protect and track virtual property efficiently. When it necessarily comes to arbitration, it can be a valid and important source of digital evidence in judicial judgment.

The rest of this paper is organized as follows. Some researches on virtual property and the protection of virtual property are briefly introduced in Section 2. Then, we illustrate how to use the integrity protection mechanism as the security basis of VPDL in Section 3. Section 4 is the main part of this paper, in which we introduce the framework and models of VPDL, the relations between each model, and how VPDL works to protect virtual property. In Section 5, we evaluate the applications of VPDL to make readers have an intensive understanding of it. The conclusion is presented in Section 6.

2 Virtual Property

In this section, we illustrate virtual property on its definition, classification, characteristics, security issues, and discuss the references on virtual property and topics on the protection of virtual properties, so as to have an in-depth understanding on virtual property.

2.1 General Description of Virtual Property

A virtual property is an intangible asset existing in a virtual world, often in the context of online gaming. Due to the supply and demand principle in real marketplace, the

[3] http://www.ordl.net/, Accessed on June 5. 2007.
[4] One of the authors, Y. C. Chen, co-authored the paper [10].

original priceless virtual properties are given economic value by online players. These properties are observed in multi-user dungeons (MUDs) or massively multiplayer online role-playing games (MMORPG). The largest virtual properties are currently found in MMORPGs, such as EverQuest, GuildWars, Dark Age of Camelot, and Lineage.

Since the number of virtual properties is limited and some virtual equipment cost time and energy for developing, some players who need these assets would like to trade for them. This causes some virtual properties to have very high values in the marketplace. Take the Project Entropia (PE) online game for instance, the virtual currency can be converted into cash through exchange with other players. The virtual currency exchange rate was, for example, 10:1[5]

2.2 Security Worries on Virtual Property

Security worries on present virtual property are presented in the following forms:

1. *Lack of security protection:* Most of online gaming vendors only employ log records on protecting players' virtual property.
2. *No integrity protection on trading*: Because the digital data subject to unauthorized insertion, copying, revision and deletion, both trading parties have to effectively maintain the integrity of the electro-magnetic records.
3. *Obscure division on ownership and right of use*: In present gaming circumstance, the owner of the virtual property is unable to assign rights to other players, such as the right of use.
4. *Limited log records reserved*: Take the online game "Lineage" for example, there are over 200,000 concurrent users per second to play the online game simultaneously[6], whose log records are also growing extremely fast and occupy the limited storages. To save the cost, vendors only keep the log for one month. If a judicial authority asks for log records before one month, it would be a problem providing such evidence.
5. *Hard to trace digital evidence*: It is not easy to trace a suspicious or illegal demeanor from merely transactional records which is stored in a huge database. Take virtual currency for instance, virtual currency is generally unable to be personalized, so tracing a virtual currency becomes an impossible task and often leads to complicate forensic works.
6. *Unable to ensure trading security*: In present trading environment, there are several trading types related to virtual properties including sell, buy, rent, exchange, lease, and so on, but the buyer is unable to ascertain whether the object of trading is a stolen goods or not.
7. *User identification*: In most gaming systems, virtual property is protected by user identification mechanism, such as a pair of account and password. Therefore, once the user identification mechanism is attacked or cracked, the accompanied virtual property would be in danger.

[5] http://en.wikipedia.org/wiki/Virtual_economy, 2007] (10 virtual currencies can be converted to 1 US dollar, December of 2004.

[6] http://lineage2.nctaiwan.com/event/news_public_both.asp?news_type=3, June 2007.

8. *Forge or modify log records with ease*: As mentioned before, log records have been the main source of digital evidence. Nevertheless, most records are not well protected with secure mechanism by vendors.
9. *No standard language for describing the use of virtual property*: Without an exact description using a standard language, it would cause unnecessary disputes and problems in virtual property treatment, for example, the object of trading may be a stolen one.
10. *No signature mechanism*: The trading process is not signed by any fair third party, the fairness of the trading cannot be guaranteed and the gaming companies cannot guarantee the rights and interests of the players.

2.3 Previous Researches on Protecting Virtual Properties of Online Games

The VMPL has seven core entities as shown in Figure 1 [11], which proposes the logic concept with this model. Nevertheless in the seven core entities, the system contents and function of the attributes in each model are not described in detail. Therefore, we are going to revise the system framework and enrich the system contents and the functions of the attributes in each model so as to effectively manage, protect, and trace the virtual properties.

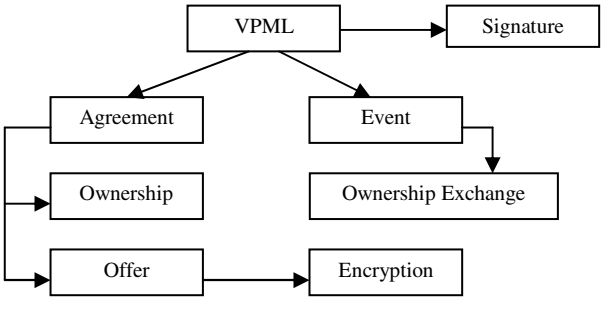


Fig. 1. VPML Core Entities [11]

How to maintain the integrity of digital records by an effective mechanism has remarkable impacts on the protection of virtual properties because the electro-magnetic records are susceptive to alternation and modification. Integrity protection method mostly applies to cryptography to add authentication code representing the original information. For example, calculate first the Message Digest by the one-way hash function [12] or calculate digital signature with private key [13], and then enclose the original file with the Message Digest and digital signature to protect the data. Hwang et al. has followed the theme and proposed a set of protection mechanism for electronic clinic reports by compound documents; they designed a framework applicable for the compound files of electronic clinic reports and proposed the protection mechanism based on this framework [14].

3 Establishment of Integrity Protection Mechanism

The definition of integrity, cryptography for integrity protection and introduction of the integrity protection mechanism into the virtual properties are generally described as follows.

3.1 Integrity

The data integrity has the following three meanings [15]:

1. Correctness of Data: The data must prevent from illegal or unauthorized modification, insertion, and deletion.
2. Authenticity of Data: The source of data must be legal, authorized and true; therefore the data transmitted need to have legal source and the creation and use of the data shall be authorized.
3. Consistency of Relevant Attributes: Presentation of the data must be consistent, for example the data structure or order. The presentation method includes encoding, for example, the Chinese uses Big5 code and English uses ASCII code etc. It further includes the authenticity of encryption key, consistency of data order relations etc.

3.2 Characteristics of Virtual Property Integrity Protection and Demands

Virtual property may be represented in a hierarchical structure and involves many unit files of proprietorship, contents, attribute, and previous event contents. Each unit file records the important information of virtual properties in detail. Therefore we must ensure that the relevant information are produced from the legal source. A compound document is composed of several unit files, and the authenticity and correctness of a virtual property must be maintained. Unauthorized operation shall be prevented in the course of information transmission and processing. In addition to the data integrity protection, the digital signature shall be used to ensure the legal source of the data.

3.3 System Implementation

In this paper we will propose a kind of integrity mechanism which will treat the virtual property as a compound document, and the compound document is composed of several unit files. The mechanism calculates the authentication code of each unit file by hash function. It further calculated the authentication code of the entire compound document in accordance with the structural relation between the virtual property and document. The examination of this compound document may enable the examiner to confirm the integrity of individual files and their order and structure, and whether they are involved in the authorized modification.

4 Virtual Property Description Language

In the following, we illustrate the language VPDL on its framework, models, and contents.

4.1 Framework of Virtual Property Description Language

VPDL is based on three core entities including virtual property, vendor and consumer. The term virtual property is a collective noun, which is uniquely identified and relates to virtual character, virtual weapon, virtual equipment, or virtual currency in online gaming. Vendor plays the role in creation, production, distribution of the virtual property, and can assert some form of ownership over the virtual property. Consumer is an end-user that consumes the virtual property over online gaming.

VPDL is a tailor-made design to meet the characteristics of virtual property and the relationship between vendor and customer. There are six functional models inside VPDL including identity model, ownership model, trading model, content model, revocation model, and security model. Within security model, two sub models such as encryption and digital signature mechanism providing a secure protection.

Each communication on models is bridged by an entity's unique identification code (UID). Entity is a term referring to dynamic objects such as a virtual property, an ownership owner, or a gaming vendor, and a unique identification code is composed of a number of alphanumeric characters referring to its virtual space, server, type, attribute, checksum, and so on.

Figure 2 depicts the overall framework and the relationship of six models for Virtual Property Description Language.

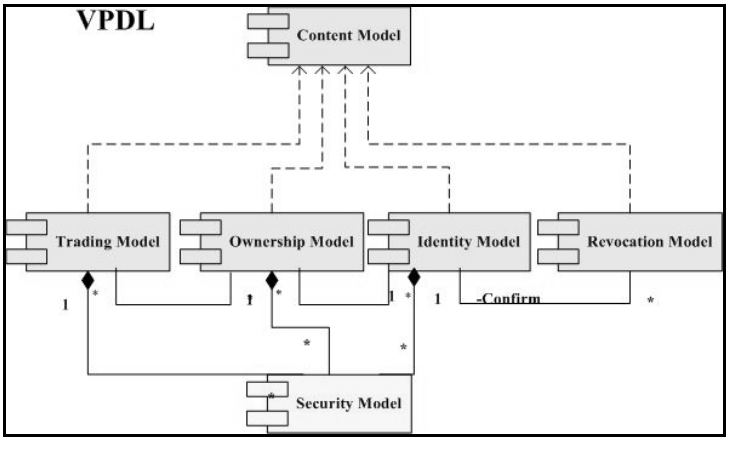


Fig. 2. Models of Virtual Property Description Language

4.2 The Functions of Six Models of Virtual Property Description Language

To avoid misunderstanding, the figures cover both the Description Language entities (shown as rectangles with "DE" in the left side) and the Data Dictionary elements (shown as rectangles with "DD" in the left side). Other namespaces such as "EN" for Encryption and "DS" for Digital Signature are also provided.

The following paragraphs discuss the six models of VPDL in detail.

1. Identity Model

In a real world, a buyer who wants to purchase valuable merchandise would ask the seller for providing its source identification or related documents in order to eliminate potential disputes and conflicts, such as avoiding stolen goods.

The main objective of identity model is to provide source identification for virtual property, which has a legal identity certifying from the gaming vendor or a trusted third party. From the source identification, it can accurately prove the virtual property is not fake, stolen, or falsified goods. Since virtual property is a magnetic or electronic data, it has the following special characteristics, such as easy to duplicate, modify and falsify. For the sake of solving this identity problem of virtual property, we design the identity model into Virtual Property Description Language to provide identification information on virtual property.

Identity model is allowed to combine with security model which deploys several cryptographic methods to provide a necessary protection on virtual property. For instance, a cryptographic hash function is used to prove the integrity of primitive message. A hash function takes a long string or message of any length as input and produces a fixed length string as output, sometimes termed a message digest or a digital fingerprint. Digital signature is also deployed in security model to provide the source identification of virtual property, and it has some benefits on its authentication, integrity, and non-repudiation. Digital signature mechanisms can be used for identifying the originator of an electronic message and ensure the legal intent being cryptographically secure. In other words, the source identification on virtual properties is similar to an identification card on human beings. Therefore, an outcome of Virtual Property Description Language is also possessed the competence of evidence and legal responsibility.

In case of any dispute, there exists a signed business arbitration authority to be followed in order to guarantee the rights and interests of customers and gaming companies. Additionally, in terms of technical level, the digital signature may have the characteristics of integrity and non-repudiation with the adoption of public key system. By virtue of the public key of the signer, it can validate whether the identity data of the virtual properties is issued really by the gaming company (or fair third party), and the relevant message digest, and the contents and operation principle of the digital signature are described in detail in the security model.

In addition, a warning mark is designed into identity model. If a virtual property is under investigation or being revoked, there would be a warning mark manifested on source authentication information that caution buyers not to buy, rent, give away, or other trading activities on it. Once a virtual property is marked with a warning sign, it represents the owner of virtual property is not allowed to use it temporarily so as to keep the evidence and the basis for post-tracing evidence.

In identity model of VPDL, it regulates the following elements:

- Unique identification code for virtual property(UID)
- Date and Time of production for virtual property(Date/Time)
- Vendor
- Special Attribute
- Warning
- Annotation

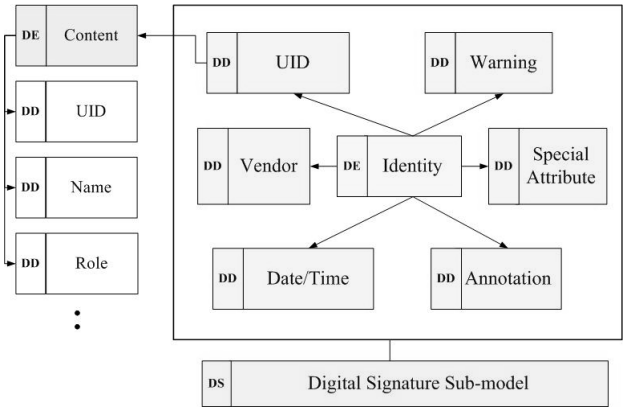


Fig. 3. Contents of Identity Model

In the identity model, various attributes are connected in series by the identity, including the UID, preparation date and time, issuing company, data of special attribute, warming marks and annotations etc. According to the UID, the content models may be reconnected to bring out the detailed relevant attribute data of the virtual properties, as shown in Figure 3 Contents of Identity Model.

2. Ownership Model

The concept of ownership can be applied to virtual property or virtual merchandise. Take the open digital rights language (ODRL) for example, there are more than five models, expressing and regulating rights for digital content, such as permission, constraint, requirement, condition, digital rights holder and offer models. For simplifying gaming environment, we do not deploy to functional models on rights regulation. In VPDL, ownership model is designed to express the virtual property ownership, and establish the complete records on each ownership transformation. In default definition of VPDL, once the owner acquired the ownership of virtual property; the owner had the entire rights on virtual property, such as sell, rent, transfer, revoke rights, and so on. An owner of ownership can be humans, organizations, and defined roles.

The right of use is also considered to employ in ownership model. Since there is still existed obscure division on ownership and right of use, the trading activities of virtual property have been hindered and unable to have a prosperous commercial application. For solving this problem, we deployed a mechanism of right of use into managing virtual property, which could easily traverse the present limitations, and provide a friendly environment for various trading requirement. For example, David has the ownership of one virtual property, and rented it out to Mary for two months one week ago. Therefore, during the period of the tenancy, David is not allowed to sell or re-rent it to others unless he revokes the lease contract. Besides, David still has the ownership of the virtual property, and Mary has the right of use during the period of the tenancy. Although Mary has the right of use, she is also not allowed to re-rent or sell the virtual property to others until she had the right of ownership.

In ownership model, there are several elements and attributes inside, such as the unique identification code for virtual property, owner with ownership and user with the right of use, complete name, credit evaluation, period of owning virtual property, current situation, the way of acquiring virtual property, source of virtual property, and related elements, so as to record and express an entire information regarding virtual property, owner, and user. A buyer, who wants to buy a virtual property, can first check the current situation of virtual property through its ownership information. After a preliminary scrutiny from ownership information, the buyer can make a right decision on renting or buying the virtual property.

When there is a dispute occurred, ownership model information can provide complete history records on ownership of virtual property, which can be an important reference information or digital evidence for a judicial arbitration. In figure 4, it depicts the content of ownership model.

In ownership model of VPDL, it regulates the following elements:

- Unique identification code for virtual property
- Unique identification code for owner of virtual property
- Unique identification code for user of virtual property
- Ownership type
- Period of owning virtual property
- Period of renting or borrowing virtual property
- Current condition
- Credit evaluation
- Source of acquiring virtual property
- Annotation

3. Trading model

VPDL supports various kinds of trading types on selling, buying, transferring, giving, renting virtual property, and so on. All processes of trading should be correctly recorded and prevent possible alteration. Any transformation on ownership of virtual

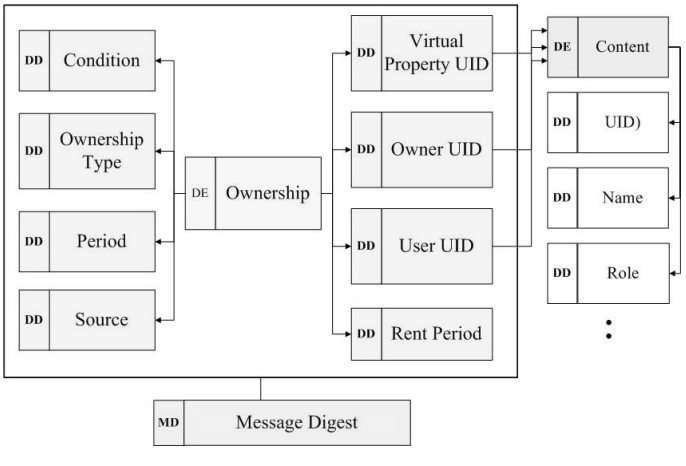


Fig. 4. Content of Ownership Model

property shall be operated through trading model as well, and it will simultaneously affect the ownership information. In fact, trading model refers to the ownership information before launching a trading, and make sure the seller has the rights to process trading in order to eliminate potential disputes on trading; for example, fist check the seller owning the virtual property and no lease contract on it. Besides, if there is any confidential or sensitive data inside the content of trading, encryption sub-model is provided to encrypt the data or related elements for offering secure protection. When a dispute appears, an intermediary can make a judgment through the detailed trading information and even recover the crime scene. The detailed information includes unique identification code for virtual property, seller and buyer, trading type, price, date/time, IP address, payment, and so on. It can provide the necessary information for arbitration. Additionally, trading model is combined with a cryptographic hash function to prevent potential modification on trading information, and it has more power of digital evidence in a court. In Figure 6, it depicts the content of trading model.

In order to meet the actual environment demands, VPDL supports various types of the virtual properties including purchase, selling, borrowing, leasing and granting. When the virtual properties are under the state of leasing or lending, the trading is conducted between both seller and buyer, and the trading model will refer to the relevant information of the Ownership Model of the virtual properties. If the current status attribute of the virtual properties displays that the present traded article is still under leasing or lending, any trading behavior is suspended, except leasing or lending status has been suspended or cancelled. In another word, when the trading model finishes both parties' leasing or lending agreement, the relevant information will be written back into the Ownership Model. If the current status attribute of the Ownership Model is changed as leasing, the information will be written into the lender's UID. In Figure 5, the contents of trading model are indicated.

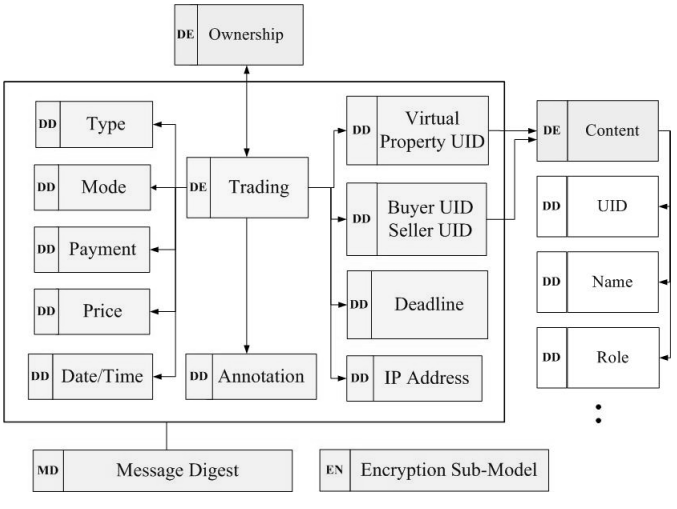


Fig. 5. Content of Trading Model

In trading model of VPDL, it regulates the following elements:

- Unique identification code for virtual property(UID)
- Unique identification code for seller of virtual property(Seller UID)
- Unique identification code for buyer of virtual property(Buyer UID)
- Trading type: such as buy, sell, borrow, give, exchange, rent, lease
- Price
- Trading date and time
- Trading deadline
- Trading mode: such as on the trading platform,
- Payment
- IP address
- Annotation

4. Content Model

The Content model plays an important role in identifying the entity on using a unique identification code to recognize virtual property, owner, buyer, etc. This ability to uniquely refer any entity can be utilized in providing linkages between entities. In Figure 6, the contents of content model are indicated.

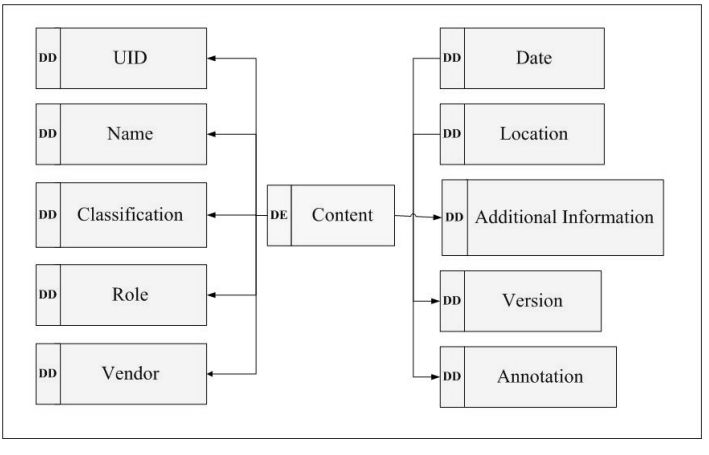


Fig. 6. Content of Content Model

In content model of VPDL, it regulates the following elements:

- UID: Unique identification code for entity: such as virtual property, owner, buyer, seller
- Name of entity
- Classification: such as virtual character, weapon, equipment, currency, space, related attributes
- Role: the role of entity play in virtual world
- Vendor or issuer
- Space name
- Date

- Location
- Additional Information: it may refer to a Uniform Resource Identifier (URI), which includes a Uniform Resource Name (URN) and Locator (URL). For example, a virtual space station at number 1002 on game.com can represent to "URN: game.com: spacestation1002" through the way of Uniform Resource Name. A virtual glove belongs to a virtual equipment on www.game-station.com may refer to "Http://www.game-station.com/equipment/glove41/" as the Uniform Resource Locator form.
- Version: version illustration on entity
- Annotation

5. Revocation model

VPDL supports revocation on different kinds of entities, such as virtual property, owner, group owner, or defined roles. The revocation of entity may be launched by gaming vendor or an end-user with ownership. Once the virtual property is revoked, the ownership and the power of use are withdrawn as well. In other words, after revocation, the property's owner or user can not operate it anymore until ownership has been applied. If the current condition of virtual property is showing lease or rent in ownership model, the owner is not allowed to revoke the virtual property as well. In Figure 7, the contents of Revocation Model are indicated.

In revocation model of VPDL, it regulates the following elements:

- Unique identification code for virtual property
- Unique identification code for owner of virtual property
- Date/Time of revocation
- IP address
- Annotation

6. Security model

VPDL supports secure protection mechanisms including encryption sub-model, digital signature sub-model, and Message Digest technique. Encryption sub-model is inherited

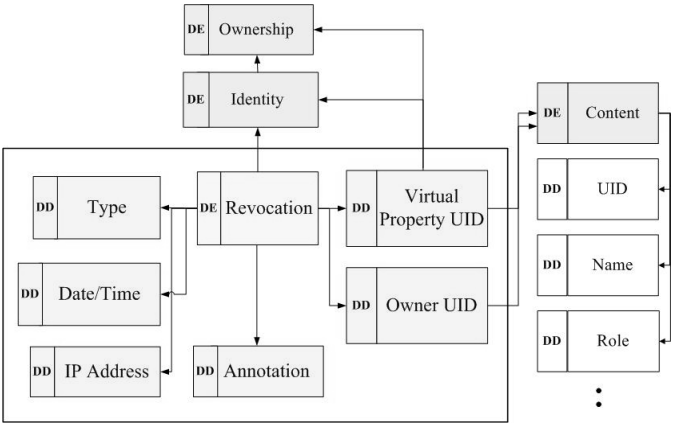


Fig. 7. Content of Revocation Model

from W3C XML Encryption (XML-ENC) standard[7]. And digital signature sub-model is inherited from W3C XML Signature (XML-SIG) standard[8]. So the security model can assure the operation among entities and the protection of virtual property.

- The encryption sub-model

The encryption sub-model provides confidential protection for the sensitive or personal data in VPDL such as credit card account and trading content. When these is a dispute occurred, the arbitrator or the trusted third party can decrypt the cipher data and recover the content of trading. In addition, the digital signature is only on the cipher data so that the decryption is not required for verification, so it makes VPDL more efficient and secure since only the seller knows the payment information.

- Encryption Method: encryption algorithm used in the encryption entity, e.g. 3DES;
- Key Info: key information or value used for the encryption, a session key value may require an encryption with public key system;
- Cipher Value: encrypted data with the above encryption algorithm and key.

- The signature sub-model

The signature sub-model provides non-repudiation protection for the trading and virtual property identity in VPDL so that any player can check if the owner is a real owner of the virtual property.

The signature mechanism of VPDL may include the following attribute contents:

- Digest Method: such as hash algorithm used in the signature sub-model, e.g. SHA-1, SHA-256;
- Digest Value: hashing result with the hash algorithm;
- Signature Method: signature algorithm used in the signature sub-model, e.g. RSA;
- Key Info: public key certificate used for the signature;
- Signature Value: signature result with the above signature algorithm and key.

5 Overall Evaluation

Compared with the traditional Log Files, the VPDL in tree data structure format using XML possesses the advantage of easily judging and reading information logically. In terms of application by the common user, the XML volume label may coordinate with .NET Web Service and use the Web interface to offer the relevant operations. For game players, only in accordance with the VPDL data frame to establish the relevant databases, they may protect the virtual properties of online games more safely and conveniently. For disputes or crimes, the investigators and examiners may find out the relevant historic data and logic relationship of the event rapidly.

6 Conclusion

We get to the bottom of on-line game brought by the serious network disputes and crime problems, largely because of the lack of virtual property protection mechanisms. In order

[7] http://www.w3.org/Encryption/2001/, June 6. 2007.
[8] http://www.w3.org/Signature/, June 6. 2007.

to solve the problem of protecting virtual property, we revised the system framework of the previous researches and enrich the system contents and the functions of the attributes in each model so as to effectively manage, protect, and trace the virtual properties. Furthermore, VPDL can guarantee the rights and interests of online gaming players, and avoid various online gaming crimes. Also, for the implementation of these measures, lots of digital evidences will be accumulated. Once any dispute or crime occurs, these evidences will be used for investigation and produced in evidence. Under this frame, six system models will be formed in order to provide an effective protection mechanism including the certification of the identity source of virtual property, ownership subordination, trading course, digital signature, and encryption Through overall evaluation, without changing the existing system mechanism of the online games, the gaming companies may attain the goal of protecting the virtual properties within the shortest time, and may use it in virtual property management, following trace and court digital evidence. If the game player wants to manage or process his virtual properties, the game player may judge and read the relevant important information of the virtual properties according to the meaning of the Virtual Property Description Language. In the future, we will establish on the WEB service and may provide owner unload, browse Virtual Property Description Language, i.e. VPDL file, and provide the interface mechanism for subscriber validation VPDL file. For game player, we expect that they can effectively manage the historic event of virtual properties by mean of the VPDL to reduce the disputes on virtual property and we expect to provide better service for game players. Because virtual property will be protected well by this system, the crimes taking the game virtual properties as the subject matter will be reduced comparatively, and the order maintenance of virtual property will be facilitated. Even if a crime derived from virtual property occurs, the VPDL may provide definite digital evidence facilitating the judicial authority to understand fully the case and treat the cases just and fair.

References

1. Lin, Y.S., Chou, S.L.: Online Games Brings up Valuable Ways (in Chinese), MIC, C6 Edition, Economic Daily News (2006)
2. Wang, S.W.: No-lives Live out their second lives through Online Games (in Chinese), Common Wealth Magazine, No.379 (2007)
3. DFC Intelligence, Online Game Market Forecasted to Reach $13 Billion by 2011 (858) 780-9680 (2006)
4. NCSoft, Earnings Release 4Q 2003, NCsoft Corp. (2004)
5. Police Administration Bureau, Police Administration Statistical Aviso (in Chinese), No. 35 (2006)
6. Chen, L.J.: Consumption Dispute Online Games Topped Last Year (in Chinese), C4 Edition, Taipei Overall, The United Daily News (January 27, 2006)
7. Jamieson, J.: Online gaming portal for crime Cyber-crime - expert warns the problem could find its way to North America, The Province Newspaper, Vancouver, Canada (2004)
8. Chen, Y.C., Chen, P.S., Hwang, J.J., Korba, L., Song, R., Yee, G.: Analysis of Online Gaming Crime Characteristics. Journal of Internet Research (2005)
9. Yan, J.J., Choi, H.J.: Security Issues in Online Games. The Electronic Library: International journal for the application of technology in information environments, 3–8 (2001)

10. Song, R., Korba, L., Yee, G., Chen, Y.C.: Protection of Virtual Property in Online Gaming. In: Proceedings of 2005 Conference on Distributed Multimedia Systems, Banff, Alberta, Canada (2005)
11. Song, R., Korba, L., Yee, G., Chen, Y.C.: Protect Virtual Property in Online Gaming System. International Journal of Software Engineering and Knowledge Engineering (IJSEKE) 17(4), 483–496 (2007)
12. Kaihara, S.: Realization of the computerized patient record - relevance and unsolved problems. International Journal of Medical Informatics (49), 1–8 (1998)
13. van der Haak, M., Wolff, A., Brandner, R., Drings, P., Wannenmacher, M., Wetter, T.: Data security and protection in cross-institutional electronic patient records. International Journal of Medical Informatics (70), 117–130 (2003)
14. Hwang, J.J., Chou, S.L., Yeh, Y.J., Tsai, R.L.: Protecting information integrity for electronic patient records. Journal of Information Management 13(4) (2006)
15. Huang, C.C.: Information Security-Basis of E-commerce (in Chinese). Hwa Tai Publishing, Taipei (2001)

Dataset Analysis of Proxy Logs Detecting to Curb Propagations in Network Attacks*

Da-Yu Kao[1], Shiuh-Jeng Wang[2,**], Frank Fu-Yuan Huang[1], Sajal Bhatia[3], and Saurabh Gupta[3]

[1] Department of Crime Prevention and Corrections, Central Police University, Taoyuan, Taiwan
[2] Department of Information Management, Central Police University, Taoyuan, Taiwan 333
sjwang@mail.cpu.edu.tw
[3] Department of Communication and Computer Engineering, LNM Institute of Information Technology, Jaipur, India

Abstract. The exponential growth of Internet has brought a monolithic change in the level of malicious attacks, leading to the emergence proxy servers, which indeed have proven to apotheosis hiding place for the Internet intruders. The collected logs of these proxy servers contain portentous information, and its dissection can help in analyzing the deviation of abnormal activities form the normal ones. How to figure out their network of networks, identify possible offenders, and strike the heartland of their safe haven has become an upcoming challenge for universal law enforcement agents. This paper considers exactly what kind of elements should be explored once an offensive behavior has been noticed in proxy logs. It scrutinizes (i) the Time Stamp gap of sequential records (ii) the parameters of digital action (iii) the appearance of special parameters (iv) the patterns in the log files.

Keywords: Proxy Logs, Linkage Analyses, Network attack, Computer Crime.

1 Introduction

The network attack analysis process involves three main procedures: initial response, media imaging duplication, and imaged media analysis. Our proposed method focuses on the procedure of imaged media analysis. This paper describes how the data mining methodologies can be applied to the numerous log based information, which can derive the top facts in each of the diverse connections and locate malicious events spread across the network.

2 The Value of Auditing Log Analysis

The main purpose of this following proposed model describes how forensic techniques have been applied to the million of daily auditing records, and provides a rapid feedback of investigation coverage for network inspection [4].

* This work was supported in part by National Science Council in R.O.C. under Grant No NSC-96-3114-P-001-002-Y.
** Corresponding author.

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 245–250, 2008.

This paper focuses on establishing strong, legitimate links between individuals and digital information by analyzing logs in an intrusion event. These logs contained data pertaining HTTP requests over the past periods. The original identifiable information is modified for the academic research purpose. Fig. 1 shows the format of a CCProxy log and illustrates the following information:

Time Stamp: Client IP Address: Digital Action I: Server IP Address: Digital Action II

No.	Time Stamp	Client IP Address	Digital Action I	Server IP Address + Digital Action II (File Location and Parameter)
1	2006/12/11 22:04	100.128.194.77	unknown Web GET	http://160.17.225.8/jkcboard/upload/HT_tw.asa
2	2006/12/11 22:04	100.128.194.77	unknown Web POST	http://160.17.225.8/jkcboard/upload/HT_tw.asa?pageName=default&theAct=chkLogin
3	2006/12/11 22:04	100.128.194.77	unknown Web GET	http://160.17.225.8/jkcboard/upload/HT_tw.asa?pageName=server
4	2006/12/11 22:04	100.128.194.77	unknown Web GET	http://160.17.225.8/jkcboard/upload/HT_tw.asa?pageName=server&theAct=showService
5	2006/12/11 22:04	100.128.194.77	unknown Web GET	http://160.17.225.8/jkcboard/upload/HT_tw.asa?pageName=cmdShell
6	2006/12/11 22:04	100.128.194.77	unknown Web GET	http://160.17.225.8/jkcboard/upload/HT_tw.asa?pageName=fso
7	2006/12/11 22:05	100.128.194.77	unknown Web GET	http://160.17.225.8/jkcboard/upload/HT_tw.asa?pageName=fso&thePath=D%3A\wwwroot\jkcboard\
8	2006/12/11 22:05	100.128.194.77	unknown Web GET	http://160.17.225.8/jkcboard/upload/HT_tw.asa?pageName=fso&thePath=D%3A\wwwroot\jkcboard\upload%2Easp&theAct=edit
9	2006/12/11 22:05	100.128.194.77	unknown Web GET	http://160.17.225.8/jkcboard/upload/HT_tw.asa?pageName=fso&thePath=D%3A\wwwroot\jkcboard\uploadok%2Easp&theAct=edit
10	2006/12/11 22:05	100.128.194.77	unknown Web GET	http://160.17.225.8/jkcboard/upload/HT_tw.asa?pageName=stream&thePath=D%3A\wwwroot\jkcboard\uploadok%2Easp&theAct=down
11	2006/12/15 19:47	100.128.194.77	unknown Web GET	http://cdtower.fcu.com.tw/HT_tw.asa
12	2006/12/15 19:47	100.128.194.77	unknown Web POST	http://cdtower.fcu.com.tw/HT_tw.asa?pageName=default&theAct=chkLogin
13	2006/12/15 19:47	100.128.194.77	unknown Web GET	http://cdtower.fcu.com.tw/HT_tw.asa?pageName=server
14	2006/12/15 19:47	100.128.194.77	unknown Web GET	http://cdtower.fcu.com.tw/HT_tw.asa?pageName=server&theAct=showService
15	2006/12/15 19:47	100.128.194.77	unknown Web GET	http://cdtower.fcu.com.tw/HT_tw.asa?pageName=server&theAct=showUser
16	2006/12/15 19:48	100.128.194.77	unknown Web GET	http://cdtower.fcu.com.tw/HT_tw.asa?pageName=fso
17	2006/12/15 19:48	100.128.194.77	unknown Web GET	http://cdtower.fcu.com.tw/HT_tw.asa?pageName=fso&thePath=C:\
18	2006/12/15 19:48	100.128.194.77	unknown Web GET	http://cdtower.fcu.com.tw/HT_tw.asa?pageName=fso&thePath=C%3A\back
19	2006/12/29 00:02	100.128.194.77	unknown Web PUT	http://100.128.113.241/HT_tw.asa
20	2006/12/29 00:04	100.128.194.77	unknown Web PUT	http://100.128.113.241/upload/HT_tw.asa
21	2006/12/29 00:12	100.128.194.77	unknown Web PUT	http://100.128.113.241/upload/HT_tw.asa
22	2006/12/29 00:13	100.128.194.77	unknown Web PUT	http://100.128.113.241/upload/HT_tw.asa
23	2006/12/31 00:39	100.128.194.77	unknown Web GET	http://www.mdhs.tc.com.tw/Web-SYS/MDDownloadDoc/data/20051231039106 62-HT_tw.asa
24	2006/12/31 00:39	100.128.194.77	unknown Web POST	http://www.mdhs.tc.com.tw/Web-SYS/MDDownloadDoc/data/20051231039106 62-HT_tw.asa?pageName=default&theAct=chkLogin
25	2006/12/31 00:39	100.128.194.77	unknown Web GET	http://www.mdhs.tc.com.tw/Web-SYS/MDDownloadDoc/data/20051231039106 62-HT_tw.asa?pageName=server

Fig. 1. An Example of CCProxy Log

- *Time Stamp:* Seeing the data in this proxy log file it could be concluded that, if the time gap of sequential records is longer than two minutes, we can believe that it could be a sign of some abnormal activity by human being. The records from No. 19 to No. 21 are one of the best examples.
- *IP Address:* The source computer of launching backdoor attack is always the '100.128.194.77' IP Address, which is also recognized as a public firewall IP address.
- *Digital Action:* The information of digital action is divided into two parts. The first part represents that the attacker uses an unknown account to access the web service and to get or post some packets by way of this proxy server. The second part describes the relevant program information, which contains the file name, location, parameter, and etc. This information is useful for further analysis.

3 Our Proposed Approach

3.1 Detect Network Attacks

The following section discusses our methodologies in reference to the three main phases: Data Pre-Processing, Pattern Discovery, and Pattern Analysis [2]. Table 1 outlines these sequential decision making procedures and their main functions.

Table 1. Sequential Decision Making Procedures

Phase	Procedures	Main Function
I. Data Pre-Processing: Clarify the Relevant Characteristics	Image Duplication	Backup the source logs
	Refine processing	Figure out specific parameters
	Input processing	Compare the parameters
II. Pattern Discovery: Find Out the Intruder Pattern	Pattern observation	Observing specific patterns
	Pattern searching	Parameter filtering
	Pattern redistribution	Active monitor parameters
III. Pattern Analysis: Pattern-oriented Analysis	Content analysis	Passive decode parameters
	Linkage network analysis	Organize as a chart

Phase I: Data Pre-Processing- Clarifying the Relevant Characteristics
This phase is done to remove useless information and has three main components [1, 2]:

- *Image Duplication*: A backup of the source log is created and the subsequent examination is undertaken on the imaged files.
- *Refine Processing:* This step works on procedures attempting to extract proactive information from massive datasets and promotes effective analysis.
- *Input Processing*: It is concerned with transferring the datasets into another format to deal with semi-automated data mining.

Phase II: Pattern Discovery- Find out the Intruder Pattern
A brief observation of the log files can outline the intruder's activities. This is illustrated as follows:

- *Pattern Observation:* The examiner can identify the entrance and exit of any intrusion. The following two items of the datasets should be considered for the further analysis of observational pattern.
  - *IP address:* After examining the auditing information, the analyst finds out that the intrusion has come from a certain IP address. The consequent scrutiny reveals that it belongs to a public firewall IP address which hides many source addresses behind it.
  - *Specific Characteristics:* If the source computer is the same, it is believed so are the version of client computer OS (Operating System) and the browsing AP (Application Program). Hence, the investigators can double check on the above characteristics. In most situations, the intruders will use the particular programs, port numbers, or program functions to penetrate the compromised server computers.
- *Pattern Searching:* Certain string matching procedures are necessary to be implemented to facilitate this process. The crucial elements of network log are serviceable references in Table 2.

Table 2. Crucial Elements in Network Log

Elements	Option1: Pattern Concerns	Option2: Action Concerns	Option3: Intruder's Concerns
IP Address	Source or Destination	Single or Multiple	Static, Dynamic, or Proxy
Time Stamp	Sequential or Range	By Program or Human-being	Time Zone, Daylight Saving Time, or Synchronization
Digital Action	Visible or Hidden	True or False	Common or Special Parameters
Response Message	Success or Failure	Recorded or Missed	Reasonable or Unreasonable

- *Pattern Redistribution:* This step parcels out the meaning of selected pattern and its investigative context. Our empirical analysis reveals that log analysis can identify as many substitutes as complements.

Phase III: Pattern Analysis- Pattern-oriented Analysis

This phase makes the investigator aware of the important enigmatic information. It consists of the following procedures:

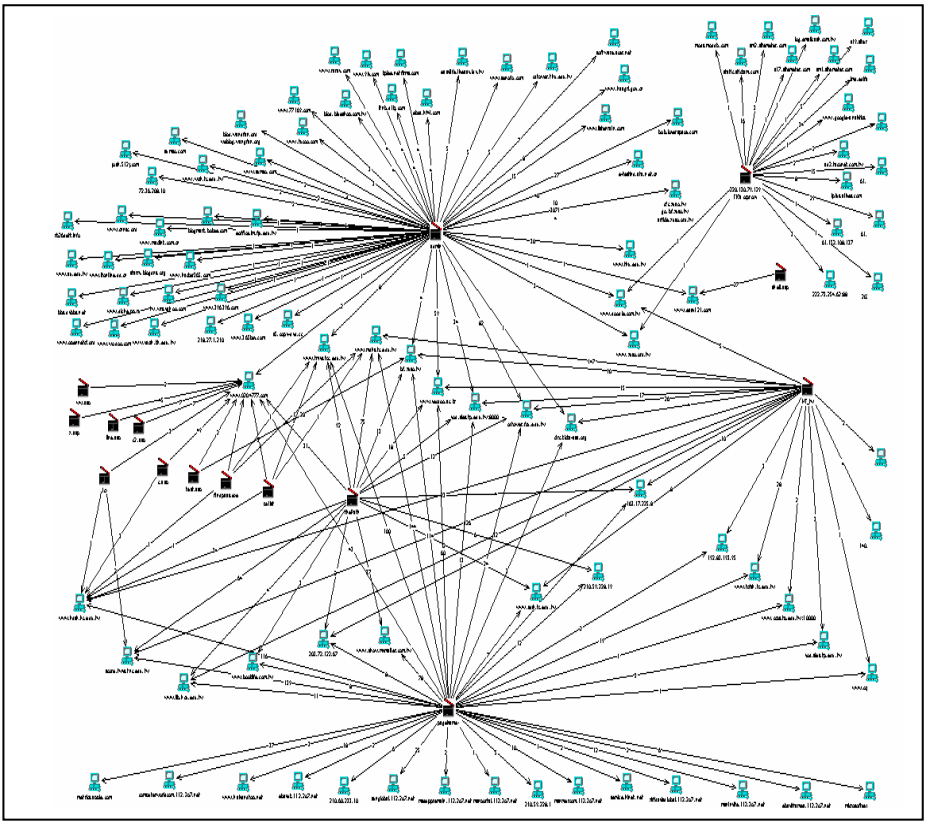


Fig. 2. A Sample of Linkage Network Analysis

- *Content Analysis:* This step focuses on investigating operating pattern, detecting anomalous behavior, and getting the evidence discovery of better insights into the data flow [3].
- *Linkage Network Analysis:* Examiners are required to discover elements that co-occur frequently within datasets consisting of multiple independent correlations of co-occurring elements, such as IP addresses, program names, and etc. Fig. 2 illustrates a sample of linkage network analysis on the elements of IP address and program name, and reduces potentially huge information to a small, understandable set of graphically supported elements.

3.2 Prevent Network Attacks

From the previous observation, the threat of malicious activity lies not in the information of unwariness but in the improper use of that information, which is already held by various auditing systems. The integration and expansion of auditing database promises greater security possibilities which are appropriate to boost future defense. Thus, we focus our attention on preventive measures to reduce the opportunities of network attacks, or take proactive steps to shield our particular information from high-risk Internet activities.

3.2.1 Limitations of Our Approach

The growth of Internet has lead to stupendous increase in the nature of network attacks. Therefore, every intrusion detection technique is bound to have some limitations like Intruder signatures of unknown behavior, current threats of zero day attack, and encrypted problem of packet data.

3.2.2 Possible Enhancements

Even though our proposed method has certain limitations, the scope of improvement is always there like regularly updating the database, implementing parallelization at the hardware level to reduce the analysis time, and using latest anomaly based tools like I2 or Adaptive Security Analyzer (ASA).

4 Conclusion

In this paper, we considered the problem of mining on proxy logs and proposed a mechanism which could detect malicious network activities and develop a preventive strategy for future attacks.

References

1. Fei, B., Eloff, J., Oliver, M., Venter, H.: Analysis of Web Proxy Logs. In: IFIP International Conference on Digital Forensics, Orlando, pp. 247–258. Springer, Heidelberg (2006)
2. Jone, K.J., Bejtlich, R., Rose, C.W.: Real Digital Forensics: Computer Security and Incident Response. Person Education, Inc. (2006)

3. Marshall, B., Chen, H.: Using Importance Flooding to Identify Interesting Networks of Criminal Activity. In: Mehrotra, S., Zeng, D.D., Chen, H., Thuraisingham, B., Wang, F.-Y. (eds.) ISI 2006. LNCS, vol. 3975, pp. 14–25. Springer, Heidelberg (2006)
4. Pollitt, M., Whitledge, A.: Exploring Big Haystacks – Data Mining and Knowledge Management. In: IFIP International Conference on Digital Forensics, pp. 247–258. Springer, Heidelberg (2006)

Identifying Chinese E-Mail Documents' Authorship for the Purpose of Computer Forensic

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

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

Abstract. Nowadays, computer crimes involving e-mail increases rapidly. To prevent these phenomena from happening, the authorship identification methods for Chinese e-mail documents were described in this paper, which could provide evidence for the purpose of computer forensic. The theoretical framework was presented. Various style features including linguistic features, structural characteristics and format features were analyzed. The support vector machine algorithm was used for learning algorithm. To validate the methods, experiments were made on limited dataset. The results were satisfying, which proved that the methods were effective and feasible to apply to computer forensic.

Keywords: Computer forensic, Chinese e-mail, authorship identification, Support Vector Machine, feature extraction.

1 Introduction

With the rapid growth in computer technology and information level, especially the increasing popularization of Internet, e-mail has become an expedient and economical form of communication. However, the crimes increase by means of e-mail, such as antisocial mail, fraud mail, racketeering mail, terroristic threatening mail, pornographic mail, virus and junk mail etc, which do harm to people's daily life, even affect social stabilization and national security. It is the time to take some effective measures to prevent these phenomena.

The current methods are merely some passive defending measures such as e-mail filtering, installing firewall, etc. These methods are not effective and cannot put an end of the e-mail crime. So the ability to provide evidence for courtroom and punish the criminal by means of law is an effective method for preventing the e-mail crime. But it is difficult to find out the real identity of e-mail by the free mailbox. Because the applicant of free mailbox can forge the information at will, when they register. The sender's address can be forged and routed through anonymous mail server, or the sender's name may have been modified. So finding some efficient methods for analyzing the content of e-mail messages and identifying or categorizing the authors of these messages automatically are becoming imperative.

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 251–259, 2008.

Stylometry is a linguistic discipline that applies statistical analysis to literature by capturing the often elusive character of an author's style, using a variety of quantitative criteria. The main assumption underlying stylometric studies is that authors have an unconscious aspect to their style. Every author's style is thought to have certain features that are independent of the author's will, and since these features cannot be consciously manipulated by the author, they are considered to provide the most reliable data for the stylometric study. Stylometry is the basis for authorship analysis, which evaluates writing characteristics to make inferences about who wrote it. Authorship identification is one approach of authorship analysis, which deals with attributing authorship of unidentified writing on the basis of stylistic similarities between the authors' known works and the unidentified piece. From a machine learning point of view, this task can be seen as single-label multi-class machine categorization problem.

One major subtask of the authorship identification problem is extracting the most optimum features for representing the style of an author. Several measures have been proposed, including attempting to quantify lexical features, syntax features and structural features. The primary problems in the field are that there is no consensus of fixed features set. Techniques 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. Furthermore e-mail documents are one especial form of text documents. E-mail documents are generally brief and to the point. Similar to written letter, the writing should obey some form of formats, though some authors ignore them, which just can be represented as a sort of features. So in this paper, Features exaction methods that adapt to Chinese e-mail documents were presented to investigate the methods for identifying Chinese e-mail documents' authorship. The support vector machine was adopted as learning algorithm. Experiments were made to validate the feasible of the methods to computer forensic.

The rest of the paper is organized as follows: Section 2 introduces the related work involving authorship identification. Section 3 describes the methods of Chinese e-mail authorship identification. Section 4 provides our experimental methodology and analysis the experimental results. Section 5 is the conclusion of the paper.

2 Related Works

Stylometry is a burgeoning interdisciplinary research area that integrates literary stylistics, statistics and computer science. The origins of stylometry can be traced back to the mid 19th century, where the English logician Augustus de Morgan suggested word length could be an indicator of authorship. The real impact did not come until 1964, when two American statisticians Mosteller and Wallance decided to use word frequencies to investigate the mystery of the authorship of The Federalist Papers [1].

Stylometry has been used in a small but diverse number of application areas. Examples include identifying authors in literature, in program code, and in forensic

analysis for criminal cases. Perhaps the most extensive and comprehensive application of authorship analysis is in literature. Several studies attempts to resolve Shakespeare's works date back many years[2]. Specific author features such as unusual diction, frequency of certain words, choice of rhymes, and habits of hyphenation have been used as testing for authorship attribution. Program code authorship has been researched in the context of software and plagiarism, software author tracking and intrusion detection. Some features such as typographical characteristics, stylistic metric, and programming structure metrics have been researched [3][4][5]. The forensic analysis attempts to match text to authors for the purpose of a criminal investigation[6]. Currently forensic analysis has become increasingly popular in identification of online messages due to augmented misused of the Internet[7][8].

Recently, e-mail authorship analysis begins to draw researchers' attention. De Vel attempted to identify and attribute authorship of e-mail messages using support vector machine by using a collection of typical stylistic features such as structural characteristic and linguistic evidence for the purpose of computer forensic. Promising results were achieved, but the approach used was limited and far from optimized[9][10]. Tsuboi studied authorship attribution of e-mail messages and World Wide Web documents written in Japanese[11]. He used the sequential word patterns or word n-grams with n=2 and 3 from each sentence in the documents as features set. Good classification performance was gained. But Chinese is different from English, and a much lesser different from Japanese, therefore techniques developed in English and Japanese, especially the former, may not be directly applicable to Chinese.

Statistical and machine learning techniques constitute the two most common analytical approaches to authorship attribution. Many multivariate statistical approaches such as principal component analysis have shown a high level of accuracy. However, these approaches also have some pitfalls, including the need for more stringent models and assumptions. Machine learning techniques emerged from the drastic increases in computational power over the past several years. These techniques include support vector machine(SVM), neural networks, and decision trees. They have gained wider acceptance in authorship analysis studies in recent years because they provide greater scalability than statistical techniques for handling more features, and they're less susceptible to noisy data. These benefits are important for working with authorship identification, which involves classification of many authors and a large features set[7][10][11].

3 The Methods of Chinese E-Mail Authorship Identification

Figure 1 provides the process design of the methods of Chinese e-mail authorship identification. The process was divided into three steps. The first was feature extraction step. The purpose of the step was extraction writing style features from e-mail set and representing these features by vector space model. The second was training step. In this step, support vector machine algorithm was used to learn the features set and transformed into classifier. The third was categorizing step. The unknown classificatory e-mail documents could be categorized into some authors list automatically by the classifier that was trained in training step. The following are the methods described in detail.

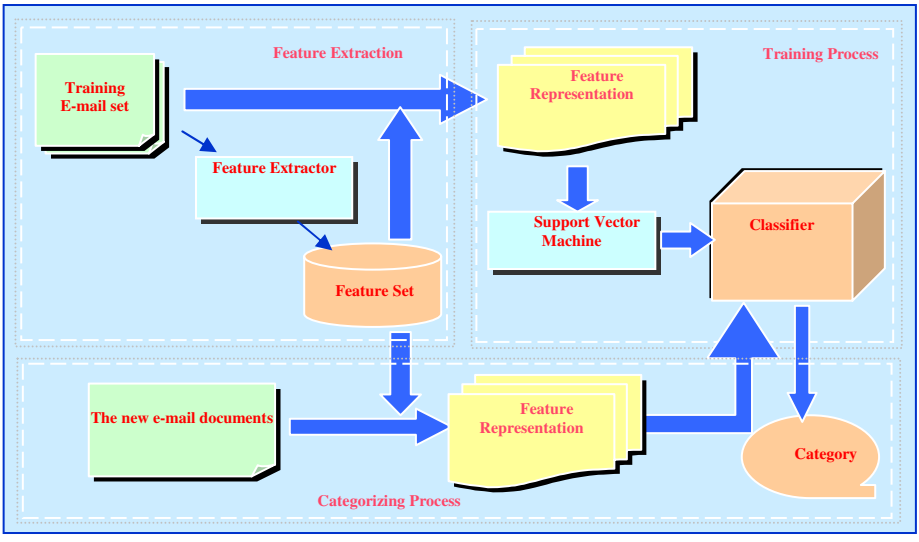


Fig. 1. Authorship identification process design

3.1 The Feature Extraction and Representation Methods

The optimum feature extraction methods contribute to the categorization accuracy. As to Chinese e-mail documents, linguistic features, structural characteristics and format features were extracted.

Linguistic features can be word based. That is to say, the features are that somebody has a preference for usage of some specific words unconsciously. The statistic of the frequencies of the words can be computed and represented as linguistic features, which is similar to lexical features in some studies.

Because Chinese text does not have a natural delimiter between words, the frequencies of words cannot be computed until Word segmentation methods are available in china, which is necessary to the research of Chinese text process. In the research, the word segmentation software named ICTCLAS developed by Chinese academy of Sciences was used for word segmentation and part of speech tagging of Chinese e-mail text.

We adopted Vector Space Model (VSM) to represent the linguistic features. The vector space model has been widely used in the traditional Information Retrieval field. The basic idea is to represent each document as a vector of certain weighted word frequencies. 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 $V_d = ((t_1, w_1), (t_2, w_2), \cdots, (t_n, w_n))$. We calculate the weight of the vector by the common technique $tf \bullet idf$ (term frequency-inverse document frequency) value:

$$W(t,\vec{d}) = \frac{tf(t,\vec{d}) \times \log(N/n_t + 0.01)}{\sqrt{\sum_{t \in \vec{d}} \left[tf(t,\vec{d}) \times \log(N/n_t + 0.01) \right]^2}} \tag{1}$$

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.

It is not necessary to extract all the words as the linguistic features, because some words contain little information. Too many dimensions may cost computation time, and may have negative effect on categorization results. So dimension reduction is essential. Document frequency (DF), information gain (IG), mutual information (MI), term strength (TS), the x^2-test (CHI) are the common feature selection methods. Yang has compared the various methods and found IG most effective[12]. So we adopted IG as the feature selection criteria.

$$\begin{aligned} Gain(w) = &-\sum_{i=1}^{m} p(c_i) \log P(c_i) + p(w) \sum_{i=1}^{m} p(c_i / w) \log P(c_i / w) + \\ &P(\overline{w}) \sum_{i=1}^{m} p(c_i / \overline{w}) \log P(c_i / \overline{w}) \end{aligned} \tag{2}$$

$Gain(w)$ denotes the information gain of term w. $\{c_i\}_{i=1}^{m}$ denotes the set of categories in the target space.

Structural characteristics deal with the text's organization and layout. The body of e-mail documents sometimes is short. There are a few sentences and paragraphs. The authors can write at will. Therefore, based on e-mail's writing characteristics, we extracted 32 structural characteristics. The examples are listed in table 1.

Table 1. E-mail documents' structural characteristics

Attribute type
Mean sentence length
Mean paragraph length
Number of blank lines/total number of lines
Number of space/total number of words
The rates of English words
The rates of digital
The rates of punctuation

Table 2. E-mail documents' format features

Attribute type
With or without have attachments
With or without have reply
With or without have date
With or without have appellation
Uses a greeting acknowledgement
Uses a farewell acknowledgement
Contain signature text block

Similar to common letters, e-mail documents have some writing formats, Such as greeting, attachment, signature, farewell etc. we extracted 21 format features. The examples are listed in table 2.

3.2 Support Vector Machine Classifier

Support Vector Machine (SVM) is a relatively new class of machine learning techniques first introduced by Vapnik[13]. Based on the structural risk minimization principle of the computational theory, SVM seeks a decision surface to separate the training data points into two classes and makes decisions based on the support vectors that are selected as the only effective elements from the training set.

The idea of the Support Vector Machine is to find a model for which we can guarantee the lowest true error by controlling the model complexity (VC-dimension). This avoids over-fitting, which is the main problem for other learning algorithm. So the distinctive advantage of the SVM is its ability to process many high-dimensional applications, such as text classification and authorship categorization. So in this study, we adopted SVM algorithm as learning algorithm.

4 Experiments

4.1 Experiments Methods

To validate the methods, experiments were made on limited dataset, which included 150 e-mail documents that were written by five persons. Each person had 30 e-mail documents. 20 e-mail documents were selected as training set. And other 10 e-mail documents were selected as testing set. Since there were only a small amount of data to produce a model of authorship, the performance of each feature was measured by 3-fold cross-validation to provide a more meaningful result. We wanted to do two experiments. The purpose of the first experiment was to validate the effect of different features set combination on results. 1000 linguistic features mainly including some words and phrases were extracted. Moreover 32 structural characteristics and 21 format features were extracted. To experiment how many dimensions of linguistic features were optimum, we have done the second experiment on different number of linguistic features.

To evaluate the experimental performance on the e-mail document corpus, macro-averaged F_1 statistic F_1 (M) was calculated, where:

$$F_1^{(M)} = \frac{\sum_{i=1}^{N_{AC}} F_{1,AC_i}}{N_{AC}} \tag{3}$$

Where N_{AC} is the number of author category and F_{1,AC_i} is the per-author-category F_1 statistic for author category $AC_i (i = 1,2,\cdots,N_{AC})$:

$$F_{1,AC_i} = \frac{2P_{AC_i}R_{AC_i}}{(P_{AC_i} + R_{AC_i})} \tag{4}$$

The classifier used in the experiments was the Support Vector Machine classifier Libsvm2.8.4. Libsvm is a simple, easy-to-use, and efficient software for SVM classification and regression. In our experiments we used the linear kernel function and the balance parameter value C was set to 1.0.

4.2 Experiments Results and Discussions

After extracting the feature values, we have done the first experiment to test the affect of the three types of features on the results respectively. The first set (F1) consisted of linguistic features, and the second (F2) denoted structural characteristics. (F3) was the format features. Table 3 and figure 2 summarize authorship identification accuracy results for the comparison of the different features set combination.

Table 3. The experimental results of different features set combination

Features set	F_1
F1	83.04%
F1+F2	92.88%
F1+F3	96.59%
F2+F3	97.59%
F1+F2+F3	98.36%

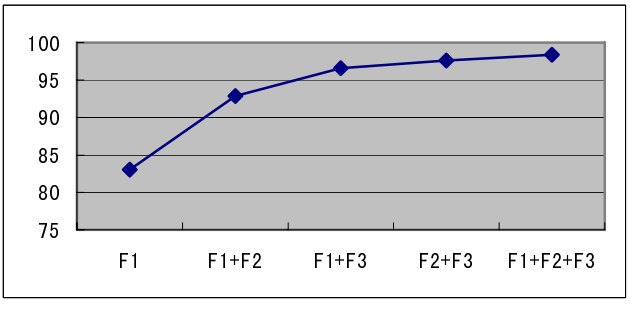


Fig. 2. The experimental results of different features set combination

From table 3 and figure 2, we could see that the best results were gained if we combined linguistic features, structural characteristics and format features together. The result of F_1 was 98.36%. This was a satisfying result. The worst results were gained if we experimented on linguistic features solely. The result of F_1 was 83.04%. We could get better results if we combined linguistic features and other two kinds of features.

To experiment how many dimensions of linguistic features were optimum, we have done the second experiment on different number of linguistic features. Table 4 and figure 3 show the affect of different number of linguistic features on results.

Table 4. The effect of different dimensions of linguistic features on experimental results

dimensions	100	300	400	500	600	700	1000
F_1(%)	64.2	86.9	87.1	87.5	85.7	86.6	83

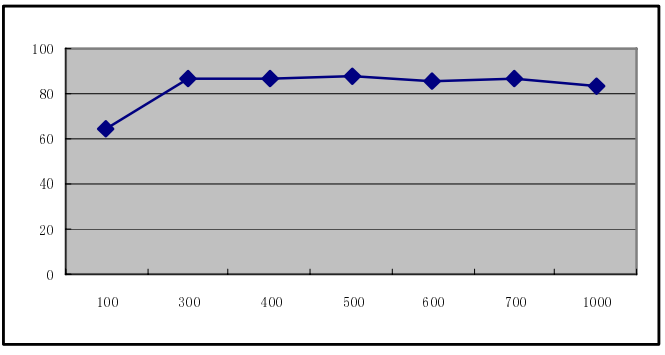


Fig. 3. The effect of different dimensions of linguistic features on experimental results

From Table 4 and figure 3, we could see that about 500 dimensions of linguistic features were optimum. Too few features could not express the author's features adequately. Too many features may result in features redundancy and have negative effect on categorization results. So the optimum features selection for linguistic features were essential.

5 Conclusion

In this paper, the computer forensic methods for identifying e-mail documents' authorship automatically were provided. Various features including linguistic features, structural characteristics and format features were analyzed. Support vector machine algorithm was adopted as learning algorithm. To validate the effects of different features on results, different features combination was experimented. The conclusions were made that the best results were gained if the linguistic features, structural characteristics and format features were combined together. The F_1 was 98.36%. Furthermore we have drawn the conclusion that about 500 dimensions of linguistic features were optimum. The results were satisfying, which proved that the methods were feasible to apply for computer forensic.

References

1. Mosteller, F., Wallace, D.L.: Inference and Disputed Authorship. Addison-Wesley Publishing Company, Inc., Reading (1964)
2. Elliot, W., Valenza, R.: Was the Earl of Oxford the true Shakespeare? Notes and Queries 38, 501–506 (1991)

3. Krsul, I.: Authorship analysis: Identifying the author of a program. Technical report, Department of Computer Science, Purdue University (1994)
4. Krsul, I., Spafford, E.: Authorship analysis: Identifying the author of a program. Computers and Security 16, 248–259 (1997)
5. 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)
6. Crain, C.: The Bard's fingerprints. Lingua Franca, pp. 29–39 (1998)
7. Abbasi, A., Chen, H.: Applying Authorship Analysis to Extremist-Group Web Forum Messages. IEEE Intelligent System 20(5), 67–75 (2005)
8. Zheng, R., Qin, Y., 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)
9. de Olivier, V.: Mining E-mail Authorship. In: KDD 2000 Workshop on Text Mining, ACM International conference on knowledge Discovery and Data Mining, Boston, MA, USA (2000)
10. de Olivier, V., 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)
11. Tsuboi, Y.: Authorship Identification for Heterogeneous Documents. Nara Institute of Science and Technology, University of Information Science, Japanese (2002)
12. Yang, Y.: An evaluation of statistical approaches to text categorization. Journal of Information Retrieval 1, 67–88 (1999)
13. Vapnik, V.: The Nature of Statistical Learning Theory. Wiley, New York (1998)

A Collaborative Forensics Framework for VoIP Services in Multi-network Environments

Hsien-Ming Hsu[1], Yeali S. Sun[1], and Meng Chang Chen[2]

[1] Dept. of Information Management, National Taiwan University, Taipei, Taiwan
{d94002,sunny}@im.ntu.edu.tw
[2] Institute of Information Science, Academia Sinica, Taipei, Taiwan
mcc@iis.sinica.edu.tw

Abstract. We propose a collaborative forensics framework to trace back callers of VoIP services in a multi-network environment. The paper is divided into two parts. The first part discusses the critical components of SIP-based telephony and determines the information needed for traceback in single and multiple Autonomous Systems (ASs). The second part proposes the framework and the entities of collaborative forensics. We also propose an algorithm for merging collected data. The mechanism used to execute collaborative forensics with co-operating units is presented and the procedures used in the collaborative architecture are described. For every entity, we suggest some interesting topics for research.

Keywords: collaborative forensics, VoIP services, traceback, SIP.

1 Introduction

The Public Switched Telephone Network (PSTN) has dominated voice communications over a long period. With the growth of the Internet, however, VoIP (Voice over IP) services based on packet-switched technology have become widely accepted and could eventually replace PSTN. Currently, a major drawback of VoIP services is that they are vulnerable to many potential security threats inherited from the Internet Protocol (IP). A taxonomy for mitigating potential VoIP security and privacy problems is defined in [1].

While VoIP services have many desirable communication features, they have also become a tool for illegal activities, as criminals can communicate via VoIP services and avoid being intercepted by law enforcement agencies (LEAs). There are a number of reasons why LEAs have difficulty intercepting and tracing back VoIP calls. Two major reasons are that 1) diverse techniques are used to access the Internet, e.g., campus networks, General Packet Radio Service (GPRS), Public 802.11 wireless network, and 3G; and 2) the dynamic addresses assigned to the caller/callee, are frequently located behind a Network Address Translation (NAT) router. Therefore, how to help LEAs identify IP packets lawfully is a major problem in various networks [2].

The goal of the VoIP traceback task is to trace the identities and geo-locations of the caller and callee of a VoIP service. To achieve this goal, Network Operators, Access Providers and Service Providers (NWO/AP/SvP) have to cooperate to record the identities of the parties and other necessary information. In this paper, we argue that

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 260–271, 2008.

the information needs to be recorded by operators of SIP- (Session Initiation Protocol) based networks. We propose a collaborative forensics framework, protocol and mechanism that automatically collects, associates, manages, and links information in order to reconstruct criminal acts. As a result, different parts of an event can be linked to build a complete picture of an incident that could be used as evidence in a court of law. By correlating related events, we can determine how a network incident (i.e., crime/attack) occurred, including the origin, the method used, and the people responsible. Ultimately, we hope to apply our findings to help prevent criminal activities on the Internet.

The proposed collaborative forensics framework is based on two assumptions: a) each NWO/AP/SvP has the administrative capability to handle event interception and to provide correct information to LEAs; and b) the collaborative forensic operating environment is secure.

The primary contributions of this paper are the follows:

- We discuss the protocol and the critical components of VoIP services and determine the information that needs to be recorded for possible forensic investigations.
- We explain how to perform traceback by using the recorded information in two scenarios, single and multiple AS networks.
- We propose a cooperative architecture, protocol and mechanism for collaborative forensics. In addition, we propose an algorithm for merging the collected data.
- We suggest several interesting research avenues related to the development of the collaborative framework.

The remainder of this paper is organized as follows. Section 2 contains a review of related works. In Section 3, we describe SIP-based VoIP services and traceback. In Section 4, we discuss the proposed collaborative forensics framework. Then, in Section 5, we summarize our conclusions and indicate future research avenues.

2 Related Work

Although VoIP provides many desirable services, such as convenient voice calls, the services are vulnerable to a number of potential security threats inherited from the root Internet Protocol (IP). In recent years, VoIP has become a tool for illegal activities as criminals have exploited the security loopholes in IP. In [1], the authors investigate the risks of VoIP technology and define a taxonomy to enhance VoIP security and mitigate threats to privacy. Because VoIP services rely on the Internet, they are vulnerable to threats from different protocol layers. In [3], attacks are categorized by the vulnerabilities of VoIP devices, configurations, infrastructures, protocols and applications. Meanwhile, some works have focused on developing a VoIP intrusion detection system [4, 5]. There has also been a substantial amount of research on how to establish an LEA architecture for VoIP services [6, 7], and how to enhance VoIP for use by emergency services [8]. Newly-developed anonymous VoIP telephone services (e.g., Skype [9]) make the traceback task even more difficult for LEAs. To resolve this problem, Wang et al. [10] proposed a method that effectively traces anonymous calls by embedding a unique watermark on the inter-packet timing of the VoIP flow in real-time.

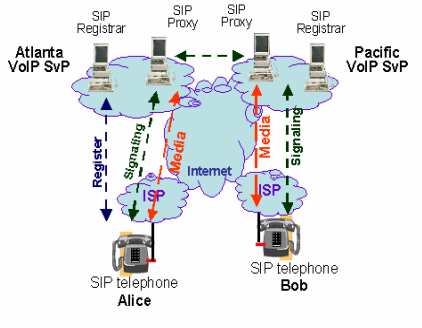


Fig. 1. SIP-based IP telephony

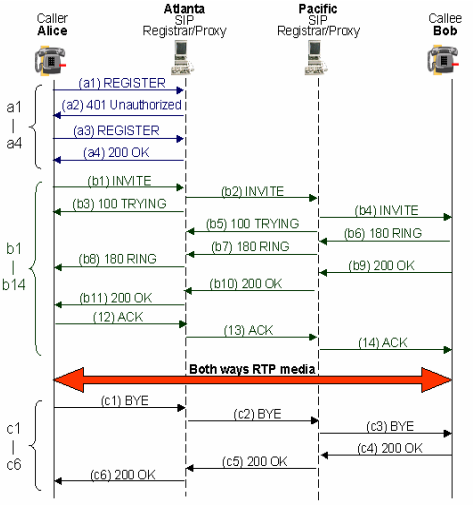


Fig. 2. SIP Signaling

In this research, we follow the approaches in [11, 12, 13] and propose a framework for collaborative forensics. We also propose sending XML (eXtensible Markup Language) [14] formatted messages via web services. Our objective here is twofold: 1) to help LEAs obtain the information necessary to trace back VoIP phone calls; and 2) to help domain experts construct domain knowledge to enhance existing systems collaboratively.

3 VoIP Traceback

In this paper, we only consider SIP-based IP telephony. To provide VoIP services, an SIP-based telephone system utilizes multiple protocols, including the Session Initiation Protocol (SIP) [15] and the Real-time Transport Protocol (RTP) [16]. As mentioned earlier, there are a number of reasons why LEAs have difficulty intercepting and tracing back VoIP calls. Two major reasons are that 1) diverse techniques are used to access the Internet (e.g., campus networks, General Packet Radio Service (GPRS), Public 802.11 wireless network, and 3G); and 2) dynamic IP addresses assigned to a caller/callee are frequently located behind the Network Address Translation (NAT) router. In this section, we describe the communications and the critical points for VoIP services, and determine the information that needs to be recorded. Then, based on the recorded information, we discuss how to perform traceback with single and multiple ASs.

3.1 The Communications and Critical Points of VoIP Services

The architecture of SIP-based IP telephony is shown in Fig. 1. The Registrars and Proxies are the SIP servers. A Registrar is responsible for registration, after which the Proxy servers relay the signaling to the callee's address and offer the service. The

Table 1. Information recorded by the SIP Registrar Server

Attributes	Description
Account	User's network-based phone account
Source IP	Obtained from the user's registered message
Timestamp	The time the call was registered

Table 2. Information recorded by the SIP Proxy server

Attributes	Description
Caller's Account	Caller's account or telephone number; obtained from the caller's INVITE message.
Callee's Account	Callee's account or telephone number; obtained from the caller's INVITE message.
Caller's IP/Port (Signaling)	Caller's IP and Port number; obtained from the caller's INVITE message.
Callee's IP/Port (Signaling)	Callee's IP and Port number; obtained from the callee's OK message.
Caller's IP/Port (media)	Caller's IP and Port number; obtained from the caller's SDP on INVITE message.
Callee's IP/Port (media)	Callee's IP and Port number; obtained from the OK message of the callee's SDP
Time: From	The time Proxy received the INVITE
Time: To	The time Proxy received the BYE
Answering time	The time Proxy received the OK

Table 3. The NAT/DHCP

Attributes	Description
Account	User's network-based phone account
Private IP/Port	The private IP/Port with NAT
Public IP/Port	The public IP/Port assigned to NAT/DHCH
Time: From	The time the private IP made the call
Time: To	The time the private IP was interrupted

signaling of the SIP protocol is shown in Fig. 2 [15]. Tables 1, 2, and 3 list the respective information that the SIP Registrar Server, SIP Proxy Server and NAT/DHCP (Dynamic Host Configuration Protocol) need to record for traceback.

3.2 Traceback within a Single AS Network

We use a scenario of a caller (Alice) and a callee (Bob) in a single AS network to explain how we perform traceback from Bob to Alice, as shown in Fig. 3.

Case 1: Both Alice and Bob are with public IPs in a single AS. In this scenario, all the required information about the caller and callee is recorded by the Registrar and Proxy service providers when the connection for the session is set up. The information, which is distributed over a number of components (as shown in Fig. 3), can be easily collected,

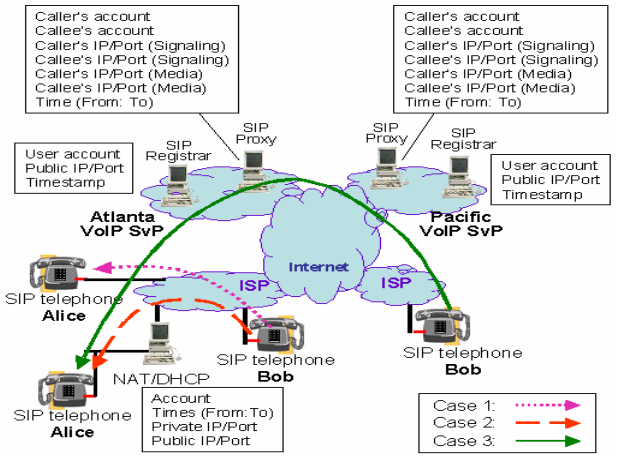


Fig. 3. The necessary information recorded by the components of SIP-based telephony to trace back calls in Single- and Multi-operators

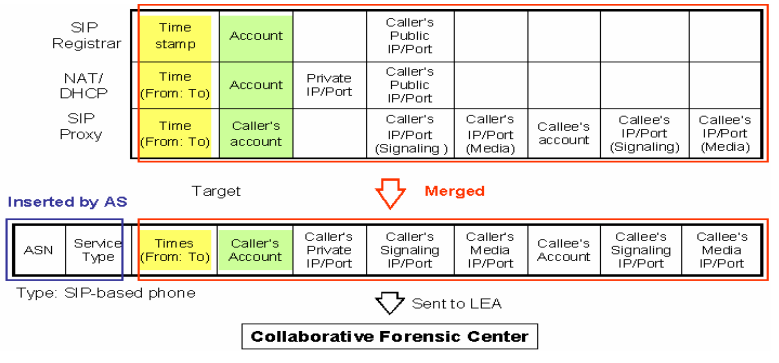


Fig. 4. The data recorded by the NAT/SIP Registrar/SIP Proxy is merged as a Local Event

extracted and merged into a Local Event (LE). The LE can be represented as an XML formatted message and reported by the administrator using the SEAL Protocol (introduced in the next section). Before the result can be sent to the LEA, the autonomous system number (ASN) and service type are inserted for classification and storage purposes, as shown in Fig. 4.

Case 2: Alice is with a private IP and Bob is with public IP in a single AS.

Because Alice is behind the NAT with the private IP, we need the NAT router to record the mappings of private IPs and public IPs during the connection period. The information gathered by the NAT router, SIP Registrar and SIP Proxy is collected, extracted and merged as a Local Event. This is same as case 1, except for the private IP, as shown in Fig. 4. Then, based on the reported events, we can perform a trace back to the caller across the NAT from the callee. For collaborative forensics, we have to decide how long information about Local Events should be kept. A tradeoff between storage requirements and the need for accuracy (mainly, the false negative rate) has to be made carefully.

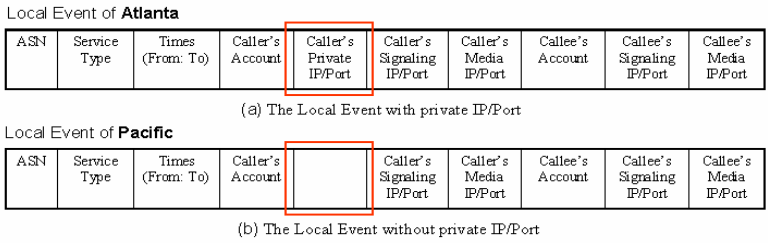


Fig. 5. The Local Events Protocol with/without private IP/Port

3.3 Traceback with Multi-AS Networks

Case 3: Traceback under multi-AS is similar that for the single AS cases described in the previous section, except that Alice and Bob's IPs are located at different service providers. In Fig. 3, the caller (Alice) and the callee (Bob) belong to different VoIP service providers, each of which has its own SIP register and SIP proxy servers. Assume that caller Alice is either behind the NAT router with a private IP or she uses the dynamic IP. After Alice completes the registration, the signaling will be relayed from the Atlanta SIP proxy to the Pacific SIP proxy, which will then relay it to Bob. Each Atlanta and Pacific SIP proxy can obtain the complete information within its AS and produce its Local Event independently, as shown in Fig. 5.

When tracing back from the callee Bob, we can get the caller's public IP from the Local Event of Pacific, but not the private IP. However, if we can match the Local Event of Pacific with the Local Event of Atlanta, we can obtain the caller's private IP, as shown in Fig. 5 (b).

4 The Framework of Collaborative Forensics for VoIP Services

Operators may not want to share their information with others for a variety of reasons (e.g., privacy concerns, commercial competition, policy, cultures, and implementation differences). One way to solve the problem is to design a mechanism that can be supervised by an independent authority. The mechanism would aggregate, integrate and correlate local information (i.e., Local Events) from operators to carry out the traceback task without violating privacy laws. Since the LE only contains information needed for traceback and the collaborative framework is under the supervision of an independent authority, network operators should not be reluctant to collaborate.

Network operators already have systems and databases for the distribution of information needed for traceback. Therefore, we only need a cooperative architecture, protocol and mechanism to automatically collect, associate, manage, link and reconstruct information about criminal activity in real-time for a fast response. The proposed collaborative framework, called SKYEYE, is designed to meet this need.

4.1 SKYEYE Entities and Their Functions

In this section, we introduce the entities of SKYEYE and their functions. Fig. 6 illustrates the SKYEYE entities and procedures, as well as the cooperating units, i.e., LEA, existing systems, NWO/AP/SvP and FRTs (Fast Response Teams).

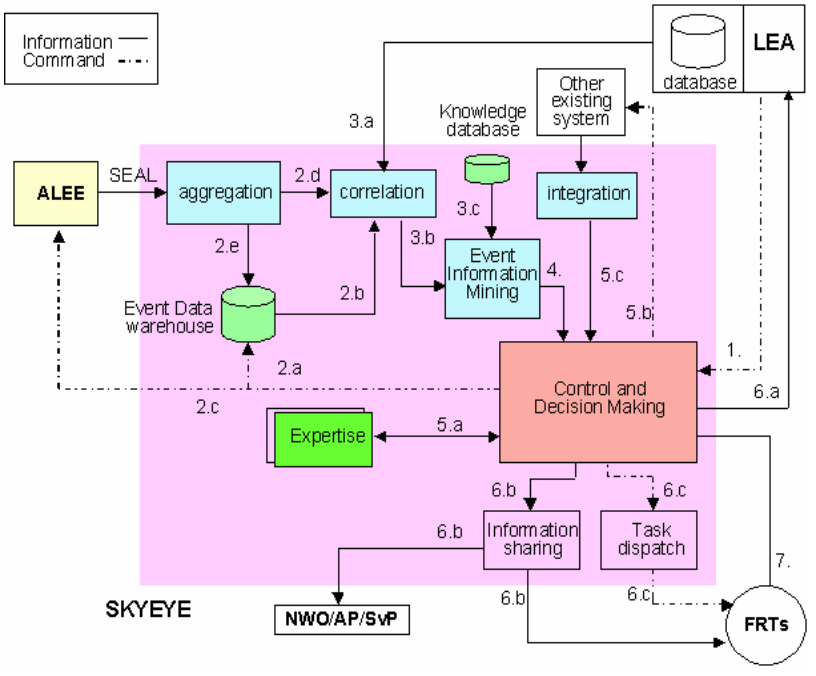


Fig. 6. The architecture and procedures of SKYEYE

4.1.1 Local Event (LE)

Local Event data is collected from the NAT/DHCP, SIP Registrar and SIP Proxy in an AS by ALEE using the specification of SkyEye-ALee (SEAL) Protocol. For one VoIP call, the ALEE will produce two copies of the Local Event; one will be stored in the local operator's database for backup, and the other will be sent to SKYEYE for forensic analysis. The service type of the Local Event (e.g., SIP, GSM, 3G) will be labeled accordingly by ALEE. No matter whether the caller and callee are located in the same AS or not, their IPs are encapsulated in the LE, as shown in Figs. 4 and 5. For traceback to the caller, we only need the LE's from the callee and caller. In other words, the LEs of the intermediate ASs are not needed.

4.1.2 Access Local Event Entity (ALEE)

ALEE is the interface of AS that connects and communicates with SKYEYE, which is independent of the AS realm. For each SIP-based phone call, the ALEE will automatically collect the required information about the caller and callee from the NAT/DHCP, SIP Registrar and SIP Proxy in the operating network (AS) and merge the pieces of information to produce the Local Event. When ALEE receives the command from SKYEYE, the queried LEs will be sent to SKYEYE again for confirmation.

4.1.3 The Flexible SKYEYE ALEE (SEAL) Protocol

The SEAL protocol is designed to transport Local Events presented in XML-formatted messages; therefore, it can easily be extended to accommodate different access network technologies and different services.

4.1.4 The SKYEYE

The SKYEYE is the kernel of the collaborative forensic mechanism. All the collaborative investigating activities are executed in this entity. It is comprised of the following modules and functions: aggregation, event correlation, event information mining, integration, and expertise repository.

Below, we describe each module and its function in the collaborative forensics framework. We also indicate issues that require further research.

- *The Aggregation Module (AGG)*: The AGG is the interface of ALEE that collects all LEs from ALEE and stores them in the event data warehouse (EDW). To improve the querying and search functions, further research is required on: a) LE classification and storage; and b) efficient aggregation methods.
- *The Correlation Module (CORR)*: The CORR's function is to correlate related LEs in order to build a complete picture and determine *how* a network incident (crime, attack) occurred. CORR tries to find out the origin, the method, the people responsible, and the identity of potential victims. Further research is required on a) an efficient algorithm for correlating related events; and b) how to identify the key attributes of a crime for correlative forensics.
- *The Event Information Mining Module (EIM)*: By exploiting data mining techniques, the EIM tries to discover useful knowledge from LEs in order to predict criminals' intentions and thereby prevent crimes occurring. Further research is required on: a) techniques for Event information mining; and b) a model of criminal behavior that can be used to predict possible criminal activity.
- *The Integration Module (INTE)*: The INTE is an interface for integrating information from existing information systems for forensic analysis.
- *Control and Decision Making Module (CDM)*: This is the core module of SKYEYE. Its main functions are to control the processes of collaborative forensics and make necessary decisions for other tasks. The CDM can look up an expertise repository or consult domain experts via virtual panels for decisions. The CDM issues orders for Fast Response Teams (FRTs) to execute tasks, and provides updates about the latest situations and information about incidents. Further research is required to a) develop standard operating procedures (SOPs) for the CDM; b) establish a decision-making procedure; and c) improve methods the used to identify potential victims and criminal companies.
- *Fast Response Teams (FRTs)*: FRTs are specialized units that perform diverse tasks. They receive and execute orders sent by the CDM. They also report the latest situations and incidents to the CDM.
- *Information Sharing*: Event information is shared with partners for event detection and prevention, defense in depth and fast responses to events, and to alert potential victims.

4.2 Collaborative Forensics Work of SKYEYE

Next, we describe the execution of collaborative forensics with cooperating units. The steps are as follows.

Step (1). LEA sends commands to the Control and Decision-Making module (CDM) of SKYEYE. Each command includes two essential elements, the Callee's

Account and the Calling time-parameter. The former is the starting point for the traceback, and the latter is used to identify the right call. Both of them are key attributes of LE searches of the Event Data Warehouse (EDW) and the operator's database.

Step (2.a). The CDM module sends a request to the EDW and ALEE for a Local Event (LE) search with Callee's Account and Calling time-parameter. If an LE exists it should be stored in the EDW, because it would have been produced and sent to the EDW by ALEE when a call was terminated. The only reasons for the non-existence of an LE are that it was deleted when the stored data expired or some transmission errors occurred.

Step (2.b). The responses of EDW depend on the number of LEs found in its database. For example, given a Calling-parameter, $19^{h}30^{m}$:$19^{h}52^{m}$, for each call, there should be two LEs, produced by the ASs of the caller and the callee respectively and stored in the EDW. The LEs will be accessed and sent to correlation module (CORR) to be double-checked with the LE information from the operator. For these LEs, the Caller's Related Information (e.g., Account, Public IP/Port) and ASN can be obtained and passed to the LEA (Step 6.a). For SIP-based phone calls, the data collection task has been completed in principle.

Step (2.c). No matter whether the LEs can be found in the EDW or not, the CDM module sends the request to the operator for Local Events (LEs) for confirmation. The LEs will be sent to Aggregation module (AGG) via SEAL. First, the AGG will check whether the LEs are still stored in the EDW. If they are, they will be sent to the CORR module (Step 2.d) for double-checking and correlating; otherwise, they will be stored in the EDW first (Step 2.e), and then forwarded to the CORRR.

Step (3.a). The CORR module double-checks the LEs sent by the operators and the EDW, and then correlates them with the data in the LEA database. For the forensic investigation, it is necessary to confirm the true identity of the caller by his/her Account during the Calling time. Based on the key attributes of Local Events, the CORR tries to fit the parts of events together correctly and build a *complete* picture of the incident, which can be used as evidence in a court of law.

Step (3.b). The related Local Events are sent to the Event Information Mining module (EIM).

Step (3.c). Based on the related Local Events, the EIM will consult the forensic domain knowledge in the Knowledge DataBase (KDB) to guide the search or evaluate the behavior models or predict criminal activity.

Step (4). The EIM mines the LEs for further useful information that could be used to predict criminal activity. All of the predicted results are sent to CDM module.

Step (5.a). The CDM may need to consult the domain experts via the virtual panel.

Step (5.b). The CDM tries to determine if any information is missing or confusing, and requests data or evidence from other systems.

Step (5.c). The other systems send their responses to the Integration module (INTE), which can integrate information in different formats. The result is sent to CDM to support decisions about subsequent action.

Step (6.a). All the decisions will be passed to the LEA. Step (6.b). The information is shared with the cooperating units to alert them in order to prevent possible criminal activity, and with FRTs (Fast Response Teams) to support their tasks. Step (6.c).

Step (7). Any changes and updates should be sent back to the CDM module.

4.3 The Algorithm for Data Merging

The data merging algorithm listed in Fig. 7 is used for VoIP services only. For the other types of service, the corresponding algorithms need to be defined according to their individual needs and the information needed for traceback and forensics. When an SIP-based phone call is terminated, all information recorded by the SIP Proxy, SIP Registrar and NAT router is collected by the AS administrator and sent to ALEE. Then, ALEE processes the data merged by the algorithm in Fig. 7. The output is the Local Event presented as an XML-formatted SEAL message.

```
Algorithm for Data_Merging
Input: SIP_Proxy, SIP_Registrar, NAT_router
Output: Local_Event

begin
   NAT_R:=NAT Record;
   SP_t:=SIP_Proxy(Time(From:To));
   SR_t:=SIP_Registrar(Timestamp);
   NAT_t:=NAT(Time(From:To));
   CRA_SP:=Caller's Account of SIP Proxy;
   CRA_SR:=Caller's Account of SIP Registrar;
   CRA_NAT:=Caller's Account on NAT router;
   CRA_U:=Caller's Account of User;
   CEA_SP:=Callee's Account of SIP Proxy;
   Pu_IP/Pt:=Public IP and Port;
   Pt_IP/Pt:=Caller's Private IP and Port;

   create a Local_Event
   LE.ASN:= ASN of operator
   LE.Service_Type:= Service_Type of calling;
   if (NAT_R is not empty)
      do (LE.time(From:To):= NAT_t;
          && LE.Caller_Account:= CRA_NAT;
          && LE.Private_IP_Port:= Pt_IP/Pt);
      else
      do (LE.time(From:To):= SP_t;
          && LE.Caller_Account:= CRA_SP);

   do (LE.caller_Public_IP_Port(signaling):= Pu_IP/Port(signaling);
       && LE.Caller_Public_IP_Port(Media):= Pu_IP/Port(Media));
   do (LE.Callee_Account:= CEA_SP;
       && LE.Callee_Public_IP_Port(Signaling):= CEA_SP(signaling);
       && LE.Callee_Public_IP_Port(Media):= Pu_IP/Port(Media));
end
```

Fig. 7. The Algorithm for Data Merging

5 Conclusion and Future Works

In this paper, we have discussed the critical components of VoIP services, and defined the information that needs to be recorded for forensic investigations. Based on the recorded information, we explain how to perform traceback in single and multiple AS networks. We propose an architecture, protocol and mechanism for collaborative

forensic tasks. In addition, we describe the entities of the architecture and their functions, and define the SEAL protocol with XML formatted messages.

The proposed SKYEYE model is the kernel of the collaborative forensic mechanism. The aggregation, event correlation, event information mining, integration and expertise modules are still under development.

Our ultimate goal is to establish a collaborative forensic center that can automatically collect, associate, manage, link and reconstruct information about possible criminal activities, as well as share the real-time information from different autonomous systems with all cooperating units.

Acknowledgements. This work is partly supported by the National Science Council of Taiwan under Grant No: NSC 96-3114-P-001-002-Y. The study would not have been completed without the help of the ANTS lab members, especially Liang-Ming Wu and Hao-Wen Ke.

References

1. Endler, D., Ghosal, D., Jafari, R., Karlcut, A., Kolenko, M., Nguyen, N., Walkoe, W., Zar, J.: VoIP Security and Privacy Threat Taxonomy, Public Release 1.0 (2005)
2. ETSI TR 101 944: Telecommunications security; Lawful interception (LI); Issues on IP Interception (2001)
3. Dhamankar, R.: Intrusion Prevention: The Future of VoIP Security. White paper. Tipping Point (2005), http://www.tippingpoint.com/pdf/resources/whitepapers/503160-001_TheFutureofVoIPSecurity.pdf
4. Sengar, H., Wijesekera, D., Wang, H., Jajodia, S.: VoIP Intrusion Detection Through Interacting Protocol State Machines. In: IEEE Dependable Systems and Networks Conference (2006)
5. Wu, Y., Bagchi, S., Garg, S., Singh, N., Tsai, T.: SCIDIVE: A Stateful and Cross Protocol Intrusion Detection Architecture for Voice-over-IP Environments. In: IEEE Dependable Systems and Networks Conference (2004)
6. Milaovic, A., Srbljic, S., Razjkevic, I., Sladden, D., Skrobr, D., Matosevic, I.: Distributed System for Lawful Interception in VoIP Networks. In: EUROCON (2003)
7. Karpagavinayagam, B., State, R., Festor, O.: Monitoring Architecture for Lawful Interception in VoIP Networks. In: Second International Conference on Internet Monitoring and Protection (2007)
8. Mintz-Habib, M., Rawat, A., Schulzrinne, H., Wu, X.: A VoIP Emergency Services Architecture and Prototype. Computer Communications and Networks (2005)
9. Skype-the Global Internet Telephony Company
10. Wang, X., Chen, S., Jajodia, S.: Tracking Anonymous Peer-to-Peer VoIP Call on the Internet. In: Proceedings of the 12th ACM Conference on Computer and Communications Security (2005)
11. Goodell, G., Aiello, W., Griffin, T., Ioannidis, J., McDaniel, P., Rubin, A.: Working Around BGP: An Incremental Approach to Improving Security and Accuracy of Interdomain Routing. In: The 10th Annual Network and Distributed System Security Symposium (2003)
12. Dawson, M., Winterbottom, J., Thomson, M.: IP Location- IP Location in Wireline Public Carrier Networks. McGraw-Hill Companies, New York (2007)

13. Nena, J.: Homeland Security Techniques and Technologies. Charles River Media, INC. (2004)
14. Bray, T., Paoli, J., Sperberg-McQueen, C., Maler, E.: Extensible Markup Language (XML) 1.0., 2nd edn. W3C Working Draft (2000)
15. Rosenberg, J., Schulzrinne, H., Camarillo, G., Johnston, A., Peterson, J., Sparks, R., Handley, M., Schooler, E.: SIP: Session Initiation Protocol (SIP). RFC 3261, IETF Network Working Group (2002)
16. Schulzrinne, H., Casner, S., Frederick, R., Jacobson, V.: RTP: A Transport Protocol for Real-time Applications. RFC 3550, IETF Network Working Group (2003), http://www.ietf.org/rfc/rfc3550.txt?number=3550

Wireless Forensic: A New Radio Frequency Based Locating System

Emmanuel Velasco[1], Weifeng Chen[2], Ping Ji[1], and Raymond Hsieh[3]

[1] Department of Math and Computer Science
John Jay College of Criminal Justice
City University of New York
New York, New York 10019
emmanuelvelasco@gmail.com, pji@jjay.cuny.edu
[2] Department of Math and Computer Science
[3]Department of Justice, Law and Society
California University of Pennsylvania
California, PA 15419
{chen,hsieh}@cup.edu

Abstract. Wireless networks are more prevalent today since they enhance the user flexibility and mobility when browsing online. However, those advantages come at a price. Wireless networks do not provide the same security features as its wired counterpart. An unprotected wireless network can be easily served as a stepping-stone in attacks through such as Warchalking. Due to the nature of wireless networks, it is very difficult to trace back to the intruder. The three current location tracking solutions: Closest Access Point, Triangulation, and Radio Frequency Fingerprinting have their own limitations in tracing wireless intruders. In this paper, we propose a more precise location tracking system - Multiple Grids System - that overcomes the limitations of the existing solutions. This new approach uses a modified wireless grid to track wireless hackers in long distance or short range at different floors of the same building. Our proposed multiple grids system can be effectively applied to detect, track and prevent wireless intruders. We also substantially present that our system can be collaborated with multiple locating applications.

1 Introduction

Today, many of us use wireless access to the Internet to communicate with others. Wireless access points or hot spots are now widely available in airports, hotels, libraries, schools and other public buildings. Compared to wired networks, wireless networks provide several advantages, enhancing the user flexibility and mobility when browsing online. However, this convenience comes at a price. Wireless networks do not provide the same security features as its wired counterpart. Anyone with the right radio frequency may eavesdrop on a wireless connection, or connect to the insecure network and use it for malicious reasons. War-chalking [2] is an example. Many wireless users do not have a technical background and their wireless devices are not configured correctly. Criminals then use these unprotected networks anonymously as a stepping-stone in their attacks. According to Urbad and Krone [3],

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 272–277, 2008.

urged security risks associated with wireless networks include Intrusion, Leeching and Exploitation.

Due to the nature of wireless networks, it is very difficult to trace back to the intruder. When one uses a wireless network, his IP address is the same as that of the network. When the hacker launches an attack, it will appear that the attack came from the network's owner instead of the hacker. Network logs on many routers are usually turned off by default. Even if they were turned on, the MAC address or any other identifiers that may determine the identity of the hacker can be spoofed. When an attack is over, the hacker can just walk away from the scene undetected.

The three current solutions for location tracking are Closest Access Point, Triangulation, and Radio Frequency (RF) fingerprinting [1]. However, these solutions have their limitations in tracing wireless hackers [4]. In this paper, we propose a more precise tracing system that overcomes the limitations of the existing solutions. This new approach uses a modified wireless grid to trace wireless hackers with long distances or at different floors. The proposed solution can be used to effectively detect, track, and prevent wireless intruders. We also show that this improved system can be used in multiple applications besides locating hackers.

The rest of the paper is organized as follows. Section 2 describes our new radio frequency based locating system. We present in Section 3 several promising applications based the new approach. Section 4 concludes the paper and provides our future work.

2 A Multiple Grids System

In [4], we studied the three current solutions for location tracking: Closest Access Point, Triangulation, and Radio Frequency (RF) fingerprinting. We found that these solutions attempt to locate wireless users within a building or controlled environment. Whereas, hackers usually try to connect to the network from outside the area. At further distances, the accuracy diminishes due to two reasons. (1) First, the interference and other effects on the signal strengths are unknown and will vary among different buildings; (2) Second, the wireless access points (APs) may not be within range to read the signal. Another problem for the three existing solutions is that the tests were done on the same floor. However, in a high-rise building, the hackers could be on another floor. RF fingerprinting would give the wrong results because the database would not have any empirical data from that position. It may not even be possible to obtain a reading from the floors above or below an office, since it may be private property and access is not allowed. Other solutions require customized equipments, which may not be possible to obtain.

Traditional APs would not work well with long distances because they cannot accurately distinguish the received signals. A signal received from a wireless card at a close range may have the same signal strength as the one received from a directional antenna at a further distance. The current locating systems would interpret the two signals coming from the same position.

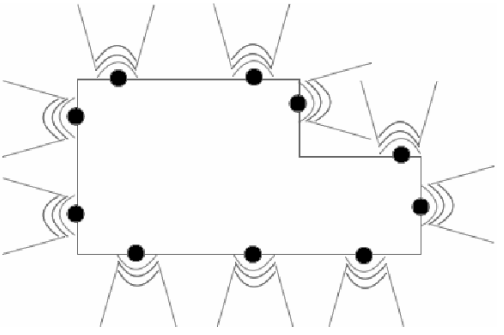


Fig. 1. Directional antennas throughout the external areas of the office

To address these limitations, we propose in this paper a new locating system to trace wireless hackers more precisely. Our improved system will use a modified wireless grid to locate the hackers. In order to detect hackers that are using directional antennas from a further distance, we must use directional antennas of our own. The new system includes a set of wireless receivers connected to directional antennas at the external points of the building (Figure 1). These devices are used strictly for receiving signals and not for any transmission. When our system receives a signal, it is able to tell whether the signal came from outside the building or inside. If the signal came from the inside, the internal APs would receive a reading, but the external receivers will receive a weaker signal or none at all because the antenna is pointed outside the building. However, if a hacker is accessing the network from outside the building, our external receivers will have a strong signal, as well as the internal APs, which receive a weaker signal. This allows us to conclude that somebody outside the building is trying to access our network.

We still have a problem with hackers from different floors. If the hacker is above or below us, it will appear as if the signal is coming from inside the building. Our current improved system should still be effective because the signal must go through floors, resulting in weak signal strength. The external receivers will also have weak signal strengths because the signals did not come from outside. Therefore, we can

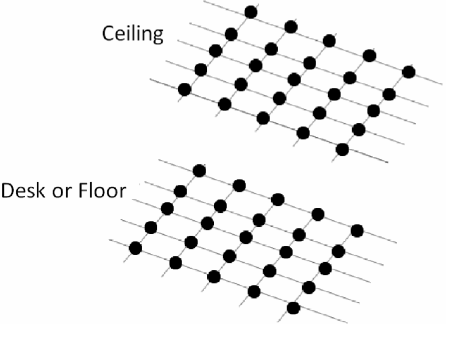


Fig. 2. Two wireless grids. One installed at the desk or floor level, one in the ceiling.

assume that the weak signals received by both the APs and the external receivers mean that user is accessing the network from different floors. However, we can increase the accuracy by using two wireless grids. One grid will be installed in the ceiling and the other grid will be installed on the floor or desk level, depending on the space that is available (Figure 2). If APs in the two grids do not match in a straight line, we can use angular information and the laws of sine and cosine to determine the device. To obtain the best accuracy, APs in the two grids should match in a straight line, as shown in Figure 2. In this case, the distances between the upper grid and lower grid are known, as well as the distances between each AP in the same grid.

The two grids are used to communicate with each other to determine the signal strength between them. If the hackers are in the floor above us, we can expect the ceiling grid to have a stronger signal than the lower grid, and vice versa if they are in the floor below us. People have used two or more access points to triangulate positions on a plain, i.e., the two-dimension (x, y) position. Because the locations of the access points are known, we can use their signal strengths as distances. A circle is drawn around each AP, with the signal strengths as the radius (Figure 3). The stronger the signal strength, the closer the distance is between the device and the wireless AP. The intersections of the circles are the possible locations of the device.

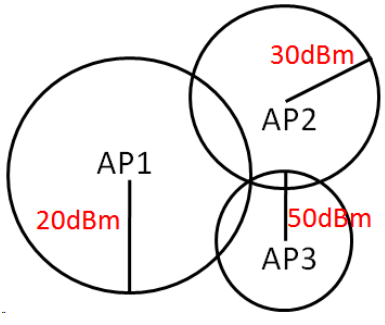


Fig. 3. Two wireless grids. One installed at the desk or floor level, one in the ceiling.

We now have two wireless grids, an upper and a lower, which allow us to calculate the third dimension. Instead of only locating the (x, y) coordinates of the user, we can now estimate the z-coordinate. In this case, each wireless access point would produce a sphere based on the detected strength and distance. Multiple spheres would then intersect to produce an estimation of the three-dimension location.

The last problem that the three existing solutions had is the following. When new objects were placed in the area that interfered with the radio frequency signals, a person was required to walk around the area to take the new signal readings and update the database in order to reconfigure these systems. Our improved system gets around this problem by automatically detecting the anomalies in the signals. Because the upper and lower grids have their own RF fingerprint, if they periodically test the signal strengths between each other, they can detect an anomaly. If it is found that the RF signal strength is different from its history, an alert will be sent to the network administrator. The administrator will then check the area to see if any new devices, furniture, or objects have caused the interference. If the cause for the anomaly is

found and accepted, the administrator simply updates the database with the new received signal strength indication values.

3 Applications of Improved Locating System

We have described our new locating approach based on multiple wireless grids (external and internal, upper grid and lower grid). The new solution can be used in various applications besides being a tool to locate and detect wireless hackers. First, it can be used to measure network coverage. We can find places in the building where the wireless coverage is weak, if we survey the area using the system.

The system can also be used to track employees in the area. Research has been conducted using infrared transmitters to track people in a building [5]. Note that, comparing to infrared technique, radio frequency technique is more reliable. Infrared transmitters require a direct line of sight with the receiver. It also has range limitations. Infrared can also be blocked by objects, or affected by the weather. Thus the new system is a better choice since it is based on RF fingerprinting. It can also be used for physical security. More specifically, if we can track employees in the area, we can track authorized users versus those that are not. Radio frequency identification (RFID) devices can be embedded in employee ID cards, which would transmit their locations to the receivers. Security features such as fingerprint scanners, iris scanners, and badges currently prevent users from entering a secured location. However, sometimes employees open a secured area for other coworkers to enter. The RFID would transmit their locations and alert security if an unauthorized user enters a secured area.

Besides employees, the system can also track assets. For example, most organizations are currently using the barcode system in their evidence room. Each piece of evidence has a barcode, which is then linked to a database stating where it is located. When somebody needs to examine the evidence, the evidence is scanned, taken out of the room, and then returned when done. However, we are uncertain of its exact location after it has been taken out of the evidence room. This problem can be solved using our new system. The advantage of our approach over a barcode system is that we can track evidence locations both inside and outside the evidence room. The improved locating system described in this paper can ensure that the evidence only goes into approved rooms. If the evidence is about to go into an area that does not have optimal conditions and may contaminate the evidence, an alert is sent out.

In a wireless network, if we want to limit the wireless users to only those within a particular area, we can have the network designed to ignore all requests from outside the designated area. For example, a university John Jay College can use this to limit wireless connections to only those inside the college buildings. Note however, while there are other security measures such as encryption, MAC address filtering, and user authentication, this system was not meant to replace them, but rather complement them to create a safer wireless network. Rogue networks are wireless APs placed inside a company's network, creating a security hole for hackers to use. Our new system can be used to detect these rogue networks. The locations of authorized devices are known. Therefore, when an unauthorized AP is transmitting, the system can detect the undetermined device and alert the network administrator. The

administrator can then block the rogue network by removing the wired connection that the unauthorized AP is connected to, or physically going to the location of the AP and removing the device.

A locating system can be very important for law enforcement also. Law enforcement currently has a hard time in locating hackers and charging them with a crime. Under this new system, their chances of finding the hacker increases. Being able to place an individual at a certain location and time can help law enforcement solve a case. A security camera may have captured the hacker in the act or it may break a suspect's alibi stating that he or she was at another location at the time the crime was conducted.

4 Conclusion and Future Work

Location tracking is a crucial component on cyber-crime investigation and computer forensics. In this paper, we have proposed a new locating system in wireless networks based on multiple grids. Our new approach effectively contrasts the limitations of the existing locating techniques. More specifically, our proposed solution is able to distinguish intruders from outside or inside the building, and even narrow down to detect hackers on different floors in the same building. We have also described other important applications in which the new system can be used.

There are some issues that need to be addressed. First is the cost of the system. While the wireless grids are made from readily available APs, the external receivers with the directional antennas are not. Commercial grade antennas are still expensive. Although there are homemade antennas, it does not always meet the strict requirements and quality testing done by professionals. Second, an office may not always be able to align the upper and lower grids, which may affect the accuracy of determining the z-coordinates.

As part of our future work in this work we plan to implement the system with real world scenarios. The real world environment can have some unanticipated effects on our proposed locating system. A real-time working system can reveal these deficiencies and our system can be improved effectively and accordingly.

References

1. Aruba Wireless Networks: Location, Location, Location: RF Distance Measurement and Location-Based Services in Corporate Wi-Fi Networks. White Paper (2004)
2. Kipper, G.: Wireless Crime and Forensics Investigation. Auberbach Publications, Boca Raton
3. Urbas, G., Krone, T.: Mobile and Wireless Technologies: Security and Risk Factors. Trends & Issues in Crime and Criminal Justice 329
4. Velasco, E., Chen, W., Ji, P., Hsieh, R.: Challenges of Location Tracking Techniques in Wireless Forensics. In: 4th International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIHMSP 2008), Harbin, China (2008)
5. Want, R., Hopper, A., Falcao, V., Gibbons, J.: The Active Badge Locating System. ACM Transactions on Information Systems (TOIS) 10(1), 91–102 (1992)

Applying Public-Key Watermarking Techniques in Forensic Imaging to Preserve the Authenticity of the Evidence

Wen-Chao Yang[1], Che-Yen Wen[1], and Chung-Hao Chen[2]

[1] National Central Police University, No. 56, Shujen Rd.,
Takang Village, Kueishan Hsiang, Taoyuan County, 33304, Taiwan
[2] The University of Tennessee, 800 Andy Holt Tower,
Knoxville, TN 37996, USA

Abstract. The traditional verifying evidence method in court is to check the integrity of the chain of custody of evidence. However, since the digital image can be easily transferred by Internet, it is not easy for us to keep the integrity of the chain of custody.

In this article, we use the PKI (Public-Key Infrastructure), Public-Key Cryptography and watermark techniques to design a novel testing and verifying method of digital images. The main strategy of the article is to embed encryption watermarks in the least significant bit (LSB) of digital images. With the designed method, we can check the integrity of digital images by correcting public-key without side information and protecting the watermarks without tampering or forging, even the embedded method is open. Finally the proposed method can be applied in court to digital evidence testing and verification, and used to check the admissibility of digital image.

Keywords: Digital Image Forensic, Digital Image, Credibility, Integrity, Public-key Cryptography.

1 Introduction

Digital image data have many advantages, such as easily to modify, transfer and store. With the technology progressing, using digital imaging (photographing) devices has become more popular. However, the property of the digital data, easy to be, becomes serious challenges when using digital images in court. The challenge is the method of verifying digital image evidence in court is not established perfectly.

Generally, the key point about if the digital evidence is admitted by court has two conditions[1]. One is the warrant, is required to search and seize evidence, and the other is verifying the evidence handle procedure, is also called the chain of custody. The evidence handle procedure (chain of custody) is usually used to make sure the credibility and integrity of evidence.

The traditional method of verifying digital evidence can divide into three steps. The first step is to hash the evidence to generate the digest of evidence during seizing, the second step is to seal the digest for safekeeping and manage it carefully, and the last step is to verify the evidence with the digest in court. Therefore it is a big overhead for processing the lots of digital image evidence.

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 278–287, 2008.

According the foregoing, we design a protocol with secure hash function[2,3,4], Public-Key Cryptography[2,3,5] and digital watermarking techniques[6,7] through the Public-Key Infrastructure (PKI)[2,3,8] for handling the chain of custody of digital image evidence. And we embed the encryption watermark in the least-significant-bit (LSB) plate of digital image[6,7] to avoid interfering the information of the digital image. We describe our method in session II and make two novel experiments, one is tampering case and the other is forging case, to show the performance of our method in session Ⅲ. Finally we conclude the result and our future work.

2 The Proposed Method

Our novel protocol can divide into two procedures, one is the embedding procedure and the other is extracting procedure.

The embedding procedure of our method can be divided into five steps (We illustrate the embedding procedure in Fig.1).

1. The first step is to read original image and binary it (or a special image) with threshold 128 to be original watermark.
2. In the second step, we use secure hash function (SHA-1) to generate the "image fingerprint" of the original image.
3. In the third step, we XOR operate original watermark and the "image fingerprint".
4. In the forth step, we encrypt operated watermark with the privacy key of the seizer.
5. In the last step is to embed the encrypted watermark in the least-significant-bit (LSB) plate of the original image to replace the original least-significant-bit (LSB) plate.

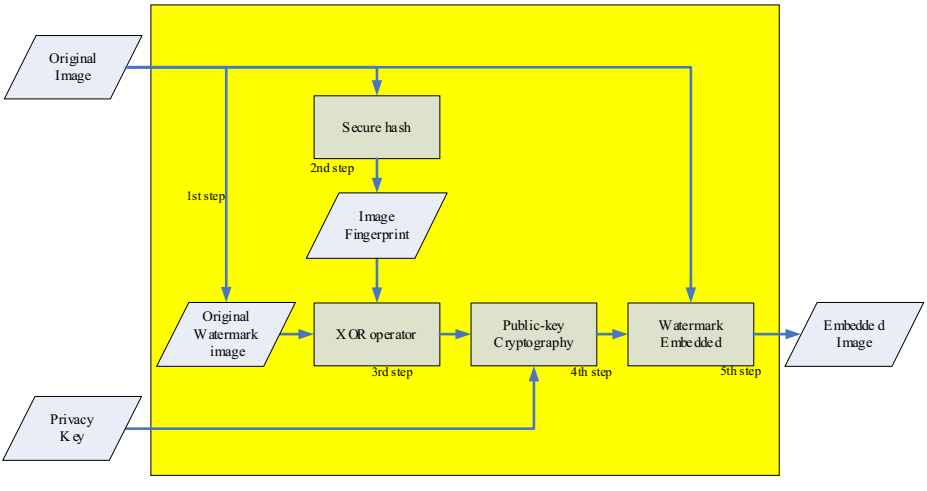


Fig. 1. The embedding procedure

The extracting procedure of our method can also be divided into six steps (We illustrate the extracting procedure in Fig.2).

1. The first step of all is to get Public-key of the seizer through Public-Key Infrastructure (PKI).
2. In the second step, we cut the least-significant-bit (LSB) plate from the embedded image.
3. In the third step, we to decrypt the LSB plate of embedded image with Public-key of the seizer to get the original watermark.
4. In the forth step, we use secure hash function (SHA-1) to generate the "image fingerprint" of the embedded image.
5. In the fifth step, we XOR operate decrypted LSB plate and the "image fingerprint".
6. In the final step, we compare the result bit plate of the fifth step and the watermark (the original watermark was generated from the embedded image with threshold 128, or we can also input the special image if the watermarking image in embedded) to get the result image.

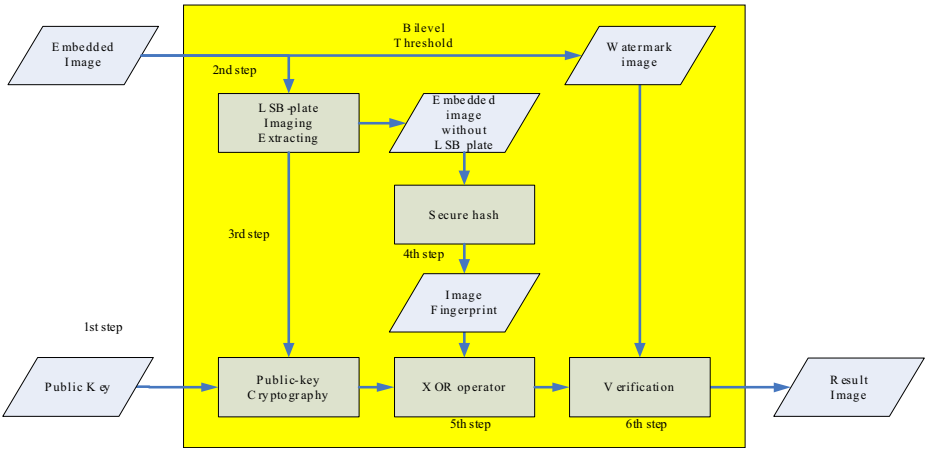


Fig. 2. The extracting procedure

3 Experiment Results

In this session, we make two novel experiments to show the correct and performance of our proposed method. The first experiment is a tampering case and the other experiment is a forging case.

3.1 The Tampering Experiment

We simulate a novel tampering test to show the correct and performance of our proposed method (use public-key (5,129) and privacy-key (17,129)). The experiment describes as follow:

3.1.1 The Embedding Procedure

In the first step, we use an image of real criminal case as Fig.3 and an original watermark with gray-level threshold 128 as Fig.4.

In the second step, we hash the original image without LSB plate and get the "image fingerprint", "6c174fa9b65f469cd8c5b7eac6fd1e169cc35c14". In the third

Fig. 3. The original image

Fig. 4. The original watermark

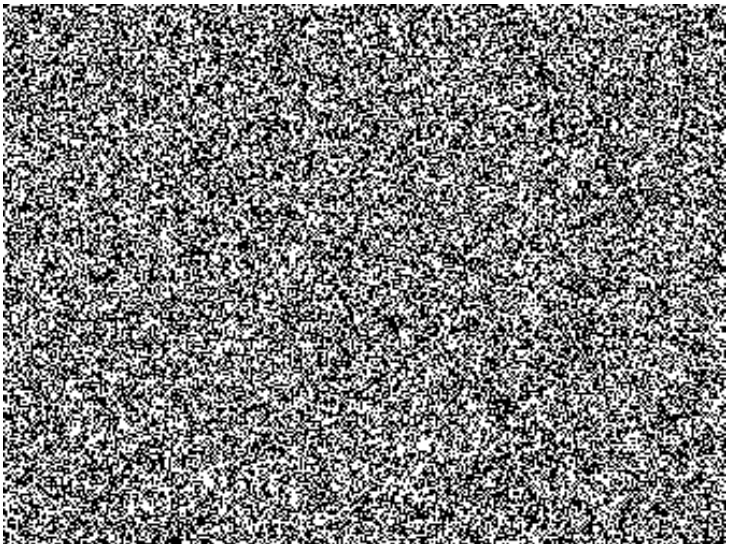


Fig. 5. The encrypted watermark

step, we XOR operate the original watermark (Fig.7) and the "image fingerprint". In the forth step, we use RSA encryption algorithm with a simple privacy-key (17,129) to encrypt our result of the third step, then generate the encrypted watermark as Fig.5.

In the final step (fifth step), we use the encrypted watermark to replace to the LSB plate of the original image, and then generate the embedded image as Fig.6.

After embedding watermark, we try to modify the license plate of the embedded image with Adobe Photoshop CS and the result shows as Fig.7.

Fig. 6. The embedded image

Fig. 7. The modified (tampered) image

3.1.2 The Extracting Procedure

In the first step, we get Public-key(5,129) through Public-Key Infrastructure (PKI). In the second step, we extract the least-significant-bit (LSB) plate of the modified image (Fig.8) and the modified image without LSB plate (Fig.9).

In the third step, we decrypt the LSB plate of embedded image with the correct public-key to get the modified watermark. In the fourth step, we hash the modified image without LSB plate (Fig.9) and get the "image fingerprint" of the modified

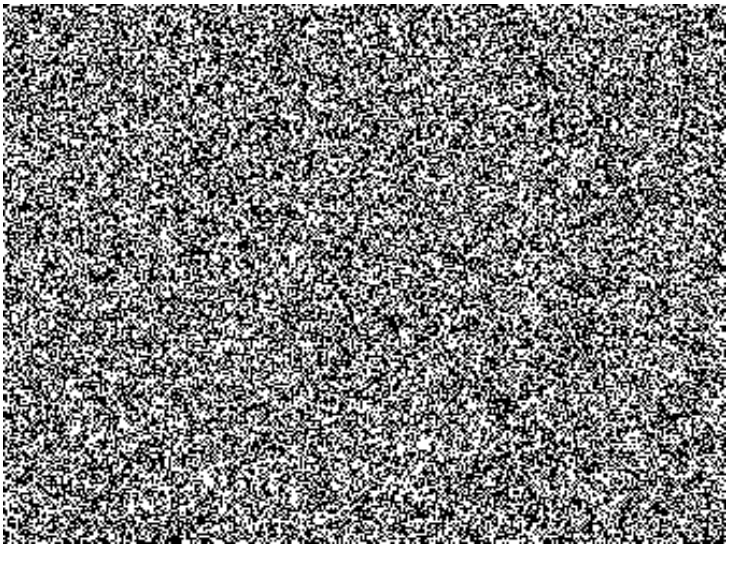


Fig. 8. The LSB plate of the modified image

Fig. 9. The modified image without LSB-plate

image without LSB plate, "d6ed6361f054be90d6ffdb778386ba4d40f746b1". In the fifth step, we XOR operate the result of the third step and the "image fingerprint" of the modified image without LSB plate, then we get the result as Fig.10.

In the last step (six step), we compare with the binary plate of the embedded image (with gray-level threshold 128) and the extracted watermark (Fig.10), then get the verify result, as Fig.11.

Fig. 10. The exacted watermark with correct public-key(5,129)

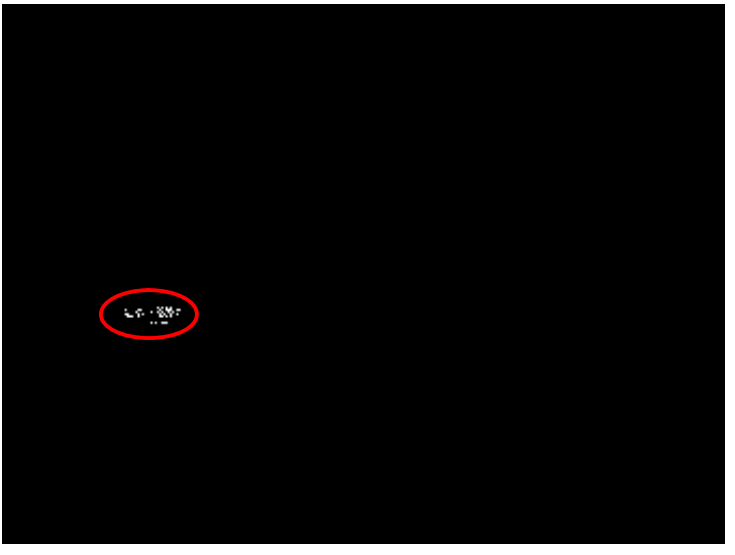


Fig. 11. The result image of verifying digital image evidence

3.2 The Forging Experiment

We simulate a novel forging test to show the correct and performance of our proposed method (assume the correct public-key is (5,129) and correct privacy-key is (17,129) and generate the forged image with wrong privacy-key (23,129), as Fig.12). After the extracting procedure, we get the extracted watermark, as Fig.13.

Fig. 12. The forged image

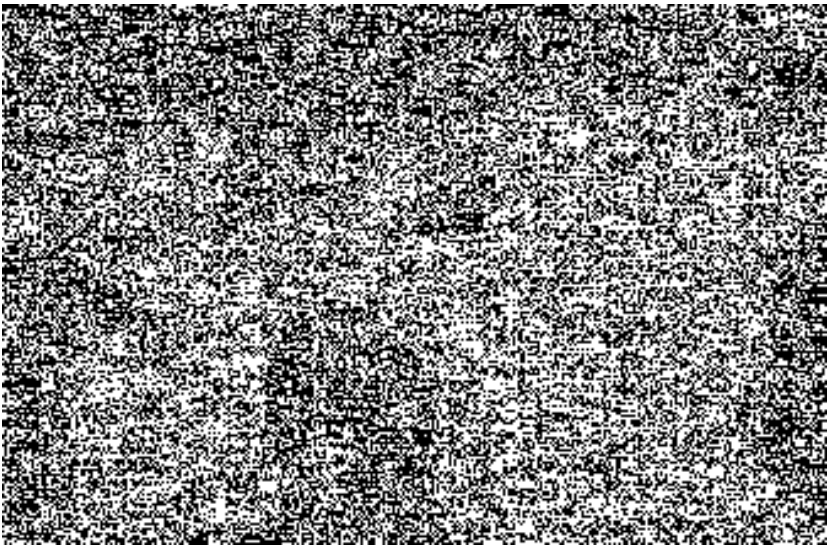


Fig. 13. The extracted watermark of the forged image

According to the forgoing experiments, we can detect the modified digital image or the forged digital image and find out the modified part through the proposed method. Therefore we show our method is helpful to enhance the credibility (the modified experiment) and integrity (the forged experiment) of digital image evidence.

4 Conclusion

In this article, we propose a novel protocol for verifying the chain of custody of a digital image evidence only need the digital image itself. It is also suited to use in digital image forensic. Then we make two experiments with real criminal images to show the performance of our protocol. By the experiments we show that we can detect the modified digital image or the forged digital image and find out the modified part through the proposed method. It is helpful to enhance the credibility and integrity of digital image evidence.

Our method is flexible. We use the SHA-1 algorithm to be our secure hash function and the RSA algorithm to be the public-key cryptography in this article. If there is any new and mature secure hash function or public-key cryptography, we can update in our protocol.

Our method is also comfortable and easy. All of the needed data in the verifying the chain of custody of digital image evidence by our method is just the digital image itself and the seizer's public-key. Everyone can get the Public-key of the seizer through Internet.

Up to now, we can apply our protocol in non-compression image, both in monochrome and color images. By the progress of digital imaging technology, the low-loss rate compression digital image format (such as JPEG) is also used to photo the crime scenes. We will devote to design a new protocol that can be helpful to

preserve the integrity of low-loss rate compression digital image format in court in next step.

Finally, we hope the proposed protocol can be applied in police offices and in forensic laboratories to preserve the integrity of digital images; it should be helpful to check the integrity of digital images in court.

References

1. Casey, E.: Digital Evidence and Computer Crime. Academic Press, London (2000)
2. Schneier, B.: Applied Cryptography, 2nd edn. John Wiley & Sons, Chichester (1996)
3. Stallings, W.: Cryptography and Network Security: Principles and Practices, 2nd edn. Prentice Hall International, Englewood Cliffs (1999)
4. Federal Information Processing Standards: SECURE HASH STANDARD (SHA), FIPS PUB 180-1 (1993)
5. Rivest, R.L., Shamir, A., Adleman, L.: A Method for Obtaining Digital Signatures and Public-Key Cryptosystems. Communications of the ACM 21(2), 120–126 (1978)
6. Cox, I., Miller, M., Bloom, J., Miller, M.: Digital Watermarking: Principles & Practice. Academic Press, London (2002)
7. Katzenbeisser, S., Fabian, A.P.P.: Information Hiding techniques for steganography and digital watermarking. Artech House (2000)
8. Adams, C., Lloyd, S.: Understanding PKI: concepts, standards, and deployment considerations, 2nd edn. Pearson Education, Inc., London (2003)

Computer Forensics and Culture

Yi-Chi Lin[1], Jill Slay[2], and I.-Long Lin[3]

[1] University of South Australia, School of Computer Information and Science, Mawson Lakes Boulevard, Mawson Lakes, South Australia 5095, Australia
Linyy021@students.unisa.edu.au
[2] University of South Australia, Defence and Systems Institute (DASI), Mawson Lakes Boulevard, Mawson Lakes, South Australia 5095, Australia
Jill.Slay@unisa.edu.au
[3] Central Police University, Department of Information Management, No.56, Shuren Rd., Guishan Shiang, Taoyuan County 333, Taiwan
Paul@mail.cpu.edu.tw

Abstract. Based on theory gained from the science and culture research domain, this paper considers the relationship between computer forensic science and culture. In order to develop a theoretical relationship between computer forensics and culture, this paper examines computer forensics as science and discusses the universal nature of science. It points to an ongoing cross-cultural work being carried out among computer forensic experts in Australia and Taiwan.

Keywords: Computer Forensics, Culture.

1 Introduction

The last two decades has seen a massive shift in the usage of consumer electronic devices, and devices that were once science fiction are now commonplace. Digital equipments such as computer workstations, notebook computers, mobile phones, and digital cameras have become an accepted part of everyday lives. Their use, in conjunction with global communications media such as the internet, has seen people connected in ways that were inconceivable in the past.

With this rise in use of technology and instant communications comes its misuse, either to take advantage of flaws in protocols or mechanisms, or as a means of facilitation, for example, of computer viruses, computer hacking or as a means of criminal communication. In this case, traditional means of criminal and civil investigations are no longer appropriate to deal with computer-based crimes or crimes utilizing computers as potential sources of evidence. The investigation of crimes that involve digital devices and communications requires new processes and technologies, so as to provide means of analyzing all data in ways that is legally acceptable and able to provide insight and accuracy. This is the field of computer forensics [1, 2, 3, 4, 5, 6, 7, 8]. The term 'forensics' refers to finding evidence for the 'trier of fact', or for the court. Thus, computer forensics can be considered as finding computer-based evidence (digital evidence) for the 'trier of fact', or for the court.

Research into computer forensics mainly focuses on technical and legal issues. Academic researchers always regard computer forensics as multidisciplinary containing

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 288–297, 2008.

at least containing computer science and law. However, some studies [9, 10, 11, 12, 13] show that culture is one of the factors which is able to affect science and its applications. Their work provides a brand new standpoint for this research to consider the role of computer forensics. Based on concepts found in their research, this paper examines the relationship between computer forensics and culture.

The concepts of computer forensics and culture are introduced in section 2 and section 3 respectively. Section 4 further discusses 'is culture universal?'. The relationship between culture and science (science applications) is discussed in section 5. In section 6, new discoveries of this paper are listed. The relationship between computer forensics and culture is confirmed. Further works is shown in section 7.

2 What Is Computer Forensics?

Whilst in practical terms the field of computer forensics essentially is the investigation and analysis of digitized data, where the arbitration may be suitable for court or tribunal, there are several intricacies and discussions, both from an operational and an academic perspective, that detail the limitations, scope and boundaries of this area. An in-depth understanding of the field of computer forensics and its following issues is the foundation of this paper, and therefore a consistent definition must be chosen for the term, as this potentially impacts upon several areas of this work.

2.1 Definition of Computer Forensics

Before any further discussions of computer forensics, the definition of this field must be clarified and discussed. In this section, several definitions of computer forensics are listed. Through analysis of these definitions, all confusions and inconsistencies can be removed and a single definition, appropriate for this work, may be agreed upon. The works of this section discuss and compare these definitions to determine the most appropriate one for this thesis.

Pollitt [14] provided one of the earliest definitions for the field, and stated that "*Computer forensics is the application of science and engineering to the legal problem of digital evidence. It is a synthesis of science and law. At one extreme is the pure science of ones and zeros. At this level, the laws of physics and mathematics rule. At the other extreme, is the courtroom.*". This definition is defining in that it does encompass both the process and possible outcomes of computer forensics, that analysis of computers is the aim but that all outcomes must be legally acceptable. However, this definition is also cumbersome and dated in several ways. Describing computer forensics as 'a synthesis of science and law' is necessarily vague for both the definitions of the terms 'science' and 'law', as any computer forensic process would only encompass a subset of scientific process. The definition put forward by Pollitt [14] similarly is vague about the outcomes at 'the courtroom'. This term does not consider the gamut of other arbitration and legal aspects, rather that computer forensic outcomes need not end up in a courtroom, but be legally acceptable. Therefore, whilst the definition developed by Pollitt [14] was at the time groundbreaking, it is not succinct and seems dated based on definitions that have since emerged.

McKemmish [4] described the concept of forensic computing as "*forensic computing which is the process of identifying, preserving, analysing and presenting digital evidence in a manner that is legally acceptable.*". Although the definition of computer forensics provided by McKemmish is very concise, this definition indeed provides necessary information for people to understand the essential meaning of computer forensics. McKemmish used the four processes to express the core concept of computer forensics, and stated that the product of these processes, digital evidence, must be legally acceptable. His definition infers that the computer forensics is a multi-discipline subject, and its contribution is that this definition is the earliest definition provided in the Australia.

Palmer [15] gave the definition of computer forensics as "*The use of scientifically derived and proven methods toward the preservation, collection, validation, identification, analysis, interpretation, documentation and presentation of digital evidence derived from digital sources for the purpose of facilitating or furthering the reconstruction of events found to be criminal, or helping to anticipate unauthorized actions shown to be disruptive to planned operations.*". The definition was developed in the Digital Forensic Research Workshop (DFRWS), and is the consensus made by academic researchers. In other words, this is to say that the biases are rule out. Thus, the quality of this definition of computer forensics is reliability and trustworthiness.

After referring the above explanations of computer forensics,, the definition of computer forensics proposed by Palmer [15] is chosen as our official definition of computer forensics in this paper. This is because the definition was created and agreed by a group of academic researchers. Their professional knowledge provides the completeness of the definition, and shows that this definition is reliable and trustworthy.

2.2 Purpose of Computer Forensics

As computer forensics forms part of an investigation into misuse, and the outcomes must be acceptable within a court of law or arbitration process, there is a heavy reliance upon non-technical areas. As given by Yasinsac et al. [16], this misunderstanding can be clarified as "*computer forensics is multidisciplinary by nature, due to its foundation in two otherwise technologically separate fields (computing and law)*". This is a position supported by the majority of academic works in the area. Given the multi-disciplinary nature of computer forensics, a working definition of the purpose of computer forensics must be discussed to give context to this paper. There are several definitions listed below, so as to show the importance the purpose of computer forensics. The voice of the researcher is listed after these definitions.

As stated by Pollitt [14], "*To get something admitted into court requires two things. First, the information must be factual. Secondly, it must be introduced by a witness who can explain the facts and answer questions. While the first may be pure science, the latter requires training experience, and an ability to communicate the science*". In addition, Pollitt [17] expanded upon this concept and said "*investigators and others have, by trial and error, evolved methods which will allow the discovery of evidence form storage media that will satisfy the twin requirements of science and law*".

Many other academic researchers provided their own arguments to describe the purpose of computer forensics. This paper lists some definitions in this paragraph. Forte [18] stated "*The simple guiding principle universally accepted both in technical and judicial spheres, is that all operations have to be carried out as if everything will one day have to be presented before a judge.*". Lin et al. [19] has agreed with the concept that the computer forensics is a multidisciplinary issue, and stated that "*the difficulty of cyber crime investigation is because two fields are involved. They are law and computer science.*". Meyers and Rogers [20] identified that "*computer forensics is in the early stages of development and as a result, problems are emerging that bring into question the validity of computer forensics usage in the United States federal and state court systems.*".

From the previous statements, it can be understood that the ultimate goal of computer forensics is to provide legitimate and correct digital evidence in the court rather than merely examining digital equipment or analyzing digital data. This is to say that computer forensics is not a subject which only contains technical issues, but it is a discipline contains at least sophisticated computer techniques and complicated legal issues. In short, since the essential goal of computer forensics is to provide legitimate digital evidence within a court, the primary consideration of the computer forensic examiners is the trustworthiness of the digital evidence. Trustworthiness is the fundamental property of digital evidence when it is presented within a court of law.

2.3 Is Computer Forensics Science?

The answer of this question is undoubtedly 'YES'. The thorough reasons are given as follows. There are numerous studies about forensic science (often shortened to forensics) [21, 22, 23, 24, 25]. This apparently shows that forensics itself is a science. Although the term 'forensics' normally refers to DNA testing, fingerprint testing, drug testing, blood testing, it can not be denied that computer forensics is a branch of forensic science. This infers that computer forensics is a subject of science. Two demonstrations are provided, and they indeed echo this standpoint.

Volonino et al. [26] stated that "*In the United States, courts at many levels and in many jurisdictions have recognized computer forensics as a bona fide scientific method of discovering and proving facts that can be used to solve and prosecute crimes. Recently, the forensic discipline of acquiring, preserving, retrieving, and presenting electronic data has been called* ***computer forensics science****. The new emphasis on this field as a science is important because it shows that computer forensics is a discipline based on generally accepted scientific methods. This recognition helps reinforce the credibility and stature of computer forensics investigators.*". It is straightforward to understand the meaning of this reference. It simply shows that the field of computer forensics is contained by the science.

Kruse and Heiser [27] stated that "*Computer forensics involves the preservation, identification, extraction, documentation and interpretation of computer data. It is often more of an art than a* ***science****, but as in any discipline, computer forensics specialists follow clear, well-defined methodologies and procedures, and flexibility is expected and encouraged when encountering the unusual.*". Although Kruse and Heiser think computer forensics is more of an art than a science, the point is that they admitted that computer forensics itself is a science.

The two book references point out that the field of computer forensics is within science. This is very helpful for the proposed discussion in this paper. At this moment, this paper already asserts that computer forensics is science. In the following sections, this paper demonstrates that culture is not universal, and culture is able to influence science applications.

3 What Is Culture?

Many different definitions of culture are found in literature ranging from anthropology, philosophy, teleology through to information system research. For example and from a business perspective, Buzzell [28] stated the definition of culture as "*A convenient catchall for the many differences in market structure and behaviour that cannot readily be explained in terms of more tangible factors*". But it is necessary to look further back and to gain a fundamental understanding. Geertz [29], a famous anthropologist stated as "*Man is an animal suspended in webs of significance he himself has spun. I take culture to be those webs and the analysis of it is not an experimental science in search of law but an interpretative one in search of meaning*.".

Hofstede [30] is another famous academic researcher who defined culture as "*the collective mental programming of the people in an environment. Culture is not a characteristic of individuals; it encompasses a number of people who were conditioned by the same education and life experience*.". Moreover, Hofstede's definition of culture was further interpreted by Shore and Cross [31], they stated the understanding of culture as "*Culture can be defined as the set of mental programs, established early in life and difficult to change, that control or influence an individual's responses in a given context*.".

After presenting the diverse definitions of culture, it is realized that culture is not a simple issue to discuss and/or to define. Culture can be considered a phenomenon rather than an entity. Therefore, if people from different perspectives observe cultural effects, they are very likely to obtain different outcomes or consequences. People can try to find out the '*best*' definition of culture, but they may never find out the '*true*' definition of culture. In conclusion, the cultural definition used in this thesis is the one proposed by Geertz [29]. This is because Geertz's cultural definition is the one accepted by most anthropologists and science educators. The credibility of this cultural definition is not a debateable issue.

4 Is Culture Universal?

The ultimate aim of this paper is to show the relationship between culture and computer forensics. In order to do this, a question, *is culture universal?*, should be asked first so as to further clarify this concept. If the answer of this question is '*no*', the link between different cultures and science then needs to be found and established. This link also represents the relationship between different cultures and computer forensics, since it is already proved that computer forensics is science.

In the following paragraphs, several academic references are presented to explain that the answer of this question is '*no*'. First of all, listed references show that culture is not universal, and culture has the ability to affect an individual's mind and thinking. In addition, the references also indicate that culture changes the way how individual constructs knowledge, and forms experience.

Vygotsky [32], Wertsch [33], and Wertsch et al. [34] had similar thoughts. They all acknowledged and agreed that culture and society are two major factors which influence cognitive development. This fact may not directly reveal the relation between culture and science, but it shows without a doubt that culture dominates the cognitive developments. This means that people with different cultural backgrounds are likely to have different cognitive developments. In other words, people with different cultural backgrounds are likely to act or to respond to an event differently. In addition, Solano-Flores and Nelson-Barber [35] made a similar statement as "*The conceptual relevance of cultural validity is supported by evidence that culture and society shape an individual's mind and thinking.*". The last but not the least, Greenfield [36] stated that the way for an individual to construct knowledge and create meaning from experience is affected by culture. Greenfield is actually saying that how an individual thinks about things, and how he or she solves problems are affected by culture.

In conclusion, the aforementioned explanations and descriptions show that the answer of the question, *is culture universal?*, is '*no*'. They point out cultures are different, and the behaviors of an individual are affected by different cultures and/or diverse societies. Moreover, this standpoint is agreed by Hofstede [12, 13, 37, 38].

5 Culture and Science (Science Applications)

After showing computer forensics is science and culture is not universal, the relationship between culture and science applications must be established so as to prove the standpoint proposed by this paper.

In the following paragraphs, sources, discussed the relation between culture and science education, are presented. The references found out that culture is one of the facts which can affect science education. From this, the researcher can claim an official link between culture and science applications. If culture has influence on science education, it is considered that culture has influences on science applications. The link between culture and science applications is based on an individual's cognition and comprehension, which affects his or her learning of science.

According to Cobern et al. [39], "*Worldview research in science education dates at least to Kilbourn [40] and Proper, Wideen, and Ivany [41]. Cobern [42] borrowed a logico-structural model of worldview from anthropologist Kearney [43] in an attempt to bring greater coherence and sophistication to worldview research in science education.*". In addition, Cobern et al. [39] stated that "*Indeed, the criticism of modern Western scientific views of Nature (e.g. Merchant [44]) provides reason to investigate the views fostered in a science class. This line of thought suggests a broad agenda for cultural studies research in science education, premised on the assertion that all ideas, including scientific ones, are expressed within a cultural setting [29]). Thus, one must ask how the cultural setting of the science teacher and curriculum*

compares with student cultural settings.". From these two accounts, two conclusions can be made. The first conclusion is to say that the worldview research in science education is actually another way to express research of culture and science education. The second conclusion is that the statements show that the research for worldview (culture) and science education exists for more than 20 years, and many scholars now devoted themselves in this area. This is to say that this issue has been established as a serious research topic, and is a worthwhile topic to be studied.

Theocharis and Psimopoules [45], Holton [46], and Levitt and Gross [47] all considered that the seriousness of the worldview differences, and these worldview differences cause scientists' attentions for antiscientific views. This means that worldview differences are indirectly shaping an individual's views of science, and also implicitly echoes the topic that the culture is not universal, and culture has certain influences on science.

Lee and Fradd [48] said in order to improve higher academic standards for all students, certain endeavors are necessary. These endeavors must contain a method which can mediate academic content with students' cultural experiences. From this, the academic content is ensured to be reachable and meaningful for every student.

According to Solano-Flores and Nelson-Barber [35], "*So-called culturally responsive pedagogy is based on developing educational methods that are situated in students' cultural experiences (Bartolome [49]) and, in the case of science, views transitioning from a student's life-world to a science classroom as a cross-cultural experience (Aikenhead and Jegede [50]). Such an approach is especially relevant in the case of cultural groups for whom knowledge is highly contextualized. Understanding a given culture's ways of knowing and traditional knowledge is the basis for establishing connections between cultural content and academic content (Nelson-Barber and Estrin [51]).*". From these statements which were provided by diverse scholars, the relation between culture and science education is confirmed.

Stanley and Brickhouse [52] also stated that "*Multiculturalism in science education has become an increasingly rich area of study as educators struggle to find answers to the question of how to teach science in a multicultural world.*". Without a doubt, this statement clearly shows that culture is a factor for educators to teach science in a multicultural world. In other words, culture is one of the reasons to affect science education.

In summary, the link between culture and science applications is built due to the relationship between culture and science education. The link between culture and science applications responds to the core concept of this thesis, and shows that the research question is a worthwhile issue to be studied.

6 New Discoveries

Computer forensics can easily be considered a pure science subject. In other words, the results of a computer forensic program are not going to be changed under any circumstance. This statement is right from technical perspective. However, this research analyses computer forensics from cultural perspective, and provide a unique insight to examine computer forensics. This research believes that since the computer forensics is operated by individuals, and individuals are affected by their exclusive

culture. They are very likely to behave differently for the same situation. For example, operator A and operator B are assumed to have different cultural backgrounds. It further assumes that the operator A will use method A for input A to produce output A. In this case, the finding in this paper shows that it can not guarantee operator B will use method A when he or she receives input A. In other words, operator B is likely to choose method B for input A. This is because different cultures make them think and work differently.

In here, there is one standpoint need to be clarified. This paper is not saying that operator A and operator B will produce different results when they both use method A for input A. In this situation, they, of course, will obtain the same outcome (output A). What is discussed in this research is that they may think and respond differently when they encounter the same digital incident or event. Thus, they may have different reactions at the end. It is believed that the different cognitive developments are the reason caused this situation. In other words, this paper is actually saying that the culture is the major reason which causes this situation.

The value of this research is to propose the possible cultural differences within the field of computer forensics. According to our best knowledge, this is the first paper discusses the issues of computer forensics from cultural perspective.

7 Future Works

In the near further, Delphi survey and case study are going to be conducted to examine the inference discovered by this research. Australia and Taiwan are two examples to demonstrate the new discovery. Five dimensions will be used to examine the cultural attitudes are: *Current Situation Dimension*, *Policy and Organization Dimension*, *Education Dimension*, *Law Dimension*, and *Personal Preference and Skill Dimension*. These five dimensions are able to examine and to demonstrate the findings of this paper from different perspectives, and are also helpful for the overall understanding of the discoveries.

Australian and Taiwanese computer forensic experts are going to be invited for Delphi survey and case study. The potential respondents are computer forensic workers which include practical workers in law enforcement agencies and/or private companies, and academic researchers. By contrasting the Australian and Taiwanese findings, it is believed that concrete proves can be provided.

References

1. Cooper, P.: Speciation in the Computing Sciences: Digital Forensics as an Emerging Academic Discipline. In: Proceedings of the 2nd Annual Conference on Information Security Curriculum Development (2005)
2. Francia, G.A.: Digital Forensics Laboratory Projects. Journal of computing Sciences in Colleges (2006)
3. Li, X., Seberry, J.: Forensic Computing. In: 4th International Conference on Cryptology (2003)
4. McKemmish, R.: What is Forensic Computing? (1999), http://www.aic.gov.au/publications/tandi/ti118.pdf

5. Mercuri, R.T.: Challenges in Forensic Computing. Communications of ACM (2005)
6. Patel, A., Ciardhuain, S.O.: The Impacts of Forensic Computing on Telecommunications. IEEE Communications Magazine, 64–67 (2000)
7. Slay, J., Hannan, M., Broucek, V., Turner, P.: Developing Forensic Computing Tools and Techniques within a Holistic Framework: an Australian Approach. In: Proceedings from the Fifth Annual IEEE SMC Information Assurance Workshop (2004)
8. Wilsdon, T., Slay, J.: Digital Forensics: Exploring Validation, Verification and Certification. In: First International Workshop on Systematic Approaches to Digital Forensic Engineering (2005)
9. Slay, J.: Implementing Modern Approaches to Teaching Computer Science: A Life Long Learning Perspective. In: 16th World Computer Congress (2000)
10. Slay, J.: Culture and Sensemaking in Information Warfare. In: The 2nd Australian Information Warfare and Security Conference (2001)
11. Hofstede, G.: Culture and Organizations. International Studies of Management and Organization, pp. 15–41 (1981)
12. Hofstede, G.: Culture Differences in Teaching and Learning. International Journal of Intercultural Relations, 301–320 (1986)
13. Hofstede, G.: Organising for Cultural Diversity. European Management Journal, 390–397 (1989)
14. Pollitt, M.M.: Computer Forensics: An Approach to Evidence in Cyberspace. In: Proceedings of the National Information Systems Security Conference, pp. 487–491 (1995)
15. Palmer, G.: A Road Map for Digital Forensic Research. In: The First Digital Forensic Research Workshop (2001)
16. Yasinsac, A., Erbacher, R.F., Marks, D.G., Pollitt, M.M., Sommer, P.M.: Computer Forensics Education. IEEE Security and Privacy Magazine, 15–23 (2003)
17. Pollitt, M.M.: Principles, Practices, and Procedures: An Approach to Standards in Computer Forensics. In: Second International Conference on computer Evidence, pp. 10–15 (1995)
18. Forte, D.: Principles of Digital Evidence Collection. Network Security, 6–7 (2003)
19. Lin, I.-L., Yang, H.-C., Gu, G.-L., Lin, A.-C.: A Study of Information and Communication Security Forensic Technology Capability in Taiwan. In: IEEE 37th Annual 2003 International Carnahan Conference on security Technology, pp. 386–393 (2003)
20. Meyers, M., Rogers, M.: Computer Forensics: The Need for Standardization and Certification. International Journal of Digital Evidence (2004)
21. Robertson, B., Vignaux, G.A.: Interpreting Evidence: Evaluating Forensic Science in the Courtroom. J. Wiley, New York (1995)
22. Ramsland, K.M.: Forensic Science of CSI. Berkley Trade (2001)
23. Saferstein, R.: Criminalistics: An Introduction to Forensic Science. Prentice-Hall, Englewood Cliffs (2001)
24. Platt, R.: Crime Scene: The Ultimate Guide to Forensic Science. DK ADULT (2006)
25. Crispino, F.: Nature and Place of Crime Scene Management within Forensic Sciences. Science and Justice (2007)
26. Volonino, L., Anzaldua, R., Godwin, J., Kessler, G.C.: Computer Forensics: Principles and Practices. Pearson / Prentice Hall (2007)
27. Kruse, W.G., Heiser, J.G.: Computer Forensics: Incident Response Essentials. Addison-Wesley, Reading (2002)
28. Buzzell, R.D.: Can you standardize multinational marketing? Harvard Business Review, pp. 102–113 (1968)
29. Geertz, C.: The Interpretation of Cultures. Basic Books, Inc., New York (1973)

30. Hofstede, G.: Motivation, Leadership, and Organization: Do American Theories Apply Abroad? Organizational Dynamics, pp. 42–63 (1980)
31. Shore, B., Cross, B.J.: Exploring the role of national culture in the management of large-scale international science projects. International Journal of Project Management, 55–64 (2005)
32. Vygotsky, L.S.: Mind in Society: The Development of Higher Psychological Processes. Harvard University Press, Cambridge (1978)
33. Wertsch, J.V.: Vygotsky and the Social Formation of Mind. Harvard University Press, Cambridge (1985)
34. Wertsch, J.V., Del Rio, P., Alvarez, A.: Sociocultural Studies of Mind. Cambridge University Press, New York (1995)
35. Solano-Flores, G., Nelson-Barber, S.: On the Cultural Validity of Science Assessments. Journal or Research in Science Teaching, 553–573 (2001)
36. Greenfield, P.M.: Culture as Process: Empirical Methods for Cultural Psychology. In: Berry, J.W., Poortinga, Y., Pandey, J. (eds.) Handbook of Cross-Cultural Psychology: Theory and Method, vol. 1, pp. 301–346. Allyn & Bacon, Boston (1997)
37. Hofstede, G.: Cultural Constraints in Management Theories. Academy of Management Executive, pp. 81–94 (1993)
38. Hofstede, G.: The Business of International Business is Culture. International Business Review, 1–14 (1994)
39. Cobern, W.W., Gibson, A.T., Underwood, S.A.: Conceptualizations of Nature: An Interpretive Study of 16 ninth Graders' Everyday Thinking. Journal of Research in Science Teaching, 541–564 (1999)
40. Kilbourn, B.: World Views and Science Teaching. In: Munby, H., Orpwood, G., Russel, T. (eds.) Seeing Curriculum in a New Light, University Press of America, Lanham (1984)
41. Proper, H., Wideen, M.F., Ivany, G.: World View Projected by Science Teachers: A Study of Classroom Dialogue. Science Education, 542–560 (1988)
42. Cobern, W.W.: World View Theory and Science Education Research. NARST Monograph NO. 3. Manhattan, KS: National Association for Research in Science Teaching (1991)
43. Kearney, M.: World View. Chandler & Sharp, Novato (1984)
44. Merchant, C.: The Death of Nature: Women, Ecology, and the Scientific Revolution. Harper & Row, San Francisco (1989)
45. Theocharis, T., Psimopoules, M.: Where Science has Gone Wrong. Nature, 595–598 (1987)
46. Holton, G.: Science and Anti-Science. Harvard University Press, Cambridge (1993)
47. Levitt, N., Gross, P.R.: The Perils of Democratizing Science. The Chronicle of Hifher Education (1994)
48. Lee, O., Fradd, S.H.: Science for All, including Students from Non-English Language Backgrounds. Educational Researcher, 12–21 (1998)
49. Bartolome, L.: Beyond the Methods Fetish: Toward a Humanizing Pedagogy. Harveard Educational Review, 173–194 (1994)
50. Aikenhead, G.S., Jegede, O.J.: Cross-Cultural Science Education: A Cognitive Explanation of a Cultural Phenomenon. Journal of Research in Science Teaching, 269–287 (1999)
51. Nelson-Barber, S., Estrin, E.: Culturally Responsive Mathematics and Science Education for Native Students. San Francisco: Far West Laboratory for Educational Research and Development (1996)
52. Stanley, W.B., Brickhouse, N.W.: Teaching Sciences: The Multicultural Question Revisited. Science Education, 35–49 (2000)

E-Commerce Security: The Categorical Role of Computers in Forensic Online Crime

Fahim Akhter

College of Information Technology, Zayed University, Dubai, United Arab Emirates
fahim.akhter@zu.ac.ae

Abstract. Computer forensic technology with the encapsulation of prevention issues could protect data and information from hackers. Despite being a new field, great advancement have been made in computer crime investigation to protect information and data. Powerful evidence processing tools have been developed and there is a move towards standardization. The purpose of this paper is to generate an interest in and awareness of computer forensics by providing some basic information. This paper emphasizes the role of computers in crime and to give guidance for dealing with computers in that role. The fundamental purpose of categories discussed in this paper is to emphasize the role of computers in crime and to give guidance for dealing with computers in that role. These categories can be used to develop procedures for dealing with digital evidence and investigating crimes involving computers. These guidelines are still in their early stages, especially with regards to digital evidence.

1 Introduction

The presence of online security is a must for financial affairs such as high speed verification and authentication, and protection of information and data. E-commerce is prone to the attack and the cheat of hackers. With the growth of e-commerce, hacker attacking becomes more sophisticated and advance so that it now turns out to be an top barrier in respect of development of e-commerce. Therefore, online vendors are striving to develop detecting & forewarning products to protect customers information and data against cheat and intrusion. There is no fully proven method against hacker intrusion, whereas hacker could continually find the flaws of the open system, and intrude into website. The term computer security focus on protecting information systems from attack, with forensic techniques used peripherally in the intrusion detection community [1] [2]. Forensic is derived from the Latin 'Forensis', meaning making public. Forensic Science is the applied use of a body of knowledge or practice in determining the cause of death. Forensic systems engineering is the discipline investigating the history of Information Technology failures [3]. In 1994, the US Department of Justice created a set of categories and an associated set of search and seizure guidelines (USDOJ 1994, 1998). These categories made the necessary distinction between hardware and information, which is useful when developing procedures from a probative standpoint. This paper analyse the final three categories that refer to information all fall under the guise of digital evidence.

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 298–303, 2008.

2 Role of Computers in Crime

Forty years ago, people did not imagine computers would be an vital part of everyday life. Now computer technology is ubiquitous as compare to crime playing different role such as instrument of the crime and the target of the crime. By the 1970s, electronic crimes were increasing, especially in the financial sector. Financial companies were using the mainframes, used by trained or educated people with specialized skills. It was much easy to manipulate computer data to commit fraud for financial gain by internal employees. One of the most famous crimes of the mainframe era is the one-half cent crime [6]. It was common for banks to track monies in accounts to the third decimal place or more. Banks used and still use the "rounding up" accounting method when paying interest. If the interest applied to an account resulted in a fraction of a cent, that fraction was used in the calculation for the next account until the total resulted in a whole cent. It was assumed that sooner or later every customer would benefit. On more than one occasion, computer programmers corrupted this method by opening an account for themselves and writing programs that diverted all the fractional monies into their accounts. In smaller banks, this practice amounted to only a few hundred dollars a month. In larger banks with branch offices, however, the amount of money reached hundreds of thousands of dollars. At this time, most law enforcement agencies didn't know enough about computers to ask the right questions or to preserve evidence for trial.

As PCs gained popularity and began to replace mainframe computers in the 1980s, may different OSs emerged. Apple released the Apple 2E in 1983, and then launched the Macintosh in 1984. Computers such as the TRS-80 and the Commodore 64 were the machines of the day. CP/M machines (the 8088 series) and Zeniths were also in demand. The disk operating system (DOS) was available in many varieties, including PC-DOS QDOS, DR-DOS, IBM-DOS, and MS-DOS. Forensic tools at that time were simple, and most were generated by government agencies such as the Royal Canadian Mounted Police (RCMP) in Ottawa, which had its own investigative tools, and the U.S. Internal Revenue Service (IRS). Most of these tools were written in C and assembly language and were not used by the general public.

In the mid-1980s, a new tool called Xtree Gold appeared on the market. It recognized file types and retrieved lost or deleted files. Norton DiskEdit soon followed and became the best tool for finding deleted files. You could use these tools on the most powerful PCs of that time; IBM-compatible computers had 10 MB hard disks and two floppy drives. In 1987, Apple produced the Mac SE, a Macintosh that was available with an external EasyDrive hard disk with 60 MB of storage. At this time, the Commodore 64 was a popular computer that still used standard audiotapes to record data, so the Mac SE represented an important advance in computer technology.

The software component of the information system comprises applications, operating systems, and assorted command utilities[6]. The exploitation of errors in software programming accounts for a substantial portion of the attacks on information. Software programs are often created under the demanding constraints of project management, which limit time, cost, and manpower. Information security is all too often implemented as an afterthought rather than developed as an integral component from the beginning. In this way, software programs become an easy target

of accidental or intentional attacks. The security of information and its systems entails securing all components and protecting them from potential misuse and abuse by unauthorized users. A computer can play one of three roles in a computer crime. A computer can be the target of the crime, it can be the instrument of the crime, or it can serve as an evidence repository storing valuable information about the crime. In some cases the computer can have multiple roles. When considering the security of information system, it is important to understand the concept of the computer as the subject of an attack as opposed to the computer as the object of an attack. When a computer is the subject of an attack, it is used as an active tool to conduct the attack. When a computer is the object of an attack, it is the entity being attacked. There are two type of attacks: direct attacks and indirect attacks. A direct attack is when a hacker uses his personal computer to break into a system. An indirect attack is when a system is compromised and used to attack other systems, such as in a distributed denial of service attack. Direct attacks originate from the threat itself. Indirect attacks originate from a system or resources that itself has been attacked, and is malfunctioning or working under the control of threat.

By the early 1990s, specialized tools for computer forensics were available. The International Association of Computer Investigative Specialists (IACIS) introduced training on currently available software for forensic investigations, and the IRS created search-warrant programs [6]. However, no commercial software for computer forensics was available until ASR Data created Expert Witness for the Macintosh. This software can recover deleted files and fragments of deleted files. One of the ASR Data partners later left and developed EnCase, which has become a popular computer forensics tool. As computer technology continued to grow, more computer forensics software was developed. The introduction of large hard disks posed new problems for investigators. Most DOS-based software doesn't recognize a hard disk larger than 8 GB. Because contemporary computers have hard disks of 40 to 100 GB and larger, changes in forensics software are needed.

The specific role that a computer plays in a crime also determines how it can be used as evidence. When a computer contains only a few pieces of digital evidence, investigators might not be authorized to collect the entire computer. Hence, when a computer is the key piece of evidence in an investigation and contains a large amount of digital evidence, it is often necessary to collect the entire computer and its contents. Additionally, when a computer plays a significant role in a crime, it is easier to obtain a warrant to search and seize the entire computer [6].

Several attempts have been taken to develop a language, in the form of categories, to help describe the role of computer in crime [5]. Categories can be useful provided they are used with an awareness of their limitations. The following section discussed the strength and weakness of three sets of categories in efforts to improve understanding of the role of computers in crime. The final three categories that refer to information all called hardware as contraband, hardware as an instrumentality, hardware as evidence, information as contraband, information as an instrumentality, and finally information as evidence.

3 Digital Evidence

A single crime can fall into more than one category [5]. It usually contains evidence of the offence. In 2002, USDOJ document was updated to keep up with changes in

technology and law and developed into a manual for " Searching and Seizing Computers and Obtaining Electronic Evidence in Criminal Investigation" (USDOJ 2002). While the guidelines gave hardware and information equal weight, the manual takes the position that, unless hardware itself is contraband, evidence, an instrumentality, it is merely a container for evidence.

3.1 Hardware as Contraband

Contraband is property that the private citizen is not permitted to possess. For example, under certain circumstances, it is illegal for an individual in the United States, to possess hardware that is used to intercept electronic communications (18 USCS 2512). The concern is that these devices enable individuals to obtain confidential information.

3.2 Hardware as an Instrumentality

When computer hardware has played a significant role in a crime, it is considered as instrumentality. This distinction is useful because, if a computer is used like a weapon in a criminal act, much like a gun or a knife, this could lead to additional charges or a heightened degree of punishment. The clearest example of hardware as the instrumentality of crime is a computer that is specially manufactured, equipped or configured to commit a specific crime. For instance, sniffers are pieces of hardware that are specifically designed to eavesdrop on a network.

3.3 Hardware as Evidence

The separate category of hardware as evidence is necessary to cover computer hardware that is neither contraband nor the instrumentality of a crime. For instance, if a scanner that is used to digitize child pornography has unique scanning characteristics that link the hardware to the digitized images, it could be seized as evidence.

3.4 Information as Contraband

A common form of information as contraband is encryption software. In some countries, it is illegal for an individual to posses a computer program that can encode data using strong encryption algorithms because it give criminals too much privacy. If a criminal is caught but all of the incriminating digital evidence is encrypted, it might not be possible to decode the evidence and prosecute the criminal.

The value of information comes from the characteristics it possesses [6]. The nature of the information characteristics depends on information changes. Some characteristics affect information's value to users more than others do. The timeliness of information can be a critical factor, because information often loses all value when it is delivered late. A minor delay in information delivery could cause a tension when the end users' need for full access to the information. The critical characteristics of information is based on availability, accuracy, authenticity, confidentiality, integrity, utility and possession [6].

Availability assist authorized users, persons or computer systems to access information without interference or obstruction, and to receive it in the required format. Information has accuracy when it is free from mistakes or errors and it has the value that the end user expects. If information contains a value different from the user's expectations, due to the intentional or unintentional modification of its content, it is no longer accurate.

Authenticity of information is the quality or state genuine or original, rather than a reproduction or fabrication [6]. Information authentic when it is the information that was originally created, placed, stored, or transferred. Information has confidentiality when disclosure or exposure to unauthorized individuals or systems is prevented. Confidentiality ensures that only those with the rights and privileges to access information are able to do so. When unauthorized individuals or systems can view information, confidentiality is breached. Information has integrity when it is whole, complete, and uncorrupted. The integrity of information is threatened when the information is exposed to corruption, damage, destruction, or other disruption of its authentic use. The threat of corruption can occur while information is being stored or transmitted. Many computer viruses and worms are designed with the explicit purpose of corrupting data. For this reason, a key method for detecting a virus or worm is to look for changes in file integrity as shown by the size of the file.

The utility of information is the quality or state of having value for some purpose or end. Information has value when it serves a particular purpose. This means that if information is available, but not in a format meaningful to the end user, it is not useful. Thus, the value of the information depends on its utility. The possession of information is the quality or state of having ownership or control of some object or item. Information is said to be in one's possession if one obtains it, independent of format or other characteristics.

3.5 Information as an Instrumentality

Information can be the instrumentality of a crime if it was designed or intended for use or has been used as means of committing a criminal offense. Programs that computer intruders use to break into computer systems are the instrumentality of a crime. These programs, commonly known as *exploits*, enable computer intruders to gain unauthorized access to computers with a specific vulnerability.

3.6 Information as Evidence

This is the richest category of all. Many of our daily actions leave a trail of digits. All service providers keep some information about their customers. These records can reveal the location and crime of an individual’s activities, such as items purchased in a supermarket, car rentals and gasoline purchases, automated toll payment, mobile telephone calls, online banking.

4 Conclusions

The fundamental purposes of categories discussed in this paper are to emphasize the role of computers in online crime and to give guidance for protecting data and

information. Depending on the nature of the case, the seizure of computer hardware and software itself can be justified on one of above theories without regard to the data it contains. In many cases, hardware may be confiscate under more than one theory. For example, if a hacker uses his computer to spread viruses into other systems, his computer may constitute both an instrumentality of the offense. The implementation of forensic technology in online commerce would enhance the customers trust. These categories can be used to develop procedures for dealing with digital evidence and investigating crimes involving computers. These guidelines are still in their early stages, especially with regards to digital evidence.

References

1. Rosenblatt, K.: High Technology Crime. KSK Publications (1995)
2. Icove, D., Seger, K., VonStorch, S.: Computer Crime: A Crimefighter's Handbook. O'Reilly & Associates, Sebastopol (1995)
3. Dalcher, D.: Forensic ECBS: A Situational Assessment, ecbs. In: 7th IEEE International Conference and Workshop on the Engineering of Computer Based Systems, p. 390 (2000)
4. Phillips, N.: Guide to Computer Forensic and Investigations, Thomson Course Technology (2006)
5. Casey, E.: Digital Evidence and Computer Crime. Academic Press, Inc., Orlando (2004)
6. Michael, W.: Principles of Information Security, Thomson Course Technology (2005)

Forensic Artifacts of Microsoft Windows Vista System

Daniel M. Purcell[1] and Sheau-Dong Lang[2]

[1] Economic and Computer Crimes Unit, Seminole County Sheriff's Office
Sanford, FL 32773, USA
[2] School of Electrical Engineering & Computer Science, University of Central Florida
Orlando, FL 32816, USA

Abstract. This paper reviews changes made to Microsoft Windows Vista system from earlier Windows operating system (such as XP) and directs attention to system artifacts that are of evidentiary values in typical computer forensics work. The issues addressed include: NTFS on-disk structure, file system's directory structures, symbolic links, and recycle bin; we also briefly mention artifacts related to Windows mail, paging file, thumbnail caching, and print spooling.

Keywords: forensics, Windows Vista, NTFS, artifacts, symbolic link, recycle bin, paging file, thumbnail, print spooling.

1 Introduction

Microsoft's newest operating system, Windows Vista, presents a new era of challenges for computer forensic examiners. Vista was released to the public in early 2007. With the introduction of new technologies and changes to existing operating system components, forensic examiners must understand the fundamental concepts of the operating system and how to identify and decode the on-disk or logical structures. Currently, many computer forensic examiners are encountering digital media with Windows Vista installed and attempting to examine and analyze the data for criminal investigations. Without a full understanding of how the operating system or file system handles the data or specific functions, examiners may assert incorrect conclusions or explanations of a given artifact. Also, examiners must be able to test and validate their forensic software tools to ensure the date is being interpreted in a proper manner.

This paper will address several components of the NTFS file system [3] which are part of Windows Vista and various operating system components that have changed since the last versions of Microsoft's operating system line, Windows XP or Windows Server 2003. Specifically, we will discuss the overall functionality and on-disk structure related to the NTFS file system, physical and logical structures, file slack, new features of NTFS, directory structure changes (default locations), and Recycle Bin; we will also briefly describe Windows Mail (replacement for Outlook Express), thumbnail caching for folders, paging file (virtual memory), and print spooling. While there are numerous other features and artifacts in Windows Vista, the scope of this paper will be limited to the topics mentioned.

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 304–319, 2008.

2 Test Environment and Software

To simplify the testing process, we installed Windows Vista Enterprise within a VMware Virtual Machine (aka *.vmdk file) [7]. VMware Workstation version 6.01 was used to create and manage the virtual machine or installation of Windows Vista Enterprise. Initially, we created the virtual machine with an approximate disk size of ten gigabytes. The resulting virtual physical disk contains 20,971,520 sectors. The installation took approximately 7.6 gigabytes of allocated space within the single, primary partition, which was formatted NTFS. All of the default options were selected during the installation process, and once completed, we "activated" the product via the Internet. It should be noted that the license is registered to our agency under a volume licensing agreement. Throughout the process, we used EnCase Enterprise Edition 6.7 to preview the virtual machine in its live state [4]. We used the logical evidence file feature of EnCase to capture the logical contents of a given file or folder, and used EnCase to view the contents of the *.vmdk (virtual machine disk file) created by VMware Workstation. When appropriate, a forensic image of the virtual disk was created and examined. Screenshots noted in this paper were taken with VMware Workstation or Windows XP as the host machine. Several screenshots of the data examined were taken from EnCase as well.

3 Physical and Logical Disk Structure

As previously noted, Windows Vista Enterprise was installed on a ten gigabyte virtual disk through VMware Workstation. The virtual disk file was allocated on the host machine, which means the space was allocated immediately on the host hard disk drive. No other data was copied to the virtual disk file prior to installation. During the installation phase, we accepted the default values and completed the installation. Prior to changing any settings or values within Windows Vista Enterprise, the virtual machine was powered off and previewed with Encase. The VMDK was added to Encase and examined under the disk view tab within EnCase. The master boot record is located at physical sector zero (0), and the general structure from previous version of Windows installations or other partitioning utilities and standards appears to be the same.

The boot code begins at file offset zero and continues for 446 bytes. The partition table begins at sector offset 446 and is sixty-four bytes in length. Each entry within the partition table is sixteen bytes in length. The typical signature of 0x55AA at the end of the master boot record still exists (see Fig. 1).

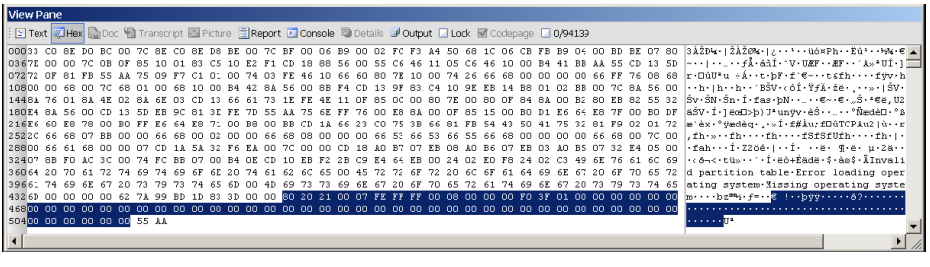


Fig. 1. Master boot record at physical sector zero – partition table highlighted (64 bytes)

The most notable change in a default installation of Windows Vista is the location of the volume boot record. It should be noted that we only created one primary partition during testing, and therefore, further testing must be performed to test the creation of additional primary partitions or extended partitions are created with Vista. The volume boot record is located at physical sector 2,048 (see Fig. 2). The structure of the volume boot record has not changed. The volume boot record is created during the formatting process of the file system; in this case, NTFS 3.1. The volume boot record contains various information about the volume or partition, including: the file system in use (NTFS), sectors per cluster, total sectors, volume name, volume serial number, bytes per sector, and other attributes of the volume or partition. It should be noted that the space between the master boot record and volume boot record, commonly referred to as the reserved area, totals 2,047 sectors and contains no relevant data. In fact, every byte within each sector in this area is 0x00.

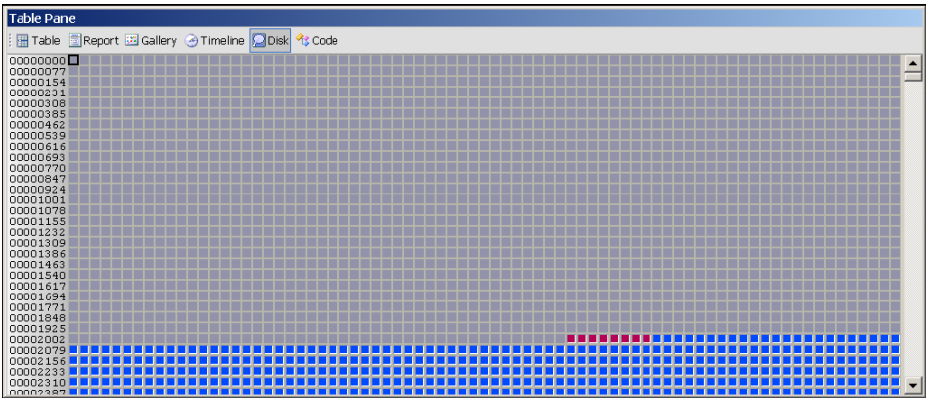


Fig. 2. Virtual view of physical sector 0 through 2,618 – VBR starts at physical sector 2,048

As with NTFS version 3.0, version 3.1 creates a backup copy of the volume boot record when the volume or partition is formatted. The backup copy is typically the last sector of the volume. In the test installation of Windows Vista Enterprise, the primary partition contains 20,967,424 sectors. The last sector of the volume is within the volume slack area (remaining sectors not allocated to a cluster) and is a backup copy of the volume boot record located at physical sector 2048. This is important for a forensic examiner to know when file system corruption is noted during analysis.

4 NTFS File System Observations

By default, Windows Vista is installed with the NTFS file system. As used with Windows XP and 2003 versions, version 3.1 is used with Windows Vista versions. The version of the file system is noted in the $Volume file on the root of the volume, and the version of the file system is 3.1. It should be noted that $Volume has an entry within the $MFT (master file table), but the file size is zero bytes in length, and therefore, it is

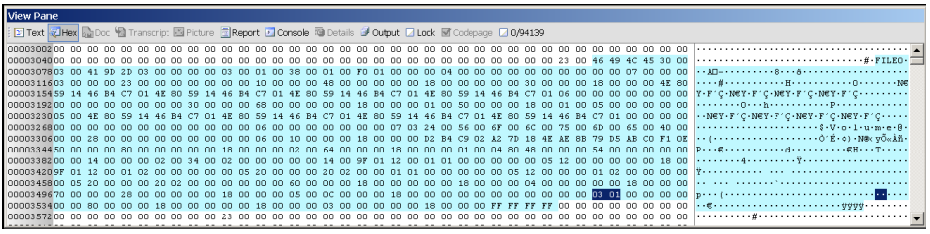


Fig. 3. $Volume entry within the $MFT – record offset 456 (2 bytes for NTFS version)

common that most forensic applications will not display any logical data. This is due to the lack of a data attribute, but at MFT record offset 456, the next two bytes are shown as 0x0301, which is interpreted as NTFS version 3.1 (See Fig. 3).

The root of the volume contains the core files of the NTFS file system. The core files did not change from recent versions of Windows, which is consistent with the use of NTFS 3.1 in Vista. The structure of the master file table ($MFT) did not change. Each record is 1,024 bytes in length by default, and the common attributes such as the standard information attribute, file name attribute, and data attribute did not change. $MFTMirr still contains a backup of the first four records of the file system, $MFT, $MFTMirr, $Logfile, and $Volume. $Bitmap still tracks cluster allocation on the volume as well. For every byte there are eight bits, and each bit tracks a single cluster's allocation (0 = unallocated, 1 = allocated). In essence, the core files of the NTFS file system have not changed.

4.1 Symbolic Links and Junctions

One of the newest additions to NTFS 3.1 under Windows Vista is the addition of Symbolic Links. According to Microsoft, a symbolic link is a "file-system object that points to another file system object" (http://msdn2.microsoft.com/en-us/library/aa365680.aspx). Like a link file, the symbolic link points to a target folder, file, or object within the file system, and essentially, the symbolic link is transparent to the user within the Windows GUI. The purpose of a symbolic link is to aid in migration and application compatibility from one operating system to another. For example, the default parent user directory in Windows XP is **<root>\Documents and Setting**. In Windows Vista, the default parent user directory is **<root>\Users**. In order to accomplish seamless migration and backward compatibility with previous version of Windows, Microsoft implemented the symbolic link within the NTFS file system.

For example, **\Documents and Settings** is a symbolic link on the root of the volume. The file identifier or $MFT record number is 6,835. In order to view the raw MFT entry, one multiplies 6,835 × 1,024 (bytes within each MFT record), which equals 6,999,040, the byte offset to the **\Documents and Settings** MFT entry. When navigated to this offset within the $MFT, we found the contents of the file name attribute to be "Documents and Settings" written in Unicode. Next, when navigated to the Symbolic Link attribute identifier, which is 0x0000000C0, we found the size of the attribute is 88 bytes. Within the symbolic link identifier, one can clearly see the actual path is directed to **C:\Users**, which is actually stored in Unicode. This is

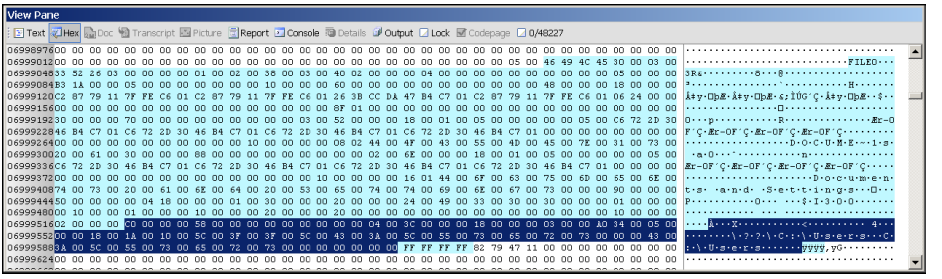


Fig. 4. MFT entry for \Documents and Settings – Symbolic Link attribute highlighted

significant for a forensic examiner as some programs written for Windows XP may make calls to **C:\Documents and Settings** within the Vista environment, and the forensic examiner must understand that the calls made by the operating system or application will be redirected to **C:\Users**. See Fig. 4 for more details.

Another feature of NTFS 3.1 that is now used in Windows Vista is junctions. Microsoft states that junctions (aka: soft link) "can link directories located on different local volumes on the same computer and are implemented through reparse points." Junctions are used to graft folders whereas symbolic links can link to files and folder. Junctions are referenced further in this document (http://msdn2.microsoft.com/enus/library/aa365006.aspx).

4.2 File Slack

Microsoft operating and file systems have made slight changes to file slack over the years. Before we identify how Vista handles file slack, it is important to define slack. As a general term, file slack refers to the space from the end of the file to the end of cluster that is not used by the logical file. File slack has two sub-terms that are applicable to this discussion, RAM and residual slack. RAM slack is the space from the end of the file to the end of the sector within the last cluster of the file; residual slack includes any remaining full sectors within the last cluster of the file. Since Windows 95b, the operating system and file system API writes 0x00 in the RAM slack area. Prior to Windows 95b (DOS 5 to Windows 95a), data from memory was written to the RAM slack area. RAM slack used to be a good location to look for data that was written to the memory buffers and later written to the RAM slack area. The residual slack area normally includes data from a previous file that existed within the cluster, but does not contain any data if the sector(s) were not written to by a previous file, folder, or object.

In Windows Vista, the behavior of slack has not changed from recent versions of Windows. The file system API still writes 0x00 from the end of the logical file to the end of the sector. The residual slack area may contain data from a previous file. Fig. 5 shows the end of a file and slack area of the physical file. The highlighted area is the RAM slack. One can see that the RAM slack contains only 0x00 while data from a previous file exists in the residual slack area (the area after RAM slack). The same observation was made after a sample of approximately fifty files.

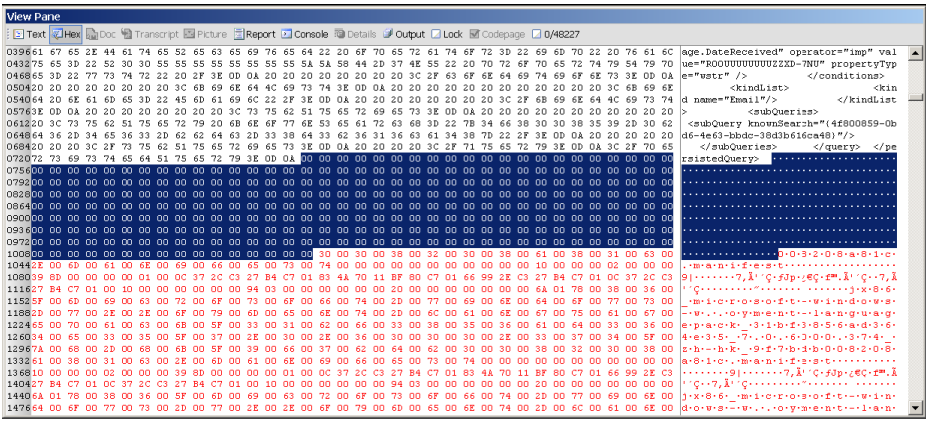


Fig. 5. File slack – RAM slack (highlighted area) and residual slack (after RAM slack) displayed.

4.3 Last Access Date and Time

By default, Windows Vista no longer records the last accessed date and time within the NTFS file system. This feature is controlled within the registry under the SYSTEM registry file. When the file is viewed with a forensic registry viewer such as Access Data's Registry Viewer [1] or EnCase version 6 (both support the registry format for Vista), the examiner must navigate to the current control set key (\ControlSet###). The subkey for examination is **System\ControlSet###\Control\ FileSystem**, and the value name is NtfsDisableLastAccessUpdate. The value data for this option is set to decimal value 1 (hex value 0x00000001), which means the option to disable the last access date from being updated is turned on. This can be changed to decimal value 0 (0x00000000), which will enable the recording of last accessed dates/times through the NTFS file system. We conducted random testing with a set of files by disabling and enabling this feature, and as expected the registry changes toggles the last access date/time for the files selected. This feature has a major impact on digital evidence as the last access date/time is commonly referred to as a reference for file activity on a given volume.

4.4 Directory Structure Changes

As noted in Section 4.1 junctions and symbolic links are used to redirect legacy paths to new directory locations. They aid in migration and application compatibility from one operating system to another. There are a number of changes to the directory structure within Windows Vista as compared to Windows 2000/XP/2003. As noted in the table below (Fig. 7), we have manually parsed the directory structure by locating the most notable symbolic links or junctions and tracing them to the new location for the folder and/or data. Since "C" is the common volume letter assigned to the boot volume containing the operating system, we used C: for all of the directory paths noted in Fig. 6.

2000/XP/2003 Directory Path (Symbolic Link in Vista)	Windows Vista Path to Data
C:\Documents and Settings	C:\Users
C:\Documents and Settings\Default User	C:\Users\Default
C:\Documents and Settings\All Users	C:\ProgramData
C:\Documents and Settings\<user>\Application Data	C:\Users\<user>\AppData\Roaming
C:\Documents and Settings\<user>\Cookies	C:\Users\<user>\AppData\Roaming\Microsoft\Windows\Cookies
C:\Documents and Settings\<user>\Local Settings	C:\Users\<user>\AppData\Local
C:\Documents and Settings\<user>\My Documents	C:\Users\<user>\Documents
C:\Documents and Settings\<user>\Nethood	C:\Users\<user>\AppData\Roaming\Microsoft\Windows\Network Shortcuts
C:\Documents and Settings\<user>\Printhood	C:\Users\<user>\AppData\Roaming\Microsoft\Windows\Printer\Printer Shortcuts
C:\Documents and Settings\<user>\Recent	C:\Users\<user>\AppData\Roaming\Microsoft\Windows\Recent
C:\Documents and Settings\<user>\SendTo	C:\Users\<user>\AppData\Roaming\Microsoft\Windows\SendTo
C:\Documents and Settings\<user>\Start Menu	C:\Users\<user>\AppData\Roaming\Microsoft\Windows\Start Menu
C:\Documents and Settings\<user>\Templates	C:\Users\<user>\AppData\Roaming\Microsoft\Windows\Templates
C:\Documents and Settings\<user>\My Documents\My Music	C:\Users\<user>\Music
C:\Documents and Settings\<user>\My Pictures	C:\Users\<user>\Pictures
C:\Documents and Settings\<user>\My Videos	C:\Users\<user>\Videos

Fig. 6. Comparison of Windows 2000/XP/2003 to Vista directory structure

There are additional folders within the user's folder that contain useful information for a forensic examiner. They are briefly discussed for reference in the table below.

C:\Users\<user>\Contacts	Similar to address book in XP. *.XML file created for contact. Associates with contacts for Windows Mail.
C:\Users\<user>\Desktop	Same as XP \Desktop folder
C:\Users\<user>\Downloads	Default folder for files downloaded from the internet via IE
C:\Users\<user>\Favorites	Same as XP. Contains *.URL files
C:\Program Files	Same as XP. Default location for programs or applications.

5 Windows Vista Artifacts

Operating systems typically produce artifacts, or features unique to the system, that are of evidentiary values during forensic examination and investigation. Many artifacts of earlier versions of Windows systems (2000/XP/2003) are well known. In this section we will describe our studies of the Recycle Bin of Windows Vista in detail. Due to space limitation we will only briefly comment on other artifacts including: thumbnail, Windows mail, page file, and print spooling.

5.1 Recycle Bin

The Windows Recycle Bin (aka recycler) is the default location for files that are deleted by a user using traditional functions, i.e., the delete icon within a given window or right-click mouse functions to delete a file. In short, these functions result in the file being moved from the current location to the Recycle Bin. Although the user may believe the file, folder, or object is deleted, the initial function is nothing more than a move function. Once the user selects the option to empty the recycle bin

within the GUI, the files are deleted from the Recycle Bin. There are a few changes to the Vista Recycle Bin from Windows XP, and we will discuss them when appropriate. The Recycle Bin is nothing more than a folder on the root of the volume. The folder's name is "$Recycle.Bin." Within the $Recycle.Bin folder are the user folders, which use the naming convention of the user's SID (Security Identifier). Just like Windows XP, the individual user folders are created when a user sends a file to the Recycle Bin the first time. This behavior was observed when user accounts were created during testing within the virtual machine of Vista Enterprise. The folder structure of three user accounts is displayed below in Fig. 7.

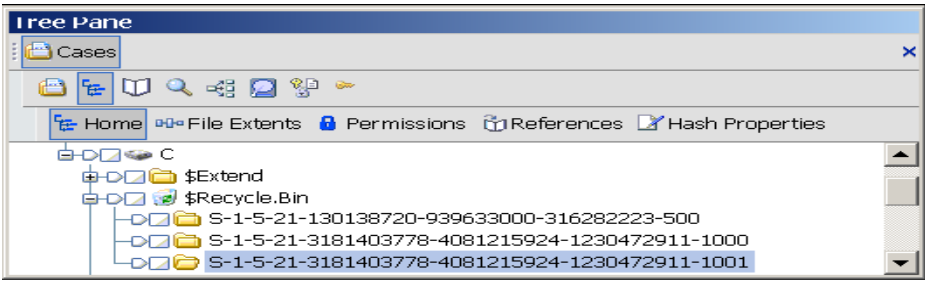


Fig. 7. Three user accounts – folder name is user's SID

In order to analyze the properties of the Recycle Bin and changes made to the file or object's MFT entry, we examined the MFT entry for the files deleted and supporting metadata file(s) at every stage of the sequence of events. The sequence of events is quite simple. We copied several graphics into the **\Pictures** directory for a given user. Next, we deleted the file(s) by executing common functions within the Windows GUI. When files were slated for deletion, they were moved to the user's Recycle Bin. Next, we either deleted individual files from the Recycle Bin or used the "Empty Recycle Bin" option to accomplish the final deletion of the file. Whenever we executed a given function within the Vista GUI, we examined the MFT entry for all relevant files and documented the findings. Although we repeated the steps numerous times to confirm the findings, only one file is used in the discussions that follow.

Example 1. The MFT entry of the file "Deputy Kevin2.JPG" prior to being deleted:

MFT Record number: **39068**
Record Status: **File, Allocated**
Name: **Deputy Kevin2.JPG**
Standard Information Attribute
File creation date: **10/17/07 22:52:54**
Last written date: **10/17/07 22:52:54**
Last MFT modification date: **10/17/07 22:52:54**
Last accessed date: **10/17/07 22:52:54**
Filename Attribute Deputy (SFN)
Parent MFT #: **36819**
Parent Directory: **Pictures**
Short File Name: **DEPUTY~1.JPG**
File creation date: **10/17/07 22:52:54**

Last written date: **10/17/07 22:52:54**
Last MFT modification date: **10/17/07 22:52:54**
Last accessed date: **10/17/07 22:52:54**
Filename Attribute (LFN)
Parent MFT #: **36819**
Parent Directory: **Pictures**
Long file name: **Deputy Kevin2.JPG**
File creation date: **10/17/07 22:52:54**
Last written date: **10/17/07 22:52:54**
Last MFT modification date: **10/17/07 22:52:54**
Last accessed date: **10/17/07 22:52:54**

After documenting the basic attributes of the MFT entry for the file, including the standard information attribute and file name attributes (short and long file names), we selected the option within the Vista GUI to delete the file and sent it to the Recycle Bin for the user. As suspected, the process of selecting a "delete" option a file within the GUI simply moves the location of the file to a new parent folder, **<system root>\$Recycle.Bin\<SID foldername>**. In fact, the file retained the original MFT record number, 39068. At MFT record offset 22 is the two-byte flag that indicates whether the file is allocated (0x0001) or deleted (0x0000). In this case, the 16-bit value resolves to decimal value 1, which indicates the file is currently allocated. This is consistent with the move function as it relates to the file system. Other notable changes include the MFT record modification date/time that is stored within the standard information attribute. When the file was moved from the original folder to the Recycle Bin, the MFT record modification date/time changed. This date and time is also known as the entry modified date and time, which is stored within the attribute in NTFS filetime format (64-bit). Prior to the file being moved to the Recycle Bin, the parent folder was **\Pictures** or MFT record number 36819. Once the file was moved to the Recycle Bin, the parent MFT record changed to 8227 or the folder name, "S-1-5-21-3181403778-4081215924-1230472911-1000." Prior to moving the file to the Recycle Bin, there were two file name attributes for the long and short file names within the MFT record. Once the file moved to the user's Recycle Bin, the file name attribute for the long file name was eliminated from the MFT record, and the file name attribute for the short file name changed to "$R78PIXA.JPG" from "Deputy~1.JPG." This is further confirmed by reviewing the entry modified date and time observed in the standard information attribute, which changed to the date/time when the file was moved to the user's Recycle Bin. Again, we have conducted numerous tests that resulted in the same changes to the MFT record of a given file.

Example 2. The MFT entry of the file after it was deleted (moved to Recycle Bin):

MFT Record number: **39068**
Record Status: **File, Allocated**
Name: **Deputy Kevin2.JPG**
Standard Information Attribute
File creation date: **10/17/07 22:52:54**
Last written date: **10/17/07 22:52:54**
Last MFT modification date: **10/17/07 22:59:49**
Last accessed date: **10/17/07 22:52:54**
Filename Attribute
Parent MFT #: **8227**
Parent Directory: **S-1-5-21-3181403778-4081215924-1230472911-1000**

Short file name: **$R78PIXA.JPG**
File creation date: **10/17/07 22:52:54**
Last written date: **10/17/07 22:52:54**
Last MFT modification date: **10/17/07 22:52:54**
Last accessed date: **10/17/07 22:52:54**

When the file selected for deletion by the user was moved to the Recycle Bin, a new "support" file was created in the user's Recycle Bin called "$I78PIXA.JPG" which accompanies the original file moved to the Recycle Bin by the user. This support file is similar to the content of the INFO2 file present in Windows 2000/XP/2003 [2], but it does not contain the same content nor does it have a similar structure. However, this file is specific to the file that was deleted. Moreover, it appears that every file that is deleted by the user and sent to the Recycle Bin, creates a file with similar contents, but the file name is slightly different and has eight bytes in length. During testing, we found that the support file begins with the first two characters of "$I", which is consistent with a hidden file that is not viewable to a normal user. As stated, the file name is random in terms of characters, but the first two characters observed were always "$I". The content of the file contains the logical size of the original file (in bytes) and an eight byte filetime date/time, which is consistent with the date and time of the file's deletion based on the local clock and stored in filetime format. In addition, the original file path and file name are stored within this file in Unicode. The details of the data contained within the support file are listed below. Based on testing and analysis, it is apparent that the support file for the target/original file that was moved to the Recycle Bin contains the logical size of the original file and date/time of the file, folder, or object's deletion along with the original file name and file path. The specifics of the support file will be discussed in detail later in this paper.

Example 3. The MFT entry of the support file in the Recycle Bin:

MFT Record number: **16187**
Record Status: **File, Allocated**
Name: **$I78PIXA.JPG**
Standard Information Attribute
File creation date: **10/17/07 22:59:49**
Last written date: **10/17/07 22:59:49**
Last MFT modification date: **10/17/07 22:59:49**
Last accessed date: **10/17/07 22:59:49**
Filename Attribute
Parent MFT #: **8227**
Parent Directory: **S-1-5-21-3181403778-4081215924-1230472911-1000**
Filename: **$I78PIXA.JPG**
File creation date: **10/17/07 22:59:49**
Last written date: **10/17/07 22:59:49**

Last MFT modification date: **10/17/07 22:59:49**
Last accessed date: **10/17/07 22:59:49**
Data Attribute
The data for this file is resident within the MFT record. The logical size of the original file is 2074631 bytes. The logical size of the file is stored within this file in an eight byte filed that is decoded little endian as a 64-bit value (2074631). We noted the presence of filetime (64-bit) date/time data, and when decoded with the proper offset from GMT (-0400), the date/time is consistent with the date and time of the file's deletion, October 17, 2007 at 22:59:49 hours. After the file was deleted, we observed the system clock, and this was the exact date/time of deletion. Inside of the same file, the

original directory path of the file deleted was present in Unicode format. The data was decoded in Unicode as "C:\Users\Dan\Pictures\Deputy Kevin2.JPG", and of course, the data is consistent with the original path and file name.

After analyzing the data in the Recycle Bin (original file and support file), we emptied the Recycle Bin, which ultimately resulted in the file being deleted within the Vista GUI. We examined the MFT entry for the original file, and at MFT record offset 22 (within the 56-byte header of the entry), we located the flag for allocation or deletion. As expected, the two-byte value changes from decimal 1 to decimal 0 (0x0000), which is the indicator for a file that is deleted. The only change to the data within the MFT entry was this single change from 0x0001 to 0x0000. The remaining attributes for the MFT record did not change. This is further confirmed by reviewing the MFT entry for the original file after the file was deleted from the Recycle Bin. Next, we navigated to the starting cluster of the file and found the data within the starting and contiguous clusters were not impacted by deletion from the Recycle Bin. Although the $Bitmap obviously changed the clusters allocated to this file to unallocated (available), the data was still present and recoverable with common forensic tools and techniques. This behavior is consistent with previous versions of Windows 2000/XP/2003 in terms of the file system minor changes and no impact on the data within the clusters on the volume.

Example 4. The MFT entry of the file after it was emptied from the Recycle Bin:

MFT Record number:	**39068**
Record Status:	**File, Deleted**
Name:	**Deputy Kevin2.JPG**
Standard Information Attribute Deleted Deputy Kevin2.jpg	
File creation date:	**10/17/07 22:52:54**
Last written date: 1	**0/17/07 22:52:54**
Last MFT modification date:	**10/17/07 22:59:49**
Last accessed date:	**10/17/07 22:52:54**
Filename Attribute Deleted Deputy Kevin2.jpg	
Parent MFT #:	**8227**
Parent Directory:	**S-1-5-21-3181403778-4081215924-1230472911-1000**
Filename:	**$R78PIXA.JPG**
File creation date:	**10/17/07 22:52:54**
Last written date:	**10/17/07 22:52:54**
Last MFT modification date:	**10/17/07 22:52:54**
Last accessed date:	**10/17/07 22:52:54**

The overall change to the Recycle Bin in Windows Vista is the method for tracking the file, folder, or object's file name, file path, and date/time of deletion within the support file. In addition to the original and support files discussed in this section, we deleted in excess of one-hundred files during testing. We found that all of the original files sent to the Recycle Bin by a user have a short file name that begins with the first two characters "$R" followed by six random characters and the file's original extension. The support file has a file name that begins with the two characters "$I" followed by six random characters that match the six random characters of the original file's short file name. In addition, the support file contains the same file extension of the original file. While more research is required to determine the random characters generated by the operating system for the original

file, there is a clear relationship between the original and support files based on the consistent file naming sequence.

Examiners can derive the relationship between the original file and support file by examining the short file name of the original file and the file name of the support file. Once identified, the examiner can examine the contents of the support file that contains the logical file size of the original file, date and time of the file's deletion, original file name, and original file path. Fig. 8 is a screenshot of EnCase version 6, which displays the deleted file and support file, $I178PIXA.JPG, in the right pane (checked). The short file name of the deleted file is also displayed in the screenshot.

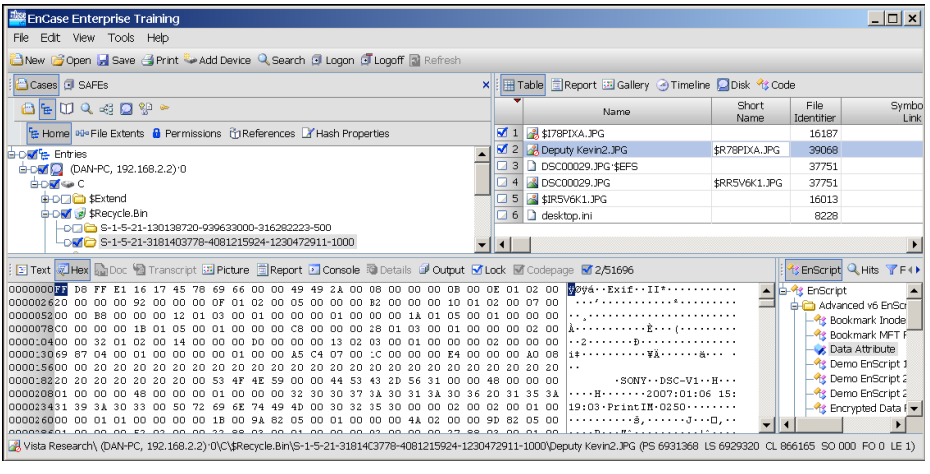


Fig. 8. Screenshot of EnCase v6 – deleted file and support file, $I178PIXA.JPG

The structure of the support file is relatively simple. All of the support files that we tested are 544 bytes in length (resident within the $MFT). We confirmed the file size by examining the MFT record for each support file and the corresponding data attribute. Each support file begins with 0x01. It appears the file header or first eight bytes of the file is 0x000000000001. The next eight bytes beginning at file offset 8 is reserved for the logical file size of the original file deleted. For example, if a file has a logical file size of 671,107,285 bytes, the entry is this section of the support file is 0x00000000280048D5. When this value converted from hexadecimal to decimal (little endian), the resulting value is 671107285, the logical file size of the original file. The next eight bytes beginning at file offset 16 is the filetime date and time value that represents the date and time that the original file was deleted by the user. The date and time is directly recorded from the local clock. For example, if the value is 0x01C81389D4F718F0 is calculated as a filetime value, the resulting date/time is "Sun, 21 October 2007 02:26:48 UTC." With most utilities, one can select an offset from GMT or UTC. Following the header of the support file, logical file size of the original file, and filetime value is the full path and file name of the original file, which is stored in Unicode.

The structure of the support file provides the forensic examiner with relevant artifacts for an examination. In the past, the Windows Recycle Bin or Recycler did

not record the logical file size of the original file. This is useful information and may assist investigators or examiners in tracking a specific file. Even if the Recycle Bin is emptied by the user, which results in the file's deletion from the file system, we discovered that the entry within the MFT is not altered by Vista. In fact, the data attribute is completely in tact and not altered by the deletion process. It is possible for a forensic examiner to parse the file system or $MFT and locate deleted support file entries. In some circumstances, it may be possible for the examiner to locate portions of the support file's data in the $Logfile, MFT record slack, file slack, or unallocated clusters. Knowing the structure and content of the support file is extremely important to understand should an examiner discover data similar to this structure.

The table below (Fig. 9) describes the offsets within the support file and their value. Fig. 10 is a screenshot of the data contained within one of the support files examined.

File Offset	# of Bytes	Value	Data
0	8	0x0000000000000001	Support file – file header
8	8	64-bit little endian	Hex to decimal = logical file size of original file in bytes
16	8	FILETIME date/time	Date & time of the file/folder/object's deletion
24	520 (max)	Unicode	Full path & file name of deleted file

Fig. 9. Table of offsets and data within the Recycle Bin's support file for the original file

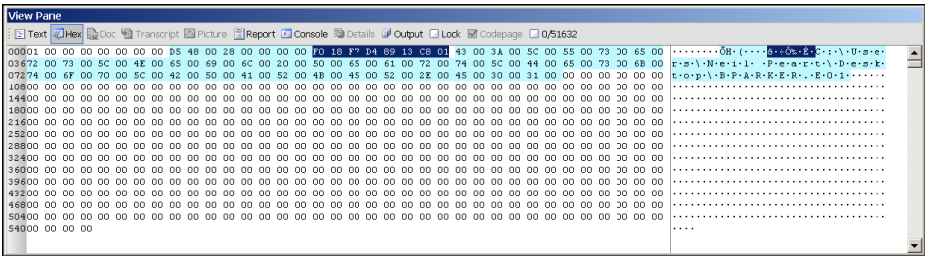


Fig. 10. Recycle Bin support file for original file – shown in hexadecimal and plain text

5.2 Thumbnail Chching

In Windows 2000/XP/2003 Microsoft implemented a thumbnail database file called thumbs.db within a folder that is viewed under the thumbnails view. This is an option that is available within any folder; thumbs.db contains a file's name, last written date/time, and a small thumbnail image of the file in JPEG format (header = 0xE0FFD8FF). The folder view option allows the user to view graphics, video, document, or other files in a thumbnail view (small graphic display). When viewed with forensic software such as EnCase or AccessData's FTK, the examiner can conduct a logical analysis of the file's content. In Windows Vista, the thumbs.db file is no longer available or present within the parent folder.

It appears Windows Vista displays user folders in thumbnail view by default. The user can globally change this setting in the Control Panel by selecting Folder Option in classic view and change the option "Always show icons, never thumbnails"

by checking the adjacent box. The registry file NTUSER.DAT for each user contains settings for Windows Explorer; it contains the subkey **NTUSER.DAT\Software\Microsoft\Windows\CurrentVersion\Explorer\Advanced** and a value "IconsOnly." By default the setting for the value is 0x00000000 (thumbnail view), but if the value is set to 0x00000001, the user will not view the files in a thumbnail view.

Windows Vista isolates the thumbnail files(s) in the folder **\Users\<UserName>\AppData\Local\Microsoft\Windows\Explorer**. The file names include thumbcache_1024.db, thumbcache_256.db, thumbcache_96.db, and thumbcache_32.db. Within these files are the actual thumbnail graphic of the file as displayed in the folder, and graphic formats include JPEG, Bitmap, and PNG file formats. The file thumbnail_idx.db appears to be an index or database file that maintains information about various thumbcache files.

Forensic examiners should be aware of the thumbcache files and the central repository of thumbnail graphics in these files for all folders in Vista. Simple hexadecimal searches will reveal the presence of bitmap, JPEG, and PNG graphics file formats in these files that are directly associated with the thumbnail images in user folders. Based on the default settings for Windows Vista's folder options, examiners can draw simple conclusions as to the presence of a graphic in a given folder if the thumbnail graphic is present within the thumbcache_###.db file(s).

5.3 Windows Mail

Windows Mail replaces Outlook Express in Windows Vista. Outlook Express centralized email into *.dbx files that were found in each user's folders in Windows 2000/XP. While user email is still isolated to each user account (in the user's account folder), the *.dbx file associated with Outlook Express is no longer used with Windows Mail. Instead, Windows Mail creates standard *.eml files for each email sent, received, or in draft format. The folder **\Users\<UserName>\AppData\Local\Microsft\Windows Mail** contains numerous files associated with the user's Windows Mail account and associated email.

The file "account{GUID}.oeaccount" is an xml file that contains the settings for the user's Windows Mail account. The file name contains a globally unique identifier that appears to be specific to the user. The file can be viewed with any browser or application capable of viewing XML data. Notably, the account name, POP server, SMTP server, user name, email address, and display name is stored within this file. The file is found in the directories **\Users\<UserName>\AppData\Local\Microsft\Windows Mail** and **\Users\<UserName>\AppData\Local\Microsft\Windows Mail\Local Folders**.

Attachments are handled in the same manner as Outlook Express's DBX file format. File attachments are stored within the EML file in Base64 format. The Base64 encoding follows the same format as used within the former DBX file format. Prior to the actual Base64 encoded data is a header that describes the file and content. Notably, the header lists the file/content type, file name, encoding type (Base64), and whether the data is an attachment or embedded in the email.

5.4 Paging File

It is well known that the Windows operating systems utilize physical memory (RAM) installed on the computer's motherboard and virtual memory that is otherwise known

as the paging file from Windows NT to the present releases of Windows Vista. The virtual memory file, pagefile.sys, is the virtual memory file used in Windows Vista, and by default, the file is located on the root of the volume where Windows Vista is installed. Pagefile.sys is a dynamic and circular file in that the data swapped into the file by Windows depends on many factors such as physical memory usage, running applications and services, processes, and a number of other programs and objects that Windows may be processing at a given time. In forensic examinations it is common to locate various types of data such as email, chat sessions, graphics, video, spooler files, and numerous other user or system data.

Windows Vista continues to manage the paging file through System Properties applet. Once the user navigates to **Advanced → Performance Options → Advanced → Change (Virtual Memory)**, the specific options for the paging file can be changed. Forensic examiners should review the System registry file located in the **\Windows\System32\Config** directory. Once the registry file is mounted or viewed using AccessData Registry Viewer or MiTec Windows Registry Viewer [6], examiners should review the **\CurrentControlSet###\Control\Session Manager\Memory\Memory Management** subkey for specific settings related to the paging file. The "ExistingPageFiles" value shows the current path and file name of the paging file, pagefile.sys.

5.5 Print Spooling

Similar to Windows 2000/XP, Vista handles printed data by creating a spooler and a shadow file. The spooler file is identified by the extension *.spl while the shadow file by *.shd. The spooler file's primary function is to store the data before it is being printed. The two most common formats for printed data in the spooler file are RAW and EMF. Vista's default print format is EMF. The printed data for EMF format is located forty-one bytes prior to the plain text of "EMF." Typically the start of the EMF data begins with 0x01. An examiner can carve the data out of the file manually and view it with any application (including EnCase and FTK) that is capable of viewing EMF data as a picture or graphic. The shadow file acts as a ticket and identifier for each print job. It contains the name of the printer (or printer driver), printed data format (RAW or EMF), path to the spooler file for each print job, friendly user name of the user who printed the data, SID number of the user who printed the data, application that printed the data, and other notable details about the print job itself. The data within the shadow file is stored in plain text and Unicode. Both the spooler and shadow files are created within the **\Windows\System32\spool\PRINTERS** directory. The file names follow a numerical sequence with the same name for each print job. For example, if the first print job is named FP0000.SPL and FP0000.SHD, the second job will create FP0001.SPL and FP0001.SHD.

Once the print job completes both the spooler and shadow files are deleted. However, since these files tend to be too large to be resident within the MFT, the data will be located in the data area of the volume. Therefore, it is possible to search file slack and unallocated clusters for specific terms related to the shadow file's contents and spoolers file's unique internal signature for EMF data. For example, in EnCase a good GREP expression is "\x01.{40,40}EMF" for EMF data since the string "EMF" appears after 0x01 and 40 arbitrary bytes.

If an examiner encounters a shadow file and spooler file within the **\Windows\System32\spool\PRINTERS** directory with the same file name, it is possible that the print job malfunctioned within Windows or the printer itself. There are a number of possibilities for print job delays or failures, including: Printer out of paper, powered off, disconnected, etc; Windows lost communication with the printer; and the service for printing malfunctioned in Windows.

6 Conclusion

In this paper we described several forensic artifacts of Microsoft's Windows Vista system which are of evidentiary value and may aid forensic examiners in their investigations. Given the proprietary nature of Microsoft operating systems it is often a tedious task to carry out experiments attempting to understand these artifacts and to validate the results. During our research we also discovered other areas of Vista to investigate such as volume shadow copy, new registry locations, ready boot, prefetch file, and more. However, these topics are still being researched, tested, and validated.

The work reported in the paper was based on research using tools such as GuidanceSoftware's EnCase, AccessData's FTK and Registry Viewer, and MiTec Windows Registry Viewer. These tools were used to view raw data and the results were used to validate the tools.

The first author (DMP) noted that Microsoft is making a concerted effort to support the law enforcement community with the release of Windows Vista and other Microsoft products. In October 2006 Microsoft held the Law Enforcement Tech Conference in Redmond, Washington, USA [5]. The event was a huge success and provided law enforcement officers around the globe with a plethora of information relating to computer forensics and investigations. It is therefore our hope that the forensic community of practitioners, digital forensics software vendors, Microsoft and other OS vendors, and academia will continue to support research and development of new tools and techniques to address the new technologies of Windows Vista and beyond.

References

1. AccessData FTK, Registry Viewer, http://www.accessdata.com/
2. Bunting, S., Wei, W.: EnCase Computer Forensics: The Official EnCE: EnCase Certified Examiner Study Guide. Wiley, Chichester (2006)
3. Carrier, B.: File System Forensic Analysis. Addison-Wesley, Reading (2005)
4. GuidanceSoftware EnCase, http://www.guidancesoftware.com/
5. Microsoft LE Tech Conference, cited in the article Legislation, Public Policy, and Enforcement (October 2006) (updated July 15, 2007), at http://www.microsoft.com/mscorp/safety/legislation/default.mspx
6. MiTec Windows Registry Viewer, http://www.mitec.cz/
7. VMWare, http://www.vmware.com/

How Useful Are Tags? — An Empirical Analysis of Collaborative Tagging for Web Page Recommendation

Daniel Zeng[1,2] and Huiqian Li[1]

[1] The Key Laboratory of Complex Systems and Intelligence Science, Institute of Automation, Chinese Academy of Sciences, China
[2] Department of Management Information Systems, The University of Arizona, USA

Abstract. As a representative Web 2.0 application, collaborative tagging has been widely adopted and inspires significant interest from academies. Roughly, two lines of research have been pursued: (a) studying the structure of tags, and (b) using tag to promote Web search. However, both of them remain preliminary. Research reported in this paper is aimed at addressing some of these research gaps. First, we apply complex network theory to analyze various structural properties of collaborative tagging activities to gain a detailed understanding of user tagging behavior and also try to capture the mechanism that can help explain such tagging behavior. Second, we conduct a preliminary computational study to utilize tagging information to help improve the quality of Web page recommendation. The results indicate that under the user-based recommendation framework, tags can be fruitfully exploited as they facilitate better user similarity calculation and help reduce sparsity related to past user-Web page interactions.

1 Introduction

Web 2.0 technology is changing the way Web is being used. As a representative Web 2.0 application, collaborative tagging, a Web-based service that allows users to provide short-phrase descriptions or classifications on Web resources such as Web pages, news, blogs, and photos, has been widely adopted. Del.icio.us is one of the most popular collaborative tagging systems, providing easy-to-use interfaces for users to publish and share their favorite bookmarks along with tags. Many general-purpose Web sites such as flickr, MyBlogLog, Youtube, and Amazon, have added user-generated tags to promote their products and services, faciliate user information search, and encourage community building. Collaborative tagging is becoming popular in non-English Websites as well. In China, for instance, some popular collaborative tagging sites (e.g., 365key.com) are starting to draw attention in the user community.

From a research perspective, the literature on collaborative tagging is rapidly expanding. Roughly, two lines of research have been pursued: (a) studying the structure of tags, and (b) using tag to promote Web search. Golder [3] summarized the patterns of user Web search activities and tagging frequencies, and

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 320–330, 2008.

pointed out the possibility of using del.icio.us as a recommendation system. Halpin [4] et. al. used the generating function method to model the tri-partite graph of tags, users, and links, and identified the power-law property of tag distribution. Ames and Naaman [5] aimed to answer the question why the users contribute and share tags, and drew an interesting conclusion that people tag partially to identify themselves with social groups. Brooks and Nancy [6] found tags are useful in grouping articles, but less effective in providing concrete clues about the actual content of the article.

Several research prototypes involving collaborative tagging have been developed. Yanbe [7] et. al. extended the PageRank algorithm to consider social tagging in order to investigate whether social tagging could promote web search. Li [8] and Bao et. al. [10] proposed a new recommendation algorithm to help user Web search in the collaborative tagging context and implemented this algorithm in a research testbed. Wu [9] et. al. integrated tagging information from del.icio.us in a semantic Web-based system to achieve better user experience.

Despite recent significant interest, research on collaborative tagging remains preliminary. One major problem with the current literature is that the analysis of tagging information tends to be superficial and remains at a relatively high level. As such, it is unclear how social tagging could help Web applications (e.g., Web search) in concrete ways and how computational approaches can be developed to mine and take full advantage of these tags.

Research reported in this paper is aimed at addressing some of these research gaps. The contribution of our work is two-fold. First, we apply complex network theory to analyze various structural properties of collaborative tagging activities to gain a detailed understanding of user tagging behavior and also try to capture the mechanism that can help explain such tagging behavior. Second, we conduct a preliminary computational study to utilize tagging information to help improve the quality of Web page recommendation. One critical research question we address is how useful social tags are through quantifying the extent to which tagging information can improve recommendation quality with a regular collaborative filtering method (without tagging information) as a baseline.

Our research is based on a tagging dataset we have been collecting from del.icio.us since June 2007. In our study, we construct a tri-partite graph and analyze various properties (e.g., degree distribution and cluster coefficient) associated with this graph. Based on the insights gained through this empirical analysis, we have developed several recommendation algorithms that take social tags into consideration. We demonstrate in a quantitative manner that tags can help better cluster users but do not add much to Web page classification. Furthermore, we explore empirically various dimensions of the utility of tags.

In Section 2, we summarize the data collection procedure and the key statistics of our del.cio.us dataset. Section 3 presents a detailed complex network-based analysis of the del.cio.us dataset. In Section 4, we report our social tagging-based algorithms for Website recommendation and their performance relative to the standard collaborative filtering-based methods that use user browsing history

alone. Section 6 concludes the paper with a summary of our research findings and future research.

2 Tagging Datasets

We have been collecting data from various collaborative tagging sites for more than a year. Research reported in this paper makes use of tagging data collected in June and July 2007, from the most popular social tagging Web site, del.icio.us.

As none of the collaborative tagging Web sites offers access to their internal tag databases, we have written specialized Web crawlers to collect information from these web sites for research purposes. These crawlers followed a breadth-first strategy. First, they are provided with certain keywords which correspond to some popular tags. The crawlers fetch the content of the web pages corresponding to these tags and retrieve the links and the titles of the pages posted with these tags. When the program follows the link of the post history of the posted pages and retrieve all the post history including the information of which user posted the page to what tags at what time. The iteration goes on until all the pages under these tags are all visited.

In our first experiment, we provided "Web2.0" as the initial seed tag. In our second experiment, we provided "music" and "finance" as the initial seed tags. Table 1 summarizes the data items used for this research.

Table 1. Summary of Tagging Data Sets Collected in June-July, 2007

Topic	Del.icio.us Data
Web2.0	• total # of users: 327,192 • total # of tags: 123,589 • total # of URLs: 4692 • total # of tuples: 6,826,755
Music	• total # of users: 192,028 • total # of tags: 43,108 • total # of URLs: 4698 • total # of tuples: 1,402,710
Finance	• total # of users: 129,080 • total # of tags: 30,626 • total # of URLs: 3603 • total # of tuples: 766,168

3 Empirical Analysis Based on Complex Network Theory

In collaborative tagging, there are three types of entities: users, items (URLs in the case of del.icio.us), and tags. We can construct intuitively a tri-partite graph where each of the three sets of nodes represents one of the entities, respectively. Existing collaborative tagging work has applied complex network theory to analyze the degree distribution of the tag-user graph [4]. However, the authors ignore the linkage between tags and URLs. We argue that such linkage information could reveal important patterns underlying user tagging behavior and information needs, and could lead to actionable information helping improve user Web search experience.

For data analysis purposes, we have projected the tri-partite graph to three bi-partite graphs. As shown in Fig. 1, we take every pair of entities of different types and the transaction/linkage between them to form these bi-partite graphs.

Computational experiments have been conducted based on all datasets summarizes in Table 1. As similar patterns have been observed across these datasets, in the discussions below, we focus on the dataset with "Music" as the seed topic. This dataset covers 192,028 users and 1,402,710 tagging activities.

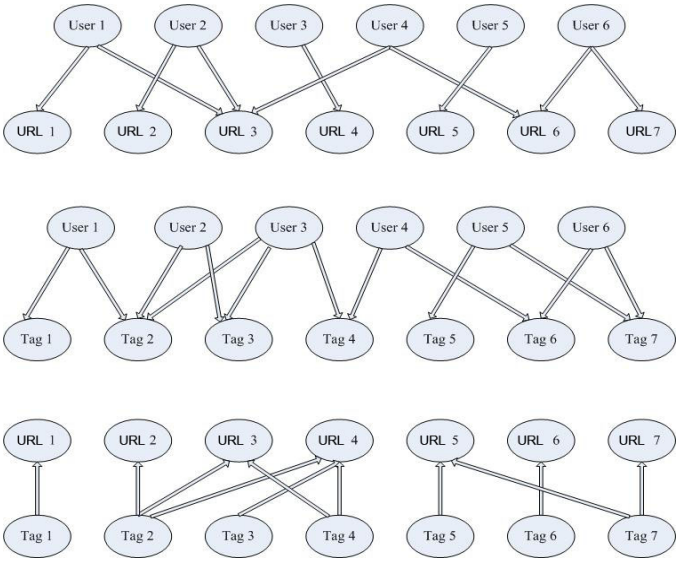


Fig. 1. Instance of bi-party graph

3.1 Degree Distribution

We have studied three bi-partite graphs constructed based on the "Music" dataset. All together, there are six degree distributions of interest:

- Degree distribution for pages in the page-tag graph
- Degree distribution for pages in the page-user graph

- Degree distribution for tags in the page-tag graph
- Degree distribution for tags in the tag-user graph
- Degree distribution for users in the page-user graph
- Degree distribution for users in the tag-user graph

The results are shown in Fig. 2. We observe that the user degree distributions deviate from the power law but not significantly. All remaining distributions seem to follow the power law pretty closely.

3.2 Cluster Coefficient

We also study cluster coefficients (CC) between URLs/pages and tags that are neighbors of a specific user. Here by a neighbor we mean an entity (either a URL or a tag) between which and the user there exists a transaction. More specifically, user i's CC is defined by the following equation

$$CC_i = \frac{T_i}{M*N}$$

where M denotes the number of the neighboring URLs and N the number of the neighboring tags, and T_i denotes the number of actual transactions between these pages and tags.

The CC for the entire graph is the average value among all users' CCs. For our del.icio.us Music dataset, with 192,028 users, the CC is 0.8744, which is very high. We have also calculated the CC among top 500 users and the result is 0.3081, which is also a relative high score. These results indicate that a URL can count for roughly 30% of a random user's interest, including those most active ones. The average number of URLs tagged by top 500 users is close to 21.

Our empirical analysis leads to a hypothesis stating that although a user might have tagged a number of Web resources, these resources tend to focus on certain areas and these areas do not change often. As such, we believe that tags are a great source of information representing user interests and information needs. This could be used to calculate accurately clusters of users sharing similar interests and in turn lead to better recommendation.

4 Tag-Based Web Page Recommendation

In this section, we investigate several tag-based Web page recommendation approaches and evaluate their performance. In particular, we are interested in providing quantitative performance measures investigating exactly how much contributions tags can make in the context of page recommendation.

4.1 Datasets for Recommendation

For algorithmic evaluation, we have used all three del.icio.us datasets (Web 2.0, Music, Finance). The summary statistics for these datasets are presented in Table 2. For each dataset, roughly 60% of the data points were reserved for training and the remaining 40% for testing.

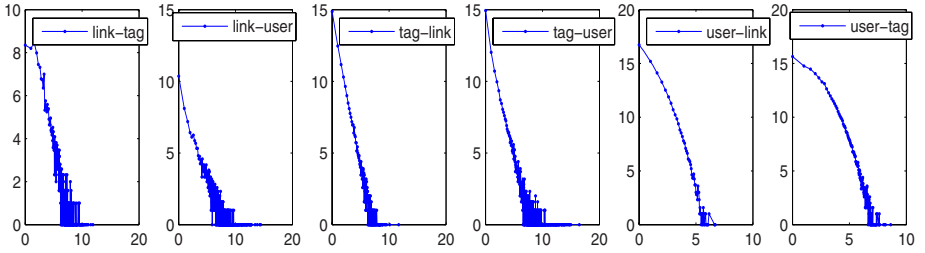


Fig. 2. Degree distribution

Table 2. Summary of Del.icio.us Datasets for Recommendation

	# of users	# of pages	# of tags	# of transactions
del.icio.us-music	423	594	2679	43,646
del.icio.us-finance	386	761	2223	33,841
del.icio.us-web2.0	472	1697	8951	317,791

4.2 Recommendation Algorithms

We have chosen the top-N recommendation approach as the baseline recommendation approach, in which the most heavily tagged pages are recommended to the users [1]. We set $N = 10$ in our experiments. In addition, we experiment with two of the most commonly used successful collaborative filtering algorithms, the user-based neighborhood algorithms [3] and the item-based neighborhood algorithms. Both algorithms are based on a user-item graph represented by a user-term interaction matrix $A = (a_{ij})$. The element a_{ij} takes the value of one if there is an transaction between user i and item j, and zero otherwise.

For a traditional user-based method, one first computes a consumer similarity matrix $WC = (wc_{st})$. The similarity score can be calculated using a vector similarity function based on the corresponding row vectors of A [3]. The high similarity score wc_{st} indicates that consumer s and t have similar preferences because they have previously purchased many common items. The product of $WC \bullet A$ gives potential scores of the items, basing on which the recommendation could be made. For a traditional item-based algorithm, an item similarity matrix $WI = (wi_{st})$ replaces the consumer similarity matrix where the similarity score can be calculated based on the corresponding column vectors of A. The high similarity score wi_{st}indicates that item s and t have similar attributes because they have many common taggers/viewers. The product of $A \bullet WI$ gives potential scores of the item.

In our new recommendation methods that take into consideration of tags, we explicitly consider tagging information as part of user profile and item profile. These profiles are the basis of similarity calculation. More specifically, the profile of a user is a vector recording the frequency of tags ever used by the user;

while the profile of a page is a vector recording the frequency of tags previously attached to the page. In both cases, we use relative frequencies, defined as the actual frequency counts divided by the largest frequency count. As the discriminating power of most popular tags tends to be low, we have applied the inverse word frequency [12] to mitigate this effect. The inverse tag frequency is derived by calculating the logarithm of the ratio between the number of URLs posted with the certain tag and the tag frequency. The final profile vector is the product of the relative frequency and the inverse tag frequency.

To evaluate the performance of our tag-based methods, we recommended 10 pages to each user based on the rankings of the URLs sorted by the user's potential interesting score. The recommended pages were then compared with the real tagging usage in the testing portion of the data to see how well our recommendation method performs. Three widely-accepted performance measures are used in our study: precision, recall and rank score [11]. Precision is simply the ratio between the number of correctly (in terms of real usage) recommended pages and the number of recommended pages. Recall is the ratio between the number of the correctly (in terms of real usage) recommended pages and the number of all pages visited by the users as indicated in the testing data. Rank score evaluates the ranking quality of the ranked list recommendation as specified the following equation:

$$RS_c = \sum_j \frac{q_{cj}}{2^{(j-1)/(h-1)}}$$

where j is the index for the ranked list; h is the viewing halflife which we set to 10 here. The performance of all approaches is based on measures averaged over all users.

4.3 Experimental Results

Table 3 summarizes our experimental results. We observe that the performance of the top-N method is lower than the user-based algorithm in both traditional (without tags) and tag-based contents in almost all experimental conditions with the exception of the Music dataset in which the traditional user-based algorithm performed slightly worse than top-N.

For user-based methods, the tagging-based method improves the precision by more than 10% over the traditional one. In particular, for the Music dataset, the improvement is as high as 24%. Similar trends are observed for recall for all datasets. For rank score, percentage improvement is more than more than 10% for Music and Web 2.0. For Finance, a slight degradation is observed (less than 2%).

Interestingly, for item-based method, we observe opposite patterns. All performance measures, precision, recall, and rank score, suffer with the recommendation approaches making use of tags for all three datasets. In particular, it seems that in general in experimental conditions where tags help more with a user-based method, the same tags managed to impact more negatively the performance of the corresponding item-based method. There is one exception for the

Table 3. Recommendation Performance Result

Data Set	Algorithm	Precision	Recall	Rank Score
del.icio.us: music	top-10	0.0598	8.3*E-4	0.4317
	traditional user-based	0.0582	8*E-4	0.3235
	tagging user-based	0.0723	10*E-4	0.3994
	traditional page-based	0.0612	8.5*E-4	0.3385
	tagging page-based	0.0222	3.1*E-4	0.1386
del.icio.us: finance	top-10	0.0158	1.9*E-4	0.1231
	traditional user-based	0.0215	2.6*E-4	0.1233
	tagging user-based	0.0244	3.0*E-4	0.1210
	traditional page-based	0.0179	2.2*E-4	0.1015
	tagging page-based	0.0168	2.1*E-4	0.1046
del.icio.us: web2.0	top-10	0.0369	1.6*E-4	0.2716
	traditional user-based	0.0496	2.2*E-4	0.2602
	tagging user-based	0.0546	2.4*E-4	0.2877
	traditional page-based	0.0451	2.0*E-4	0.2498
	tagging page-based	0.0346	1.6*E-4	0.1830

Finance dataset, in which rank score improves with the tagging-based method for item-based recommendation. Note, however, for the same dataset, the performance of user-based recommendation decreases with the tagging-based method. This opposite effect of tagging information on user-based and item-based recommendations seems to be consistent and strong. In the next section, we provide some preliminary analysis trying to account for this interesting phenomenon, along with other empirical observations.

5 Empirical Observations and Discussions

In this section, we offer additional insights as to user tagging behavior and the value of user-provided tags based on our empirical analysis of the del.ico.us datasets and related computational studies. This analysis is driven by three questions:

- Does the set of user-provided tags converge over time?
- What is the nature of tags? Do tags provide informative summaries or characterization of the corresponding Web contents? Do tags reflect user interests?
- Why do tags help with user-based recommendation? Why do tags hurt item-based recommendation?

The first question is important in the context of studying social tags as folksonomy [2]. The second question tries to capture the meaning of tags and partially the motivation behind user tagging actions, which certainly impose on user time and cognitive load. The last question is about explaining the observed performance of tag-based approaches and is closely related to the second question. To

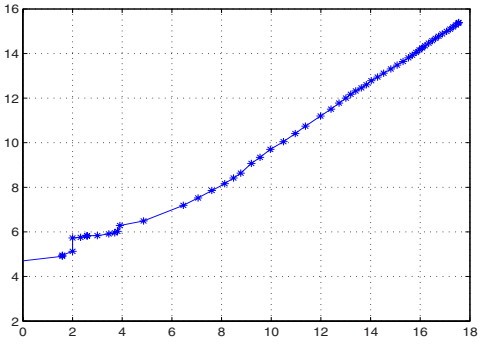


Fig. 3. User and tag number in log-log space

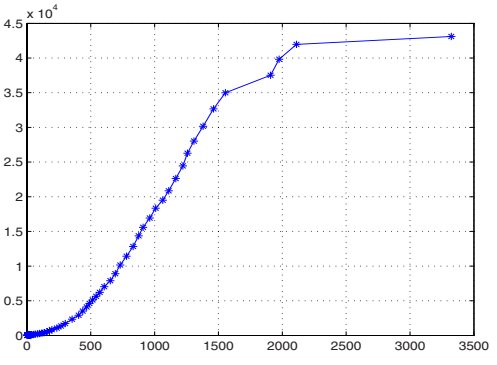


Fig. 4. Link and tag number

answer the first question, we counted the number of users and the number of unique tags for every month in the time period covered by our datasets. We find that the number of unique tags is almost an exponential function of the number of users with a less than 1 exponent. Plotting these two numbers in the log-log scale, as shown in Fig. 3, we observe an almost straight line. This result seems to indicate that it is very likely that the number of tags used will keep increasing as more users join in. We also counted the number of Web pages tagged monthly and plotted it against the number of tags. We find that in the last few months when the number of Web pages increased at a fast pace, the number of tags increased at a slower rate, as shown in Fig 4. We also observe that the number of users did not increase significantly in these months. This indicates that more active users tend to use a more stable set of tags.

To provide insights to the second question, we have examined several random samples from our del.icio.us datasets. Coupled with the evidence from our negative experience with using tags for item-based recommendation, as well as the empirical findings regarding the number of the tags used as an exponential function of the number of users, we suggest that tags do not seem to provide meaningful summaries or characterization of the corresponding Web contents in

a user-independent manner. Rather, tags reflect much more strongly about user information needs and interests.

We believe that there are two major contributors to the observed performance improvement when using tags in user-based recommendation. First, as we discussed above, tags reveal user interests and information needs and thus facilitate better user similarity calculation. Second, relative to the number of users, the density of Web pages is in general less than that of tags as many users attach several tags to a page. When calculating potential scores as the last step of recommendation through multiplying the past interaction matrix with the similarity matrix, the denser user-tag interaction history results in better performance.

6 Conclusion

In this paper, we present our preliminary analysis of data collected from the most popular collaborative tagging site. We conclude that tags tend to reveal user interests and information needs but do not provide much useful information as to the actual contents of Web pages. We have conducted a computational experiment to evaluate in a quantitative manner the actual value of tags for the task of automatic recommendation of Web pages to the users. The results indicate that under the user-based recommendation framework, tags can be fruitfully exploited as they facilitate better user similarity calculation and help reduce sparsity related to past user-Web page interactions.

The algorithmic work as presented in this paper can be extended in multiple directions. Although the proposed naive methods serve the important purpose of evaluating the value of tags in a recommendation context in a comparative setting (using vs. not using tagging information), these methods are far from being the most effective ones. Our current work is focusing on developing and evaluating more sophisticated tag-based recommendation approaches, using the methods discussed in this paper as the benchmarks.

Acknowledgements

This work is supported in part by NNSFC #60621001 and #60573078, MOST #2006AA010106, #2006CB705500 and #2004CB318103, CAS #2F05N01 and #2F07C01, and NSF #IIS-0527563 and #IIS-0428241.

References

1. Huang, Z., Zeng, D.D., Chen, H.: Analyzing Consumer-Product Graphs: Empirical Findings and Applications in Recommender Systems. Management Science 53(7), 1146–1164 (2007)
2. Myung, J., Lee, W., Srivastava, J., Shih, T.K.: Tag-Splitting: Adaptive Collision Arbitration Protocols for RFID Tag Identification. IEEE Transactions on Parallel and Distributed Systems 18(6), 763–775 (2007)

3. Golder, S., Huberman, B.A.: The Structure of Collaborative Tagging System (2005)
4. Halpin, H., Robe, V., Shepherd, H.: The Complex Dynamics of Collaborative Tagging. In: Proceeding of the 16th international conference on World Wide Web, Banff, Alberta, Canada, ACM Press, New York (2007)
5. Ames, M., Naaman, M.: Why we tag: the motivations for annotation in mobile and online media. In: Proceedings of the SIGCHI conference on Human factors in computing systems, San Jose, California, USA, pp. 971–980. ACM Press, New York (2007)
6. Brooks, C.H., Nancy, M.: Improved annotation of the blogosphere via autotagging hierarchical clustering. In: Proceedings of the 15th international conference on World Wide Web, Edinburgh, Scotland, pp. 625–632. ACM Press, New York (2006)
7. Yanbe, Y., Jatowt, A., Nakamura, S., Tanaka, K.: Can social bookmarking enhance search in the web? In: Proceedings of the 2007 conference on Digital Libraries, Vancouver, BC, Canada, pp. 107–116. ACM Press, New York (2007)
8. Li., R., Bao, S., Yong, Y., Fei, B., Su, Z.: Toward effective browing of large scale social annotations. In: Proceedings of the 16th international conference on World Wide Web, Banff, Alberta, Canada, pp. 943–952. ACM Press, New York (2007)
9. Wu, X., Zhang, L., Yu, Y.: Exploring social annotations for the semantic web. In: Proceedings of the 15th international conference on World Wide Web, Edinburgh, Scotland, pp. 417–426. ACM Press, New York (2006)
10. Bao, S., Xue, G., Wu, X., Yu, Y., Fei, B., Shu, Z.: Optimizing web search using social annotation. In: Proceedings of the 16th international conference on World Wide Web, Banff, Alberta, Canada, pp. 501–510. ACM Press, New York (2007)
11. Breese, J.S., Heckerman, D., Kadie, C.: Empirical analysis of predictive algorithms for collaborative filtering. In: Proc. Fourteenth Conf. Uncertainty Artificial Intelligence, Madison, WI, pp. 43–52 (1998)
12. Robertson, S.: Understand inverse document frequency: on theoretical argument of IDF. Journal of Documentation 60(5), 503–520 (2004)

A Generative Model for Statistical Determination of Information Content from Conversation Threads

Yingjie Zhou[1], Malik Magdon-Ismail[2], William A. Wallace[1], and Mark Goldberg[2]

[1] Department of Decision Sciences and Engineering Systems, Rensselaer Polytechnic Institute, Troy, NY 12180
{zhouy5,wallaw}@rpi.edu
[2] Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY 12180
{magdon,goldberg}@cs.rpi.edu

Abstract. We present a generative model for determining the information content of a message without analyzing the message content. Such a tool is useful for automated analysis of the vast contents of online communication which are extensively contaminated by uninformative content, spam, and broadcast. Content analysis is not feasible in such a setting. We propose a purely statistical methodology to determine the information value of a message, which we denote the Information Content Factor (ICF). Underlying our methodology is the definition of information in a message as the message's ability to generate conversation. The generative nature of our model allows us to estimate the ICF of a message without prior information on the participants. We test our approach by applying it to separating spam/broadcast messages from non-spam/non-broadcast. Our algorithms achieve 94% accuracy when tested against a human classifier which analyzed content.

1 Introduction

With ever-increasing Internet accessibility, various electronic media, such as online forums, message boards, blogs, and emails, are available for people to exchange ideas and opinions worldwide. People utilize these tools to communicate with strangers, friends, or experts, to just socialize or to seek help. The volume of such electronic data has increased tremendously. The enormous data can easily overwhelm people interested in analyzing the data for social science purposes [1,2]. Needless to say, the data contain valuable information. For example, the sentiments from a stock message board have been analyzed to show that they could influence the stock market [3,4,5]. On the other hand, huge amounts of spam and noise also exist in the data. Consider a stock analyst observing a stock message board to extract useful tips. It is not feasible to analyze every post given that there is much spam. How should the analyst determine which posts stand a good chance of being "interesting"? Removing the spam from the data

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 331–342, 2008.

set is as important as identifying the important messages. When studying interactions between people (e.g., social group dynamics) by looking at senders and recipients of messages, the spam should be removed since it does not represent interactions. The existence of a significant number of spam and broadcasts will distort the communication pattern that forms the basis for Social Network Analysis. The task of distinguishing useful information from spam among millions of messages is difficult [5]. A straightforward method to separate the informative messages from uninformative ones is to examine the content of the messages; however, for large data sets this approach is not practical.

We propose a generative model to determine the information value of a message, which we call the Information Content Factor (ICF). Our approach does not examine message content. We take as input, a set of conversation threads which have been preprocessed from the raw digital data. A conversation thread is defined as a collection of messages in response to a message. The message, which initiates the conversation, is called the root message. The parent-child relationship between messages is determined by the reply function. All replies to a message are children of that message, and a message is the parent of its replies. Thus, a root message generates a tree of replies (the thread). The depth of the thread is the depth of the tree. The total number of replies to the root message is the summation of messages at each generation summed over all generations. The general intuition behind our generative model is that the more replies, the larger the ICF of the root message is. We propose the ICF, which ranges from 0 to 1, to measure how informative a message is based upon the reply structure to that message. The ICF can be used to separate the informative messages from uninformative messages without examining the content. We apply this methodology to identify broadcasts in the Enron email data set, and we test against a human who has access to the content. Our approach gives a 94% success rate, treating the human as ground truth.

The outline of the paper is as follows. In Sect. 2, two essential elements of our generative model are described, then three reply processes and their expected number of replies are presented. The statistical method to determine the ICF of the root message is discussed in the last part of this section. In Sect. 3, we apply the method under the framework of our generative model to Enron emails to identify broadcast messages. We conclude with suggestions for future research.

2 Generative Statistical Model

We assume a message with ICF 1 will be replied with probability 1 by each of the recipients of that message, and a message with ICF 0 has no replies. More generally, the ICF is related to the probability of obtaining a reply. There are two elements in our generative model. The first determines the probability that a message is replied given its ICF b. The second determines how the ICF of a reply is related to the ICF of the parent message. Intuitively the higher b, the more likely a reply, and the ICF of a reply should be smaller than the ICF of the parent. Let p^* denote the probability of being replied when ICF is 1, then by our

definition $p^* = 1$. For any root message with ICF b, whose probability of being replied is denoted as p, we assume p is proportional to p^*, that is, $p = bp^* = b$. To capture the decay in ICF from parent message to child message, we define the ICF-propagation function $f(b)$, $0 < f(b) < 1$. Thus, for a message with ICF b, p[reply] $= b$, and ICF[reply] $= bf(b)$. That is, if the ICF of the root message is b, the probability of a reply occurs is b, and its ICF is $bf(b)$. If we let b_i denote the ICF at depth i, it is a function of b as follows:

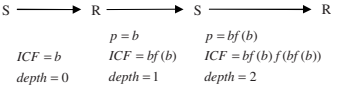
$$b_i = \begin{cases} b & \text{if } i = 0 \\ b_{i-1} f(b_{i-1}) & \text{if } i > 0 \end{cases} \tag{1}$$

The probability of a message at depth $i \geq 1$, p_i, is given as:

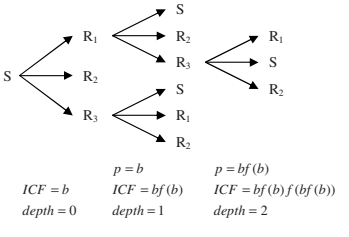
$$p_i = b_{i-1} \tag{2}$$

One interesting special case is $f(b) = f$, where f is a constant decay factor. The ICF of a message at depth i will be bf^i.

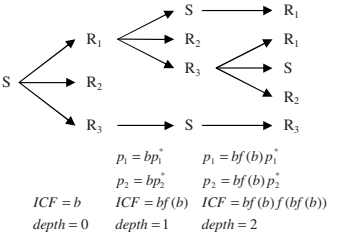
Assume a sender S initiates a message M, two cases in terms of number of recipients may happen: the message M has one recipient; and, the message M has multiple recipients. Let R denote the recipient set. In the case of multiple recipients, $R = \{R_1, R_2, ..., R_i, ..., R_n\}$, where $n \geq 2$ is the number of recipients. The generative model is applied to the three reply processes, namely, single recipient, multiple recipients, and mixed reply process. To better understand their characteristics, Fig. 1 illustrates how they work with one example for each reply process. The detail will be described in Sect. 2.1, 2.2, and 2.3 respectively. The idea behind the generative model will become clear from the three reply processes. The specific details are however application dependent and it should be possible to extend our framework to accommodate different domains.



(a) Single recipient reply process



(b) Multiple recipients reply process



(c) Mixed reply process

Fig. 1. Three examples for three reply processes

2.1 Single Recipient Reply Process

In this reply process, the sender S initiates the root message to the recipient R, and R may take 2 actions: reply to S or don't reply. If R chooses to reply, S again has two options, reply to R or don't reply, and so on. The conversation between S and R continues until one of them fails to reply. An example of such a conversation between S and R is given in Fig. 1(a). The ICF, depth, probability of each message are indicated in the figure. For example, when S initiates a message at depth 0, the probability p that R replies is b. If R replies, the ICF of this replied message is $bf(b)$, and its depth is 1 in the reply process, and so on.

This reply process is recursive with decreasing ICF. The recursion shows that the expected number of replies of the parent message is a function of the expected number of replies of the child message. Let X denote the total number of replied messages to the root message with one recipient. Let $\mathcal{E}(b) = E[X|b]$ be the expected number of replies to the root message with ICF b. We derive $\mathcal{E}(b)$ recursively as

$$\mathcal{E}(b) = b\,(1 + \mathcal{E}(bf(b))) \tag{3}$$

The first term is the expected number of replies to the root message, and the second recursive term captures the expected number of messages for the single recipient reply process initiated by the first reply. It turns out that it is hard to solve (3) analytically. We give an approximate recursion to calculate $\mathcal{E}(b)$. First we approximate $\mathcal{E}(b)$ when b is small using a Taylor series expansion to second order, and then use (3) to calculate $\mathcal{E}(b)$ recursively. The Taylor series expansion of $\mathcal{E}(b)$ at $b = 0$ is given by

$$\mathcal{E}(b) = \mathcal{E}(0) + \mathcal{E}'(0)b + \frac{\mathcal{E}''(0)b^2}{2} + \dots \tag{4}$$

Since b is small, we ignore the orders higher than 2. When $b = 0$, the probability of reply for each recipient is 0, therefore, $\mathcal{E}(0) = 0$. To find $\mathcal{E}'(0)$ and $\mathcal{E}''(0)$, the first and second derivative of $\mathcal{E}(b)$, $\mathcal{E}'(b)$ and $\mathcal{E}''(b)$, are obtained first.

$$\mathcal{E}'(b) = 1 + \mathcal{E}(bf(b)) + b(f(b) + bf'(b))\mathcal{E}'(bf(b)) \tag{5}$$

$$\mathcal{E}''(b) = 2f(b)\mathcal{E}'(bf(b)) + b\,(4f'(b) + bf''(b))\,\mathcal{E}'(bf(b)) + b\,(f(b) + bf'(b))^2\,\mathcal{E}''(bf(b)) \tag{6}$$

From (5) and (6), we have $\mathcal{E}'(0) = 1$ and $\mathcal{E}''(0) = 2f(0)$. Thus, $\mathcal{E}(b)$ can be calculated numerically as in Algorithm 1. The expected number of replies

Algorithm 1. Numerical analysis of $\mathcal{E}(b)$

```
if b ≤ 10^−5 then
    E(b) ← b + b^2 f(0)
else
    E(b) ← b (1 + E(bf(b)))
end if
```

to any of the messages in the stream can be obtained by replacing b with the corresponding ICF for that message.

2.2 Multiple Recipients Reply Process

In this reply process, the sender S initiates the root message to the recipient set $R = \{R_1, R_2, ..., R_i, ..., R_n\}$, where $n \geq 2$. R_i may take 2 actions: reply to the sender and the other recipients or not to reply. We assume that a recipient chooses to reply or not independently of the other recipients. The conversation among $\{S\} \bigcup R$ along a particular message path dies when a recipient fails to reply. The conversation ends when every message path dies. An example of such a conversation between S and $R = \{R_1, R_2, R_3\}$ is given in Fig. 1(b). The ICF, depth, probability of each message are indicated in the figure. For example, when S initiates a message to R at $depth = 0$, the probability p that each of R_1, R_2, and R_3 replies is b. If R_i replies, the ICF of this replied message is $bf(b)$, and its depth is 1, and so on.

This reply process is also recursive with decreasing ICF. Let Y denote the total number of reply messages to the root message. Let $\mathcal{F}(n, b) = E[Y|n, b]$ denote the expected number of replies to the root message with ICF b and number of recipients n. We derive $\mathcal{F}(n, b)$ recursively as

$$\mathcal{F}(n, b) = nb\,(1 + \mathcal{F}(n, bf(b))) \tag{7}$$

The logic behind this expression is that S starts n independent threads of the same form (note the factor n). For each thread, the expected number of messages is $1 + \mathcal{F}(n, bf(b))$ with probability b because R_i starts exactly the same process with lower ICF $bf(b)$. Following the same procedures in Sect. 2.1, we can obtain $\mathcal{F}(n, 0) = 0$, $\mathcal{F}'(n, 0) = n$ and $\mathcal{F}''(n, 0) = 2n^2 f(0)$. When b is small, $\mathcal{F}(n, b)$ can be approximated by Taylor series to second order. Thus, $\mathcal{F}(n, b)$ can be calculated numerically as in Algorithm 2.

Algorithm 2. Numerical analysis of $\mathcal{F}(n, b)$

if $b \leq 10^{-5}$ **then**
  $\mathcal{F}(n, b) \leftarrow nb + n^2 b^2 f(0)$
else
  $\mathcal{F}(n, b) \leftarrow nb\,(1 + \mathcal{F}(n, bf(b)))$
end if

2.3 Mixed Reply Process

The mixed reply process is a mixture of the single recipient and multiple recipient reply processes. In the mixed reply process, sender S initiates the root message to the recipient set $R = \{R_1, R_2, ..., R_i, ..., R_n\}$, where $n \geq 2$. R_i may take 3 actions: reply to the sender only ("Reply Sender"), reply to the sender and all the other recipients ("Reply All"), or not to reply. We assume that each

recipient acts independently. We denote the probability of reply to the sender only as $p_1 = bp_1^*$, and the probability of reply to the sender and the other recipients as $p_2 = bp_2^*$, where p_1^* and p_2^* denote the probability of reply using the "Reply Sender" and "Reply All" options respectively when the ICF is 1. Note that $p_1^* + p_2^* = 1$ because the probability of reply is assumed to be 1 when ICF is 1. The conversation among $\{S\} \bigcup R$ along a particular message path dies when a recipient fails to reply. The conversation ends when every message path dies.

An example of such a conversation between S and $R = \{R_1, R_2, R_3\}$ is given in Fig. 1(c). In this particular reply process, at depth 1 R_1 chooses "Reply All", R_2 chooses "No Reply", and R_3 chooses "Reply Sender". The ICF, depth, probability of each message are indicated in the figure. For example, when S initiates a message to R_1, R_2, and R_3 at $depth = 0$, the probability of replying to the sender only, $p_1 = bp_1^*$, and the probability of replying to the sender and the recipients, $p_2 = bp_2^*$, and the probability of no action is $1 - p_1 - p_2$, which is $1 - b$. The ICF of this replied message is $bf(b)$, and its depth is 1 in the reply process, and so on. What we should notice is that once the "Reply Sender" option is chosen, the reply process followed will be the single recipient reply process.

Let $R_i' = \{R_1, R_2, ..., R_{i-1}, R_{i+1}, ..., R_n\}$ denote $R \setminus \{R_i\}$. We assume that when a message is replied, two options, "Reply Sender" and "Reply All", are to be used, which correspond to "S" and $\{S\} \bigcup R_i'$ respectively as the recipient(s) in the reply message of R_i. Let p denote the probability of reply, $p = p_1 + p_2$, in which $p_1 = bp_1^*$ is the probability of reply using the "Reply Sender" option, and $p_2 = bp_2^*$ is the probability of reply using the "Reply All" option.

Of the three actions, "No Reply", "Reply Sender", "Reply All", the last two actions may result in more replied messages. Take recipient R_3 in Fig. 1(c) as an example, at depth 1 R_3 chooses "Reply Sender", the reply process followed is the single recipient reply process; the probability of reply to a parent message is the summation of probability of "Reply Sender" and "Reply All", i.e, $p = bp_1^* + bp_2^* = b(p_1^* + p_2^*) = b$. On the other hand, R_1 chooses "Reply All", the reply process followed is the recursive process with a lower ICF.

Let Z denote number of replies of this process, and X denote number of replies when a "Reply Sender" option is selected when multiple recipients exist. Let $\mathcal{G}(n, b) = E[Z|n, b]$ be the expected number of replies to the root message. Since the "Reply All" option leads to a recursive reply process with a lower ICF, $\mathcal{G}(n, b)$ is given recursively as

$$\mathcal{G}(n, b) = n(bp_1^*(1 + \mathcal{E}(bf(b))) + bp_2^*(1 + \mathcal{G}(n, bf(b)))) \tag{8}$$

The logic behind this expression is that S starts n independent threads. The term $bp_1^*(1 + \mathcal{E}(bf(b)))$ captures the expected number of messages for the single recipient reply process initiated by the "Reply Sender" option at depth 1; the term $bp_2^*(1 + \mathcal{G}(n, bf(b)))$ captures the expected number of messages for the mixed reply process initiated by the "Reply All" option at depth 1 because the same process with lower ICF $bf(b)$ is started. Since $\mathcal{E}(b) = b\,(1 + \mathcal{E}(bf(b)))$ from (3), $\mathcal{G}(n, b)$ can be written as:

$$\mathcal{G}(n, b) = np_1^*\mathcal{E}(b) + nbp_2^* + nbp_2^*\mathcal{G}(n, bf(b)) \tag{9}$$

Again, we approximate $\mathcal{G}(n, b)$ using a second order Taylor expansion. We can obtain $\mathcal{G}(n, 0) = 0$, $\mathcal{G}'(n, 0) = n$ and $\mathcal{G}''(n, 0) = 2np_1^* f(0) + 2n^2 p_2^* f(0)$. The calculation of $\mathcal{G}(n, b)$ is shown in Algorithm 3. The expected number of replies for any message in the thread can be obtained by replacing b with the corresponding ICF of the message.

Algorithm 3. Numerical analysis of $\mathcal{G}(n, b)$

if $b \leq 10^{-5}$ **then**
  $\mathcal{G}(n, b) \leftarrow nb + nb^2 p_1^* f(0) + n^2 b^2 p_2^* f(0)$
else
  $\mathcal{G}(n, b) \leftarrow np_1^* \mathcal{E}(b) + nbp_2^* + nbp_2^* \mathcal{G}(n, bf(b))$
end if

2.4 Statistical Determination of ICF

Given a thread generated from root message M, we would like to determine the ICF b of message M. We select b to match the observed tree. The approach we propose here is to select b to match the expected number of messages to the observed number of messages. This can be done at every level of the tree, treating each node as the root of its subtree-thread. The ICF of this subtree-root is determined from b and the ICF propagation function.

For a given root message with ICF b, assume the depth of the reply process is m, and there are n_i messages at each depth i. Let x_{ij} denote the total number of observed replies to the j^{th} message at depth i, b_i denote the ICF of messages at depth i, and $E[X_{ij}|n_{ij}, b_i]$ denote the expected number of replies to this message. We select b to minimize the summation of the squared difference between expected and observed number of replies for every message in the reply process. Thus, we define the error function

$$\Sigma(b) = \sum_{i=0}^{m} \sum_{j=1}^{n_i} \left(x_{ij} - E[X_{ij}|n_{ij}, b_i]\right)^2, \tag{10}$$

where b_i is defined in (1). The ICF b is then selected as argmin $\Sigma(b)$.

3 Detecting Broadcasts in Enron Emails

The methodology is applied to a subset of Enron emails to detect broadcasts. A broadcast is defined as an email which is sent to multiple recipients, but the conversation triggered by this email dies down quickly. The purpose of detecting broadcasts is to eliminate the emails that inspire little or no interaction between sender and recipients and hence are misleading for Social Network Analysis. In this paper, only those root messages with 5 or more recipients are tested.

3.1 A Brief Description of Enron Email Data

Enron Corporation was founded in 1985. It became the seventh largest business organization in the USA in fifteen years [6,7]. Enron's stock price was as high as $90 in August of 2000, however, Enron declared bankruptcy in December 2001 without any warning [6, 7]. After Enron's bankruptcy numerous investigations were conducted by authorities. Many employees' emails were also collected and released by Federal Energy Regulatory Commission (FERC) to the public for investigation [8].

The data set we are testing on is extracted from the March 2, 2004 Version of Enron emails posted by Cohen [9]. This corpus contains 517,431 messages dated from November 1998 to June 2002 organized into 150 employee folders. The researchers identified 156 employees from this data set, and most of them were senior managers of Enron [10]. Because the communication among these 156 employees was our interest, 22,099 emails among these 156 employees were extracted from the March 2, 2004 Version. The conversation threads are constructed and tested by our methodology.

3.2 Constructing Email Threads

Methods for threading emails into conversations have been discussed in previous research [11, 12, 13]. Although it is argued that language processing should be used to thread electronic messages [11, 12], we adopt a simpler but efficient method.

In an email system, usually two options, "Reply Sender" and "Reply All", are available for replying a message. Note that in most email software the "Reply Sender" option is symbolized as a "Reply" button. We ignore the slight possibility that neither of them is used in replying. We assume that when one of these two options is used to reply a message, the subject will not be changed except a "Re:" may be added. We examine the "Subject", "From", "To", "Cc", and "Date" headers to construct the parent-child relationship. If the "Subject" header of a message contains "Re:", we consider it as a child message. To find its parent message, we compare header fields of two messages. If the "Reply Sender" option is used, the recipient of the replied message will be the sender of the parent message; and if the "Reply All" option is used, the recipient of the replied message will be the sender and the other recipients of the parent message. For both options, the sender of the child message should be one of the recipients of the parent message. The "Date" field will be used as a time constraint to determine the parent-child relationship for any two messages, since the response time should not be long. We use 96 hours as the response time window.

3.3 Experiments

In this experiment, we assume the ICF propagation function $f(b) = f$ is a constant ranging from 0 to 1. Nine settings, $f \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ are tested. The probabilities of "Reply Sender" and "Reply All" when

ICF is 1 (p_1^* and p_2^*) are approximated by their relative frequencies of having been used. It turns out p_1^* and p_2^* don't differ significantly, therefore, $p_1^* = p_2^* = 0.5$ is used in this experiment. We randomly select 50 threads as a training set, and another 50 as a testing set. For each thread in the test and training sets, we read the content of the email to determine if it is a broadcast message; if a message is to inform of a decision, a result, news, a meeting time, or anything that doesn't require a reply, we categorize it as a broadcast, otherwise it is considered as a normal message. The ICF of each thread at each f setting is calculated by minimizing (10). A threshold is then chosen to determine if a thread is a broadcast, i.e., a thread is a broadcast if the ICF of the thread is not larger than the threshold, and it is a normal message otherwise.

Let C and H be the variables indicating if a message is a broadcast from the statistical method and the content of the message respectively. They can be either 0 or 1, in which 0 represents normal message, and 1 represents broadcast. Let T denote the threshold. For any message i, C_i is defined as:

$$\mathrm{C}_i(T) = \begin{cases} 0 \text{ if } ICF_i > T \\ 1 \text{ otherwise} \end{cases} \tag{11}$$

The error is defined as the cumulative absolute difference squared between C_i and H_i in (12). The threshold T^* is determined as argmin $\Sigma(T)$.

$$\Sigma(T) = \sum_{i=0}^{50} (C_i(T) - H_i)^2 \tag{12}$$

3.4 Results

The result shows that the magnitude of f doesn't effect the error greatly for a given threshold. Figure 2 shows the nine f values with their corresponding optimal threshold T^* and error $\Sigma(T^*)$. When $f \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$, the error is 2; however, when $f = 0.9$, the error is 3. A linear relationship between f and T is clear from Fig. 2. The regression analysis shows the adjusted $R^2 = 99.4\%$ with slope -0.1596 and intercept 0.2910. Therefore, $T = 0.2910 - 0.1596 * f$ can be used to estimate the threshold for a given f value. However, we notice that the error when $f = 0.9$ is larger than the error when f takes the other eight values. We further investigated the effect of f on the ICF with an example.

In this example four threads are illustrated in Fig. 3(a). All of their root messages have 5 recipients, but the generated threads are different. Thread (a) shows one of the five recipients replies to the sender; thread (b) has two of the five recipients reply, one chooses "Reply Sender" and the other chooses "Reply All"; thread (c) shows one of the recipients replies to the sender, and the sender follows up with another message; thread (d) shows one of the recipients replies to the sender and the other recipients, and one of them follows up. The ICFs for thread (a), (b), (c), and (d) are shown in Fig. 3(b) for $f \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$. As expected, the ICF decreases when f increases, and the rate of decrease is almost a constant for each thread. However, the decreasing rate varies from thread

f	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
T*	0.28	0.26	0.24	0.22	0.21	0.20	0.18	0.17	0.15
Σ	2	2	2	2	2	2	2	2	3

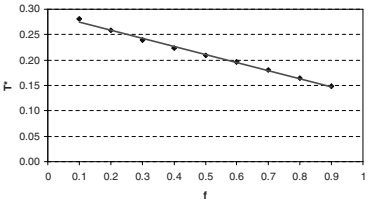
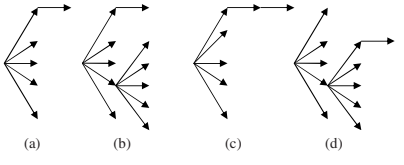
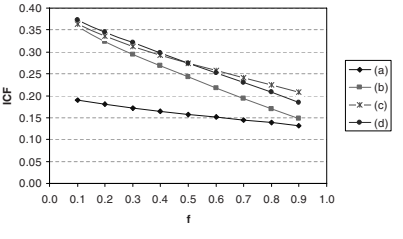


Fig. 2. f and its associated threshold T

to thread. For instance, thread (a) has a flatter slope compared with (b), (c), and (d). As a result, the ICF of (a) is very close to that of (b) at $f = 0.9$ although they are quite apart from each other at $f = 0.1$ because thread (b) expects more replies when f is large. Thread (b), (c), and (d) cluster when f is small but separate when f is large. Threads (c) and (d) intersect when f is around 0.55 because the ICF of (d) is more sensitive with higher f values. The difference in slopes is identified as the reason that the error changes with f. Compared with two replies in (b), (c), and (d), thread (a) has only one reply; therefore, (a) should be distant from the others in terms of ICF. From Fig. 3(b), we notice that when f is small, (a) is indeed separated from the other three threads, therefore, small f values are recommended.



(a) Four threads as an example



(b) ICF vs. f for four threads

Fig. 3. An example to illustrate the selection of appropriate f

Another reason that we recommend small f values (≤ 0.3) is justified as follows. Consider the thread "A→B→A→B", three emails between A and B. Communication patterns like this are very common in real life. Intuition suggests that the ICF of the root message should be high (≥ 0.9). It shows that if ICF is at least 0.9, f should be no more than 0.3. On the other hand, f cannot be too small since we also need to differentiate the interesting root messages which have triggered heated discussions. Therefore, $f \in [0.1, 0.3]$ is our recommendation.

The Relative Operating Characteristic, or ROC curve, is plotted for $f = 0.1$ in Fig. 4 using the training set data. The X axis is the false positive rate (FPR),

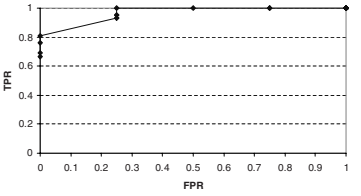


Fig. 4. ROC curve of the training data when $f = 0.1$

		Human	
		B	NB
Our Methodology	B	41	2
	NB	1	6

Fig. 5. Confusion matrix for the test data

which means it is categorized as a normal message from the content but it is classified as a broadcast using our methodology; and the Y axis is true positive rate (TPR), which means the message is categorized as a broadcast from both the content and our methodology. The area under the ROC curve is larger than 90%, which proves that our methodology is very effective (on the training set). We applied the combination of $f = \{0.1, 0.2, 0.3\}$ and its associated threshold $T^* = \{0.28, 0.26, 0.24\}$ to the test data set, which produced 3 disagreement out of 50 threads with accuracy 94%. The confusion matrix is shown in Fig. 5, in which "B" represents Broadcast and "NB" represents non-broadcast. Since the error doesn't deviate much from the error of the training data, our method is believed to be robust.

4 Conclusions

We have developed a statistical method to evaluate how informative a message is by the conversation thread it triggered. This method is then applied to a subset of Enron email data to detect the broadcast messages. We conclude that the threshold to differentiate the broadcast from the normal message is a linear function of information decay factor f, and $f \in [0.1, 0.3]$ is recommended. The method is proved to be effective and robust in detecting broadcast messages by applying it on both the training and testing data. The proposed method, in general, helps to process the data for various analyses and achieve a better understanding of the data. Our future research includes applying the methodology to detecting interesting topics from conversation threads.

Acknowledgment

This material is based upon work partially supported by the U.S. National Science Foundation (NSF) under Grant Nos. IIS-0621303, IIS-0522672, IIS-0324947,

CNS-0323324, NSF IIS-0634875, the U.S. Office of Naval Research (ONR) Contract N00014-06-1-0466, and the U.S. Department of Homeland Security (DHS) through the Center for Dynamic Data Analysis for Homeland Security administered through ONR grant number N00014-07-1-0150 to Rutgers University. The content of this paper does not necessarily reflect the position or policy of the U.S. Government, no official endorsement should be inferred or implied.

References

1. Berghel, H.: Cyberspace 2000: dealing with information overload. Communications of the ACM 40(2), 19–24 (1997)
2. Losee Jr., R.M.: Minimizing information overload: the ranking of electronic messages. Journal of Information Science 15(3), 179–189 (1989)
3. Tumarkin, R., Whitelaw, R.: News or noise? internet message board activity and stock prices. Financial Analysts Journal 57, 41–51 (2001)
4. Antweiler, W., Frank, M.Z.: Is all that talk just noise? the information content of internet stock message boards. Journal of Finance 59(3), 1259–1294 (2004)
5. Gu, B., Konana, P., Liu, A., Rajagopalan, B., Ghosh, J.: Predictive value of stock message board sentiments. In: The Social Science Research Network Electronic Paper Collection, Social Science Electronic Publishing, Inc. (2007)
6. McLean, B., Elkind, P.: Smartest Guys in the Room: The Amazing Rise and Scandalous Fall of Enron, 1st edn. Portfolio (2003)
7. Swartz, M., Watkins, S.: Power Failure: The Inside Story of the Collapse of Enron, 1st edn. Doubleday (2003)
8. Federal Energy Regulatory Commission: Addressing the 2000-2001 western energy crisis, http://www.ferc.gov/industries/electric/indus-act/wec/enron/info-release.asp
9. Cohen, W.W.: Enron email dataset, http://www.cs.cmu.edu/~enron/
10. Zhou, Y., Goldberg, M., Magdon-Ismail, M., Wallace, W.A.: Strategies for cleaning organizational emails with an application to Enron email dataset. In: 5th Conference of NAACSOS, Emory, Atlanta, GA, June 7-9 (2007)
11. Comer, D.E., Peterson, L.L.: Conversation-based mail. ACM Transactions on Computer Systems 4(4), 299–319 (1986)
12. Lewis, D.D., Knowles, K.A.: Threading electronic mail: A preliminary study. Information Processing and Management 33(2), 209–217 (1997)
13. Venolia, G.D., Neustaedter, C.: Understanding sequence and reply relationships within email conversations: a mixed-model visualization. In: CHI 2003: Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 361–368. ACM Press, New York (2003)

Using "Cited by" Information to Find the Context of Research Papers

Chun-Hung Lu[1,2], Chih-Chien Wang[1], Min-Yuh Day[1,2],
Chorng-Shyong Ong[2], and Wen-Lian Hsu[1]

[1] Institute of Information Science, Academia Sinica, Taiwan, R.O.C.
[2] Department of Information Management, National Taiwan University, Taiwan, R.O.C.
{enrico,myday,vincent,hsu}@iis.sinica.edu.tw,
ongcs@im.ntu.edu.tw

Abstract. This paper proposes a novel method of analyzing data to find important information about the context of research papers. The proposed CCTVA (Collecting, Cleaning, Translating, Visualizing, and Analyzing) method helps researchers find the context of papers on topics of interest. Specifically, the method provides visualization information that maps a research topic's evolution and links to other papers based on the results of Google Scholar and CiteSeer. CCTVA provides two types of information: one type shows the paper's title and the author, while the other shows the paper's title and the reference. The goal of CCTVA is enable both novices and experts to gain insight into how a field's topics evolve over time. In addition, by using linkage analysis and visualization, we identify five special phenomena that can help researchers conduct literature reviews.

Keywords: Automatic meta-analysis, Human factors, Link mining.

1 Introduction

Scholars and students face several problems when they start a research project. The first problem is information overload. Compared with the traditional means of reviewing literature, the World Wide Web (WWW) has wrought enormous changes in the way information is provided. As a result, the sheer amount of information can be overwhelming at times. Thus, how to help researchers find information they need has become a critical research issue.

The second problem is information organization. Although search engines are easy to use, they do not provide tools to help filter, interpret, and organize individual items of information. Jacobson & Prusak [1] found that employees of business enterprises spend more than 80% of their time and effort eliciting, interpreting, and applying knowledge, while actually searching for knowledge only occupies 15% of their time. In academic work, information organization is also important in helping researchers find the context of a research topic. Scholar tools, such as Google Scholar provide a great deal of information; however, researchers do not have time to browse documents one by one.

Despite the rapid increase in the volume of conference and journal publications, the WWW makes accessing such documents relatively easy. In this paper, we propose a

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 343–354, 2008.

novel method that integrates information search tools for scholarly publications and identifies the dynamic relationships in a research domain. We focus on systems that map the evolution of a research topic using the condensed result sets from Google Scholar and CiteSeer.

The remainder of this paper is organized as follows. In Section 2, we review several related papers and systems, and discuss our data schema. In Section 3, we define our research questions and explain our methodology. In Section 4, we discuss five phenomena that can help researchers conduct literature reviews. We also detail the results of experiments on the dataset complied from Google Scholar and CiteSeer. Finally, in Section 5, we present our conclusions.

2 Literature Review

Bibliographic management, citation indexing/extraction, and co-authorship analysis are important research topics in many fields. To the best of our knowledge, the first paper to propose the use of citation indexing for historical research was published in 1955 [3]. Since then, many researchers have used 2D graphs to demonstrate the relationship between scholarly papers and the evolution of research fields, as shown by the example in Fig. 1. Citation analysis has also become an important research topic in the field of information extraction. Giles et al. [4] developed a tool called CiteSeer [2], which can parse citations, identify citations of the same paper in different formats, and identify the context of citations in the body of an article. Lin et al. [6] used novel network paths to find information of interest. For example, they used bibliographic citation data from the Open Task of the 2003 KDD Cup to analyze and answer questions like: "Which people are connected to C.N. Pope?"

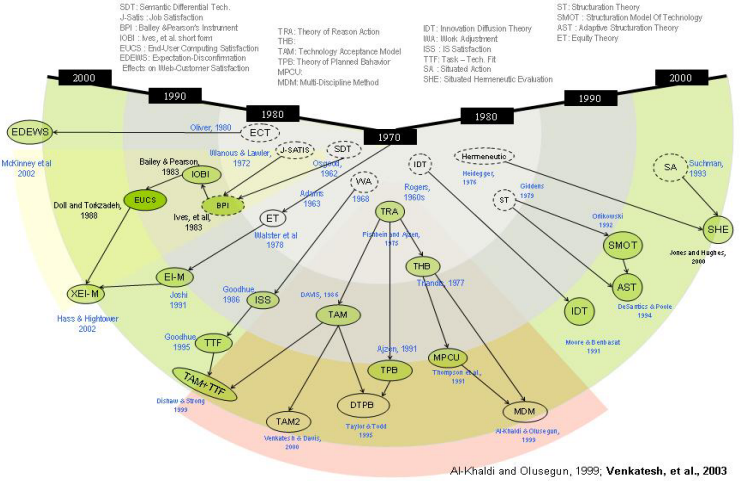


Fig. 1. Historical overview of Information System Evaluation (ISE) research streams

Table 1. Comparison of information provided by Google Scholar & Citeseer

	Citeseer	Google Scholar
Domain	Focus on computer science	General (Web, Digital Library, Publisher)
Items	Title name Paper Link Author name Journal/conference name Abstract Cited by Similar documents Active bibliography Similar documents based on text Related documents from co-citations BibTeX entry Citations Years of citations(Graph)	Title name Paper Link Author name Journal/conference name Snippet Cited by Links to British Library Direct, where the paper can be purchased Links to Web search Related articles
Characteristics of Information Provided	Provides rich historical information	Provides information about what happened after a paper was published
Constraint	N/A	Constrained at 1,000 entries

Smeaton et al. [7] analyzed the content of papers published in SIGIR proceedings to identify research trends. Their objective was to determine the topics that appear most frequently; however, they did not visualize the results or include any citation analysis. Nascimento et al. [8] constructed a co-authorship graph of all the papers published by SIGMOD between 1975 and 2002. Lee et al. [9] developed a visualization tool called PaperLens, which allows researchers to identify the trends and topics in a field.

Fortuna et al. [10] found that extracting the main concepts from documents using techniques like the Latent Semantic Index (LSI) yields more useful visualization results. For example, given a set of descriptions of European Research projects (6FP), one can find the main areas that these projects cover, e.g., the Semantic Web, e-learning, and information security.

In this section, we discuss the use of scholar tools to find information of interest and compare various tools. Scholar tools are programs that enable users to search the Web for articles in academic journals and databases. The following are some of the most popular tools: Google Scholar [1], OAIster [11], Windows Live Academic Search [12], Citeseer [2], and DBLP [13]. There are also several digital libraries, such as the ACM Digital Library [14], IEEE Xplore [15], Arxiv.org [16] and PubMed [17]. These libraries provide various kinds of information, e.g., article names, author names, papers linked to journals and conference proceedings, references, abstracts, and keywords. We conduct a data analysis of "Google Scholar" and CiteSeer, as shown in Fig. 2 and Fig. 3. We compare the information provided by "Google Scholar" and CiteSeer in Table 1. The most widely used features are author, paper

Fig. 2. Data Analysis of Google Scholar

Fig. 3. Data Analysis of CiteSeer

title, and publisher's information. To trace a theory's evolution and data convergence, we determine the relationships between an original paper and papers that have cited it and use the result to form our data schema.

3 Empirical Study and Research Method

We use research publications for our empirical study. Before the advent of the WWW, researchers had to expend a great deal of effort searching for papers to learn about developments in their field. When they started a research topic, they had to depend on one or two journals and review all the papers to find relevant information. In contrast, the WWW enables researchers to access e-journals, personal publication

lists, and other important information with relative ease. However, the availability of large volumes of information has caused an "information explosion", which poses tremendous challenges in terms of the intelligent organization of data and visualization.

3.1 Research Questions

We use Teece at al.'s "Dynamic Capabilities" theory [18] as the basis of our empirical investigation. As with any field of study, we need to address the following research questions:

RQ1: "How can we determine which papers we should review?"
RQ2: "How can we discover the evolution of a specific domain?"
RQ3: "Does the "cited by" number mean a highly cited paper is very important?"
RQ4: "Is there a tool that can locate the information we require?"

3.2 The Proposed Method - CCTVA

To address the above questions, we have developed a method called CCTVA (Collecting, Cleaning, Translating, Visualizing, and Analyzing), which is comprised of the following five processes.

Collecting: To obtain "cited by" information, we used two scholar search tools, Google Scholar and Citeseer. We developed a focused spider to crawl Google Scholar's data and translated CiteSeer OAI Compliance (http://citeseer.csail.mit.edu/oai.html) into our repository with our schema, as shown in Fig. 4. As a result, we obtained 15,717 records from Google Scholar.

id	Paper	PaperTitle	PaperTit	paperC	pape	paperCites	pape	Authors	Source	Location
132	17952	Dynamic capabilities: what are they?	/url?sa=l	9 個群	/sch	被引用 865 次	/scl	KM Eisenhardt, JA Ma	Strategic Management Journal, 2000	doi.wiley.com
133	17952	企業核心能力: 理論溯源與課輯結構剖析	/url?sa=l	2 個群	/sch	被引用 110 次	/scl	王毅, 陳勁…	管理科學學報, 2000	萬方資料資源
134	17952	Knowledge of the firm and the evolutio	/url?sa=l	9 個群	/sch	被引用 613 次	/scl	B Kogut, U Zander	Journal of International Business St	palgravejourr
135	17952	Creating and Managing a High-Performan	/url?sa=l	9 個群	/sch	被引用 461 次	/scl	JH Dyer, K Nobeoka	Strategic Management Journal, 2000	doi.wiley.com
136	17952	Value creation in e-business	/url?sa=l	18 個群	/sch	被引用 405 次	/scl	R Amit, C Zott	2000	doi.wiley.com
137	17952	Strategy Research: Governance and Comp	/url?sa=l	5 個群	/sch	被引用 391 次	/scl	OE Williamson	Strategic Management Journal, 1999	doi.wiley.com
138	17952	企業的競爭優勢來源及其戰略選擇	/url?sa=l	2 個群	/sch	被引用 41 次	/scl	李海艦, 盖耀華	中國工業經濟, 2002	維普資訊
139	17952	Deliberate Learning and the Evolution	/url?sa=l	14 個群	/sch	被引用 317 次	/scl	M Zollo, SG Winter	Organization Science, 2002	atyponlink.co
140	17952	公司治理, 內部控制, 組織結構互動關係研	/url?sa=l	none	none	被引用 33 次	/scl	程新生	會計研究, 2004	維普資訊
141	17952	我國企業核心能力實證研究	/url?sa=l	3 個群	/sch	被引用 38 次	/scl	王毅	管理科學學報, 2002	萬方資料資源
151	17952	Knowledge transfer: A basis for compet	/url?sa=l	4 個群	/sch	被引用 267 次	/scl	L Argote, P Ingram	Organizational Behavior and Human De	dxslab.mis.cc
152	17952	Bridging Ties: A Source of Firm Hetero	/url?sa=l	6 個群	/sch	被引用 217 次	/scl	B McEvily, A Zaheer	Strategic Management Journal, 1999	doi.wiley.com
153	17952	Toward a synthesis of the resource-bas	/url?sa=l	4 個群	/sch	被引用 199 次	/scl	R Makadok	Strategic Management Journal, 2001	doi.wiley.com
154	17952	Research partnerships	/url?sa=l	4 個群	/sch	被引用 195 次	/scl	J Hagedoorn, AN Link	Research Policy, 2000	arno.unimaas.
155	17952	International Expansion by New Venture	/url?sa=l	2 個群	/sch	被引用 176 次	/scl	SA Zahra, RD Ireland	The Academy of Management Journal, 2	JSTOR
156	17952	Integration and Dynamic Capability: Ev	/url?sa=l	none	none	被引用 176 次	/scl	M IANSITI, KIMB CLAR	Industrial and Corporate Change	Oxford Univ P
157	17952	Towards a competence theory of the reg	/url?sa=l	5 個群	/sch	被引用 171 次	/scl	C Lawson	Cambridge Journal of Economics, 1999	cje.oupjourns
158	17952	Direct and moderating effects of human	/url?sa=l	none	none	被引用 168 次	/scl	MA Hitt, L Bierman,	Academy of Management Journal, 2001	aom.pace.edu
159	17952	Theory and research in strategic manag	/url?sa=l	5 個群	/sch	被引用 146 次	/scl	RE Hoskisson, MA Hit	Journal of Management, 1999	jom.sagepub.c
160	17952	Beyond local search: boundary-spanning,	/url?sa=l	6 個群	/sch	被引用 144 次	/scl	L Rosenkopf, A Nerka	Strategic Management Journal, 2001	doi.wiley.com
161	17952	Human resources and the resource based	/url?sa=l	4 個群	/sch	被引用 144 次	/scl	PM Wright, BB Dunfor	Journal of Management, 2001	jom.sagepub.c
162	17952	The Satisficing Principle in Capabilit	/url?sa=l	4 個群	/sch	被引用 143 次	/scl	SG Winter	Strategic Management Journal, 2000	doi.wiley.com
163	17952	Product sequencing: co-evolution of kn	/url?sa=l	10 個群	/sch	被引用 138 次	/scl	CE Helfat, RS Raubit	Strategic Management Journal, 2000	doi.wiley.com
164	17952	Transaction Cost Economics: How It Wor	/url?sa=l	13 個群	/sch	被引用 134 次	/scl	OE Williamson	De Economist, 1998	Springer
165	17952	Capabilities, cognition, and inertia:	/url?sa=l	5 個群	/sch	被引用 134 次	/scl	M Tripsas, G Gavetti	Strategic Management Journal, 2000	doi.wiley.com
166	17952	Managing organisational knowledge by d	/url?sa=l	7 個群	/sch	被引用 130 次	/scl	N Bontis	International Journal of Technology M	Inderscience
167	17952	Resource-based theories of competitive	/url?sa=l	3 個群	/sch	被引用 131 次	/scl	JB Barney	Journal of Management, 2001	jom.sagepub.c
168	17952	Understanding dynamic capabilities	/url?sa=l	4 個群	/sch	被引用 130 次	/scl	SG Winter	Strategic Management Journal, 2003	doi.wiley.com
169	17952	Alliance capability, stock market resp	/url?sa=l	4 個群	/sch	被引用 127 次	/scl	P Kale, JH Dyer, H S	Strategic Management Journal, 2002	doi.wiley.com
170	17952	Avoiding Complexity Catastrophe in Coe	/url?sa=l	4 個群	/sch	被引用 123 次	/scl	B McKelvey	Organization Science, 1999	JSTOR
3645	17952	基於知識的動態能力演化模型研究	/url?sa=l	2 個群	/sch	被引用 11 次	/scl	董俊武, 黃江圳, 陳	中國工業經濟, 2004	維普資訊
3646	17952	NEBIC: A Dynamic Capabilities Theory f	/url?sa=l	14 個群	/sch	被引用 47 次	/scl	BC Wheeler	Information Systems Research, 2003	atyponlink.co
3647	17952	The organizational impact of technolog	/url?sa=l	8 個群	/sch	被引用 46 次	/scl	M Chesbrough	Industrial and Corporate Change, 199	bbk.ac.uk
3648	17952	Competence-building, technology fusion	/url?sa=l	none	none	被引用 45 次	/scl	DT Lei	International Journal of Technology M	Inderscience
3649	17952	International Entrepreneurship: The Cu	/url?sa=l	none	none	被引用 46 次	/scl	SA Zahra, G George	Strategic Entrepreneurship: Creating	instruction.b

Fig. 4. Raw data in the repository

Cleaning: The purpose is to construct a temporal matrix from the cleaned bibliographic dataset. In this stage, our hypothesis is that there must be a paper that is particularly important in each research field. Most researchers cite an important paper in their published works. We call that paper the *root paper*. In our empirical study, we found that Teece's "Dynamic capabilities and strategic management" paper [18] had the highest citation rate (2,928). Hence, we adopted it as the root and used the 2,928 papers that cited it as correlation elements to find the relations.

Translating: To form an *N***N* correlation matrix, the correlation matrix can be reduced to a *Partially Ordered Set* (or poset) that contains a partially order relation. The relation formalizes the intuitive concept of the ordering, sequencing, or arrangement of the set's elements.

Visualizing: The objective is to translate a matrix into a network. We use AT&T's Graphviz (http://www.research.att.com/sw/tools/graphviz) format to visualize the result. In the operation, we provide two functions to facilitate the analysis. The first function allows users to scale the number of papers. Users can specify a number (*n*) (Top *n*), which represents the most frequently cited papers. The system will then analyze the percentages of each paper's "cited by" number among the Top *n* papers. The other function is called grouping. When a user moves the mouse on to any node or inputs a search query, the system will group the related papers by changing the color of the font. This function helps researchers find the context of papers more easily. Fig. 5 and Fig. 6 show the network topology of the top 100 and 200 papers respectively.

Analyzing and Weighting: This function allows users to sort papers by the year of publication, or "cited by number" information, or specify some conditions like "select

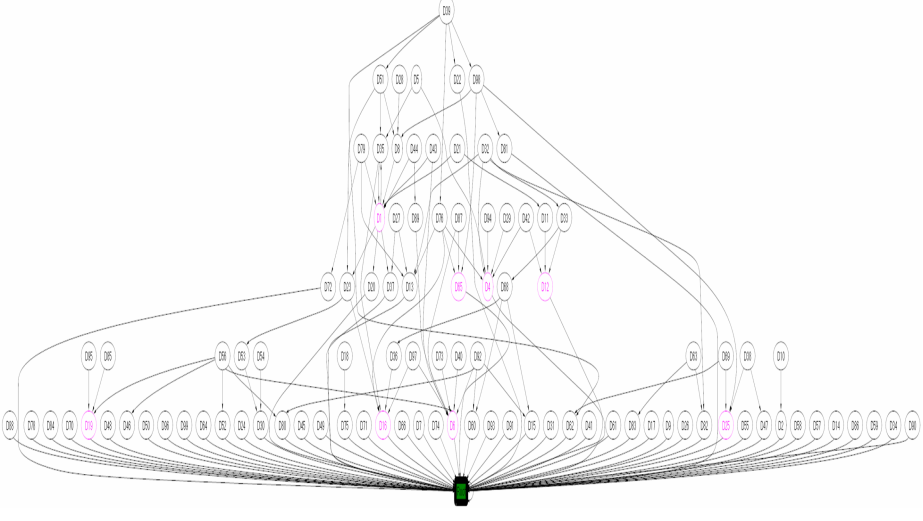


Fig. 5. The network topology of the top 100 papers

top *M* cited by number” papers or find papers with more than *N* links. For example, Fig. 5 shows the results for the case of the “Top 100 papers with more than three links”. We found that D1, D4, D6, D12, D16, D19, D25, and D65 are more important than the other papers.

4 Result Analysis and Discussion

We use the theory of Social Network Analysis (SNA) to analyze the cited-by graph. The theory views social relationships in terms of nodes and ties. By using the number of links as the condition, we observed several phenomena in the visualization results, as shown in Fig. 5 and Fig. 6.

4.1 Cited-by Data Analysis: Social Network Analysis

By changing the scale of the correlation matrix and linkage, we observed the following five phenomena.

Phenomenon 1: If paper A is cited by different groups, it means this node has several branches; that is, it is not a single node. No matter whether A’s “cited by” number is high or low, the paper is very important, as shown in Fig. 7(a). If A is widely quoted by researchers in different areas, we call this phenomenon a *diffusion model*. The real cases are shown in Fig. 5 and Fig. 6, i.e., D0, D1, D8, etc. Specifically, D0 is the original paper written by Teece et al. [18]. We find that if a node has more branches, extending from it, then it is more important than the other nodes.

Phenomenon 2: If paper A is cited by Paper B, and both papers have high “cited by” numbers, then A and B are both important, as shown in Fig. 7(b). This model is very

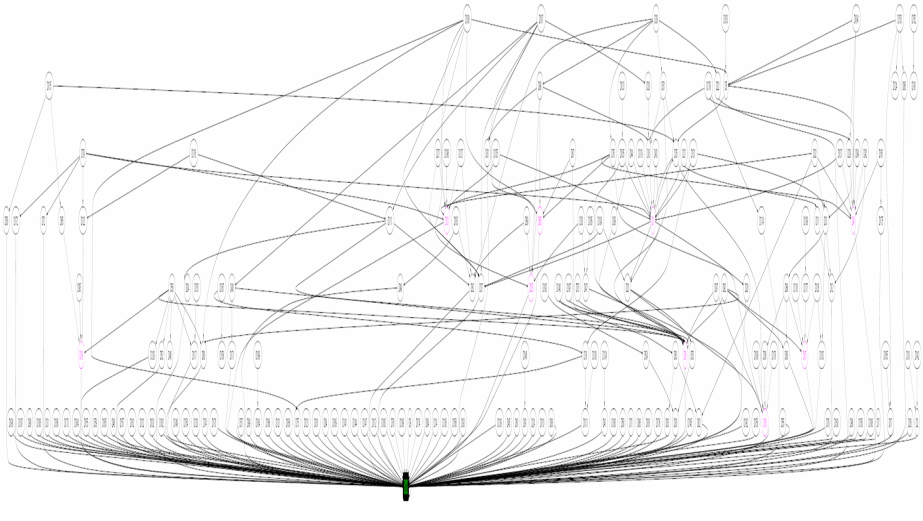
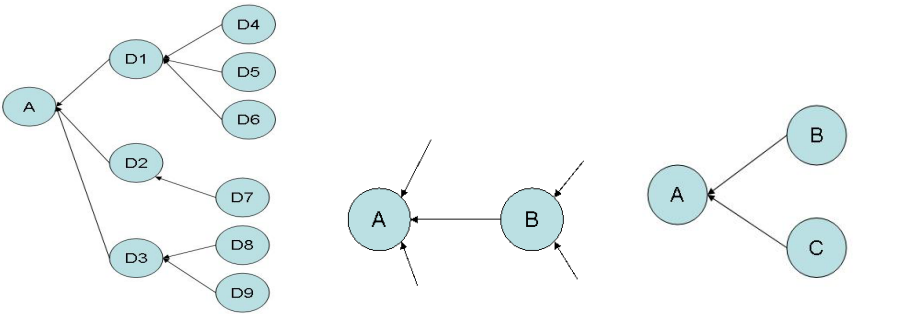
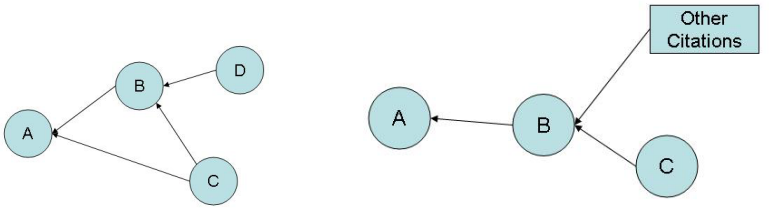


Fig. 6. The network topology of the top 200 papers



(a) Diffusion model **(b)** Structural cohesion model **(c)** " Structure hole" like model



(d) Phenomenon 4 **(e)** Phenomenon 5

Fig. 7. Cited-by data analysis: Social Network Analysis

similar to the "Structural Cohesion" model. The real cases shown in Fig. 6 are D1-D8, D4-D5, D16-D35. We observe several trends starting from the root, which means the research topic of "dynamic capabilities" has evolved into several subgroups. D1-D8 focus on *organization learning*; D4-D5 focus on the relationship between *dynamic capabilities* and *performance of a firm*; and D16-D35 focus on the relationship between *dynamic capabilities* and *innovation*.

Phenomenon 3: If paper A is cited by papers B and C, and both B and C have high "cited by" numbers, it does not matter whether A's "cited by" number is high or low because it is very important, as shown in Fig. 7(c). This phenomenon is similar to the "Structure Hole" theory in social network analysis. The cases shown in Fig. 6 are D1, D23, and D51. Note that D23 is Helfat & Raubitschek's paper [19], which addresses the relationship between knowledge, capabilities, and products.

Phenomenon 4: In Fig. 7(a), paper A is cited by papers B, C, and D, and papers A and C have high "cited by" numbers. This means that some papers cited A and C, but not B. Meanwhile, if paper D has a small "cited by" number, but it cites A and B, then paper B is important in some cases.

Phenomenon 5: We add more information, such as the value of the original paper's "cited by number," or we subtract the paper cited in the root paper, as shown in Fig. 7(e). If paper B has a high citation rate, but it does not have a high linkage in our results (shown in Fig. 5 and Fig. 6), then B may be important in another domain. Papers D11, D14, and D17 are the real cases, as shown in Fig. 6.

4.2 Discussion

By applying visualization to the different scales of the matrix elements and the linkage frequencies, some of which are shown in Figs. 5 and 6, we obtain the statistical results listed in Table 2. The results, which are based on "keyword search" and "cited by number" queries, provide a different view of the original correlation table described in Section 2. We find that the "cited by" number is not the only important evaluation criterion for deciding whether a paper is important. Some papers with high citation rates, such as D11 (cited 267 times), D14 (Cited 195 times), and D17 (cited 171 times) are not important in the visualization result. This answers the question: Does "cited by number" mean a paper is very important?

In addition, if we consider the "Structural cohesion" phenomenon, we find that several groups, e.g., D1-D8, D1-D105, D4-D5, D6-D40, D16-D35, and D30-D52-D23 are independent of the other papers. This answers the question: How can we discover the evolution of this domain?

By using linkage frequency analysis, we find that some nodes in Figs. 5 and 6 and some items in Table 2 are important in this domain. This answers the question: How can we determine which papers we should review?

Finally, we provide a user interface (shown in Fig. 8) to help researchers conduct literature reviews using CCTVA. The researcher can use the "search for a paper", "select a paper from the list", "move the mouse to select a node" functions to find the context of a research field. When a node in the panel is selected, the system will show related nodes and information.

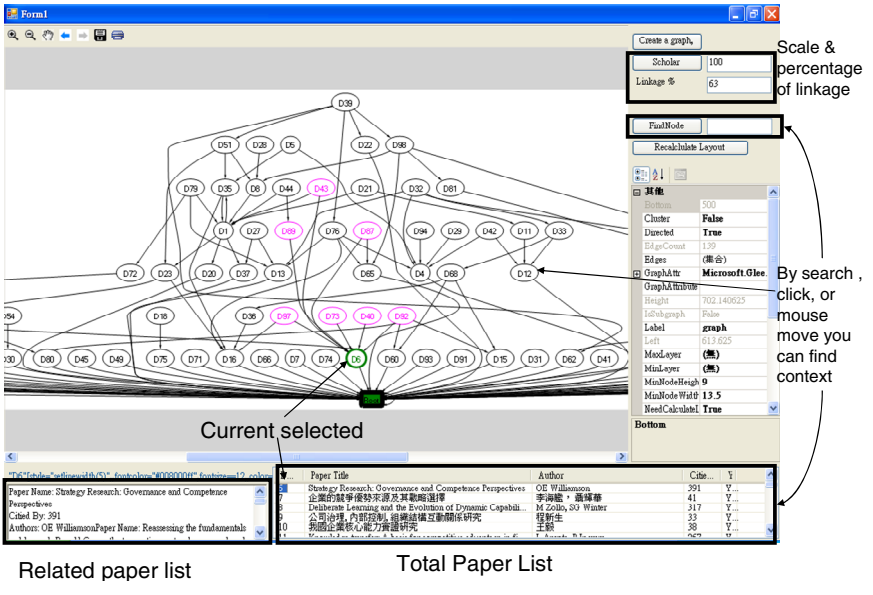


Fig. 8. The CCTVA system

Table 2. The query results

ID	Author	Title	Cited by number	Top 100, Links >3	Top 200, Links > 4	Top 300, Links> 5
D1	KM Eisenhardt, JA Martin	Dynamic Capabilities: What are they?	865	1	1	1
D4	JH Dyer, K Nobeoka	Creating and Managing a High-Performance Knowledge-Sharing Network: The Toyota Case	461	1	1	1
D5	R Amit, C Zott	Value creation in e-business	405		1	1
D6	OE Williamson	Strategy Research: Governance and Competence Perspectives	391	1	1	1
D8	M Zollo, SG Winter	Deliberate Learning and the Evolution of Dynamic Capabilities	317		1	1
D12	B McEvily, A Zaheer	Bridging Ties: A Source of Firm Heterogeneity in Competitive Capabilities	217	1		1
D13	R Makadok	Toward a synthesis of the resource-based and dynamic-capability views of rent creation	199		1	1
D15	SA Zahra, RD Ireland, MA Hitt	International Expansion by New Venture Firms: International Diversity, Mode of Market Entry, …			1	
D16	M IANSITI, KIMB CLARK	Integration and Dynamic Capability: Evidence from Product Development in Automobiles and Mainframe …	176	1	1	1
D19	RE Hoskisson, MA Hitt, WP Wan, D Yiu	Theory and research in strategic management: Swings of a pendulum …	146	1	1	
D22	SG Winter	The Satisficing Principle in Capability Learning	143			1
D25	M Tripsas, G Gavetti	Capabilities, cognition, and inertia: evidence from digital imaging	134	1	1	1
D37	S Karim, W Mitchell	Path-Dependent and Path-Breaking Change: Reconfiguring Business Resources Following Acquisitions in …	97			1
D65	D Holbrook, WM Cohen, DA Hounshell, S Klepper	The nature, sources, and consequences of firm differences in the early history of the semiconductor …	63	1	1	1
D75	TH Brush, KW Artz	Toward a Contingent Resource-Based Theory: The Impact of Information Asymmetry on the Value of …	53			1

5 Conclusion

In this paper, we propose a novel method called CCTVA (Collecting, Cleaning, Translating, Visualizing, and Analyzing) to help researchers find the context of papers on topics of interest. The method also enables users to gain insight into how a field's topics have evolved over time.

The contribution of this paper is three fold. First, the proposed CCTVA method helps researchers find the context of a research field and reduces the complexity of citations. Second, we generalize five phenomena to help researchers find important papers. The observed phenomena can help users explore the evolution of a research field, and find the most frequently referenced papers and the most published authors in that field. Finally, we provide an interface to help researchers find a topic's branches in a research field and select the branch of interest.

In our future work, we will add some algorithms to help users find more information like co-authorship, keywords, and key phrases in papers. We will also use text mining techniques to analyze papers in order to extract more information about the contexts of the papers.

Acknowledgments

We are indebted to Google Scholar and CiteSeer for providing useful search tools. This research was supported in part by the National Science Council of the R.O.C. under Grant NSC 95-2752-E-001-PAE and the Thematic Program of Academia Sinica under Grant AS95ASIA02.

References

1. Google Scholar, http://scholar.google.com.tw
2. CiteSeer, http://citeseer.ist.psu.edu
3. Jacobson, A., Laurence, P.: The Cost of Knowledge. Harvard Business Review 84(11), 34–34 (2006)
4. Adair, W.C.: Citation indexes for science. American Documentation 6, 31 (1995)
5. Giles, C.L., Bollacker, K.D., Lawrence, S.: CiteSeer: An Automatic Citation Indexing System. In: Proceedings of the Third ACM Conference on Digital Libraries, pp. 89–98 (1998)
6. Lin, S.D., Chalupsky, H.: Using Unsupervised Link Discovery Methods to Find Interesting Facts and Connections in a Bibliography Dataset. SIGKDD Explor. Newsl. 5(2), 173–178 (2003)
7. Smeaton, A.F., Keogh, G., Gurrin, C., McDonald, K., Sodring, T.: Analysis of papers from twenty-five years of SIGIR conferences: what have we been doing for the last quarter of a century? In: ACM SIGIR Forum, pp. 49-53 (2003)
8. Nascimento, M., Sander, J., Pound, J.: Analysis of SIGMOD's CoAuthorship Graph. SIGMOD Record 2003 32(3) (2003)
9. Lee, B., Czerwinski, M., Robertson, G., Bederson, B.B.: Understanding Research Trends in Conferences Using PaperLens. In: Extended Abstsracts of CHI 2005, pp. 1969–1972 (2005)

10. Fortuna, B., Grobelnik, M., Mladenic, D.: Visualization of Text Document Corpus. Informatica (Slovenia) 29(4), 497–504 (2005)
11. OAIster, http://oaister.umdl.umich.edu/o/oaister/
12. Windows Live Academic Search, http://academic.live.com
13. DBLP, http://dblp.uni-trier.de/db/index.html
14. ACM Digital, Library, http://portal.acm.org/dl.cfm
15. IEEE Xplore, http://ieeexplore.ieee.org/Xplore/dynhome.jsp
16. Arxiv.org, http://arxiv.org/
17. PubMed, http://www.pubmedcentral.nih.gov/
18. Teece, D.J., Pisano, G., Shuen, A.: Dynamic Capabilities and Strategic Management. Strategic Management Journal 18(7), 509–533 (1997)
19. Helfat, C.E., Raubitschek, R.S.: Product Sequencing: Co-evolution of knowledge, capabilities and products. Strategic Management Journal 21 (2000)

Online Communities: A Social Computing Perspective

Xiarong Li, Daniel Zeng, Wenji Mao, and Fei-yue Wang

Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
lxr606@gmail.com

Abstract. In recent years, the growth of the Internet has facilitated the rapid emergence of online communities. In this paper, we survey key research issues on online communities from the perspectives of both social science and computing technologies. We also sample several major online community applications, and propose some directions for future research.

Keywords: online community, social computing, social theory, computing technology.

1 Introduction

In the last twenty years, the growth of the Internet has greatly facilitated the rapid emergence of online communities. Community Memory of Berkeley, California, initially started in the mid-1970s, was viewed by many researchers as the first practical online community [1]. The phrase, online community, however, was created only later in Hiltz's book in 1984 [2]. An online community is a group of people interacting in a virtual environment, who have a purpose, are supported by technology, and are guided by norms and policies [3]. In this paper, we extend the scope of 'online community' to refer to people, the technological infrastructure of virtual environments, and what is produced during online interactions.

Whittaker *et al* [4] identified the core characteristics of online communities. Online communities can be viewed as technology-supported groups of people with a shared goal, interest, need, or activity. With intense interactions and strong emotional ties, its members share resources and provide information, support and services to each other in an established context of social conventions, language, and protocols. An online community consists of the technological environment, members and their social activities, and the study of online communities has to synthesize social theories and computing studies.

Doug Schuler and some other researchers published a Special Issue on Social Computing in Communications of the ACM (1994), in which social computing is described as any type of computing application in which software serves as an intermediary or a focus for a social relation [5]. Schuler [1] discovered the potential of community network and took online communities as the center of the social and political architecture. In the more recent study, Wang *et al* [6] continued the research on social computing, and identified online communities as one of the four main applications in social computing.

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 355–365, 2008.

The rest of this paper is organized as follows. In Section 2, we introduce the social theories of online communities. Related computing technologies are introduced in Section 3. Through various applications, we discuss various issues related to improving and extending online communities in Section 4. In Section 5, we summarize the paper and propose several future research directions.

2 Social Theories of Online Communities

In September 1998, the first Advanced Research Workshop of the Joint European Commission/National Science Foundation Strategy Group was held in [7]. The group recommended a series of research workshops to enable early identification of key research challenges and opportunities in information technology. Online community is one of the proposed research priorities. The group also stated that theories from sociology, psychology, social psychology, linguistics, communications research, and psychotherapy can help inform research and development in online communities.

2.1 General Theories

Anthropology
Anthropology is the study of human beings by exploring the differences and similarities between people in terms of cultures, societies, etc. As the new culture of human life is driven by information technology, online phenomena share important similarities with other types of human experience and are amenable to relatively conventional anthropological concepts and assumptions [8]. New communicative practices such as those manifested through online communities fit right into the research domain of anthropologists.

For instance, the online game community is a major representative of online communities. One popular example of these games is the World of Warcraft (WOW). Just like in the early days of the human history, there are many communities composed by gamers called guilds [9]. Researchers have found that a community in the WOW exhibits similar characteristics of a new tribalism [10]. Social fairness in a guild can motivate the gamers to make more contributions. The evolution of online communities in many cases also shares many similarities with the evolution of the human society.

Social Psychology
Hundreds of online communities emerge everyday. Participants of these communities share their experiences, and offer and receive emotional support in a climate of trust, equality, and empathy. Social psychology can also be applied to online community research. Preece and Ghozati [11] found that a community's focus of interest, gender ratio, and hostile or moderate atmosphere can influence a community's empathy. A well designed empathic community will improve the quality of life for many people.

Participation and contributions of individual members are essential to the success of an online community. As such, researchers have paid much attention to member motivation. They found that calling users' attention to their uniqueness and to the benefits they provide have a great impact on increasing their contributions [12]. The participation stimulant is also influenced by offline interaction. Users who perceive greater degree of social presence are more likely to share their information [13].

Social Network Theory
Social connections are a distinct feature of people's interactions. The well-known "six degrees of separation'' phenomenon implies that the distance between any two individuals in terms of direct personal relationships is relatively small. Social network theory examines the patterns and characteristics of social connections and their relationship to individual's lives and societal organization. Social network theory can be used to analyze the online communities from a sociological perspective [14]. Assembling data and benefits for users are two key questions in online social network studies [15]. This area has been drawing much attention due to many new emerging research questions.

Computer-Mediated Communication (CMC) Theory
Compared to face-to-face (FTF) communication, CMC is a more egalitarian medium, with greater equality of participation, relatively less intense normative pressures, and higher incidence of uninhibited behavior [16]. Increasingly, CMC technologies are being used to solve a wide range of problems and have been tapping into the Internet to as a technological platform and an operating environment [17]. CMC provides one of the underlying theories for the study of online communities and from a practical perspective, advances made in CMC could directly lead to new activities and models of online communities.

Sociolinguistics Theory
Linguisticians predicted that online interaction would have a long-term effect on the evolution of language. Development of an online community can bring about changes in linguistic interaction patterns. These patterns could converge as the members of this community converge on a linguistic style [18]. "In the Internet, nobody knows you are a dog." (However, in a recent study, based on sociolinguistics theory, men and women can be identified [19].) Combined with the social network analysis, sociolinguistic research on the online interaction holds a lot of potentials.

In an online community, the members' behavior is multifarious, such as communicating with each other, organizing small groups, constituting rules and custom, using special but uniform language, sharing their information and experience. Such phenomena are quite similar to what we usually do in real life, making existing social theories a proper analysis tool for investigation purposes. In the meanwhile, online communities provide a great test bed to verify these social theories. There exist, however, phenomena that seem to be unique to online communities. Such phenomena present great research opportunities and may lead to theoretical contributions that would enrich the understanding of human societies in general. The next subsection summarizes several prominent social phenomena identified from online communities.

2.2 Social Phenomena Specific to Online Communities

Lurkers
A lurker is a person who reads discussions on a message board, newsgroup, chatroom, file sharing or other online environments, but rarely participates [20]. It is reported that about 90% members of online communities are lurkers. Lurking is a common

activity in online life. Preece and Nonnecke have written a series of papers on lurkers online [21] [22] [23]. They believe that lurking is a systematic and idiosyncratic process, with well-developed rationales and strategies. The top 5 reasons for lurking are as follow [24],

- Shy about posting
- Want to remain anonymous
- Join wrong group
- Fear of being treated poorly
- Poor quality interaction

Lurkers have something in common: frequent login without posting, remaining on standby, posting collected information in order to eliminate the fear of posting something in error [25]. Lurking is usual behavior of community members. Thus it need not necessarily be viewed as passive participation. Lurkers' seemingly silent participation conveys deeper level of engagement than that of non-lurkers.

Other Empirical Findings

Facebook is popular nowadays, on which users keep their own profiles and friend lists. Researchers found that populating profile fields on Facebook is positively related to the number of friends a user lists [26]. There is a tendency of members to join online communities as a target, who seeks support, help and sympathy. However, a recent study showed that online people very rarely ask for help and support directly [27]. There is another interesting finding that users rate fairly consistently across rating scales and tend to rate toward the prediction the system shows, whether the prediction is accurate or not [28]. In other words, users' mind and choices could be manipulated. In addition, many online communities such as forums provide anonymity option for users. Anonymity policies can have a significant effect on the professionalism and productiveness of comments posted in an online community. In an experiment, researchers found that eliminating anonymity option nearly eliminated negative comments [29].

More in-depth work is needed to explore these online community-specific phenomena. Nonetheless, the emerging literature already contains many interesting findings based on social theories. In addition to social phenomena, computing technologies are an essential part of online communities. In the next section, we discuss related research from a computational angle.

3 Computational Studies of Online Communities

The Internet is extending the types of networked communities that have already become prevalent [30]. Many communities have moved from the real world to the virtual environment, from offline to online. Computing technologies play a key role in this transition. In this section, we discuss two lines of computational studies of online communities: community mining and computing community characteristics.

3.1 Community Mining

Although many online communities can be easily identified, there are online communities that are inconspicuous or even hidden. To find them in the huge amount of data on the Internet, researchers have developed a range of computing techniques such as searching, web crawling, social network analysis and data mining. In the context of online community, these techniques are referred to as community mining.

Web Crawling
Unlike the traditional methods such as search or resource-gathering algorithms that find information on a specified topic, Kumar *et al*. [31] proposed a novel method called Web crawl, which helps to identify all instances of graph structures that are indicative signatures of communities. A community on the web is defined as a set of sites that have more links to members than to non-members, which can be efficiently calculated in a maximum flow framework. Under this concept, a maximum flow-based web crawler can approximate a community by directing a focused web crawler along link paths that are highly relevant [32]. Besides the formal communities, the self-organization link structure communities are more complex. A novel method was developed by Flake *et al* [33]. Since this method does not make use of any text-based approaches, identified communities can be used to infer meaningful text rules and to augment text-based methods.

Community Network Analysis
There are various communities in different forms. However, at the heart of each community lies a social network. Applying social network analysis in community mining presents interesting opportunities. Most of the traditional methods on community mining assume that there is only one kind of relation in the network, and the mining results are independent of the users' needs or preferences. Cai *et al*'s [34] approach to social network analysis and community mining represents a major shift in methodology from the traditional ones, a shift from single-network, user-independent analysis to multi-network, user-dependant and query-based analysis. In addition, Yang *et al* [35] have developed a new algorithm to compute signed social networks that contain both positive and negative relations.

Communities in Blogs
A blog is written by a single author and uniquely identified by that person. Thus, a blog functions as an amalgam of document and person, and blogs link hypertext networks with social networks. The novel blogging software is reshaping the online community. Mining virtual communities in blogs is a new branch in community mining [36].

3.2 Computing Community Characteristics

When a person joins an online community, he or she may be involved in a number of activities such as learning about the community structure, searching for information, finding friends with similar hobbies, asking experts and leaders in the community, etc. Computational methods can help characterize such community activities.

Role Identification
Every member plays a role in a community. Some of them are the leaders and experts who are the main information providers at the core of community. Users turn to them for help. Researchers have been developing methods to identify these leaders and experts automatically. A set of network-based ranking algorithms have been proposed, including PageRank and HITS, to identify users with high expertise [37]. It is found that, structural information can be used for evaluating an expertise network in an online setting, and relative expertise can be automatically determined using social network based algorithms. In another study, the novel approach to combining weighting with social computing helped identify key members at a deeper level in Usenet groups [38].

Relationship (trust) Finding
Tens of millions of users participate in Web-based social networking. Privacy and trust are becoming prominent topics for research on online interactions. For example, if Alice highly trusts Bob, and Bob highly trusts Chris, can we recommend that Alice should have some level of trust for Chris? Early research on the topic of trust has focused largely on digital signatures, certificates, and authentication. Golbeck *et al* [39] suggested integrating social network analysis and other traditional methods to create a trust network. She then presented two sets of algorithms for calculating these trust inferences [40].

Information Sharing
Information and communication technologies (ICTs) are dramatically enhancing our ability to collect, organize, and share information. In a recent study, a new approach has been proposed, which allows a group of like-minded people to share documents in an implicit and intelligent way. The approach can facilitate the document recommendation and avoid broadcasting requests so as to protect users' privacy [41]. Wikipedia is a famous open community on the Internet. Users in the community can easily edit, review and publish articles collaboratively. Researchers have developed two models, namely basic model and peer review model, to measure the quality of the articles and the authorities of their contributors [42]. These methods can help the users easily find out valuable information without reading entries one by one.

In summary, computing technologies play a key role in online communities' development and growth, while social theories provide useful guidance and characterization. Synthesizing them properly is essential to the study of online communities.

4 Applications of Online Communities

Various kinds of communities have been developed to meet people's needs for communication and information sharing all over the world. Hiltz and Wellman have predicted that the development of CMC would make virtual communities replace the real ones and connect more people geographically dispersed [43]. The end results are part of the continuing social transformation toward global connectivity. Online communities actually make the world flat.

4.1 Community Evaluation and Improvement

The community members and the platform are the two important constitutional elements of an online community. To build an online community, a thorough understanding of an audience's distinctive demographic, psycho-demographic, and Internet experience characteristics are critical to crafting solutions helping to build sustainable online communities [44].

In recent years, social networking services (SNS) are becoming quite popular. SNSs provide an online private space for individuals and tools for people to interact with others in the cyberspace. Ahn *et al* [45] compared the structures and features of three popular online social networking services: Cyworld, MySpace, and orkut. The results show that online networks are quite similar to the real social networks. Fu *et al* [46] performed an empirical analysis of two Chinese online social networks—Sina blogging network and Xiaonei network (SNS). They found that the blogging network shows the disassortative mixing pattern, whereas the SNS network displays an assortative one.

Online professional communities such as open source software (OSS) communities have flourished in conjunction with the rise of the Web. A major problem for online professional communities is that it is difficult for their members to find the vast amount of information as well as other members' activities. Semantic Web technologies can make unstructured or semi-structured Web information meaningful [47]. The ontology of creating and maintaining a Semantic Web requires one to reliably predict how other members of the community would interpret the symbols of an ontology based on their limited description. Mika's study [48] extended the traditional bipartite model of ontology with the social dimension, leading to a tripartite model of actors, concepts and instances. With more content and more advantages, online communities based on the Semantic Web will possibly replace the more traditional ones.

4.2 Online Commerce

By providing customers with the opportunity to interact with each other and with companies, online communities can foster meaningful customer relationships by customizing products and services to meet consumers' demands and interests [49], so as to, for example, help companies increase the brand loyalty. Online communities' potentials in E-commerce include three main aspects: the building of trust, the collection and effective use of community knowledge and the economic impacts of accumulated buying power [50] [51]. By increasing community commitment, companies can improve their financial performance through consumer rephrasing and word-of-mouth marketing. Online communities have the potential of becoming a strong social structure that promotes trust building and facilitates the growth of electronic commerce.

4.3 Mobile Online Communities

Online communities are being developed through mobile technologies such as PDA, Pocket PC and mobile phones. Coupled with the rapid uptake of mobile phone technology in the developing world and the growing popularity of Internet-based

SNS, social networking applications developed for mobile phones could leverage both existing technology usage patterns and information seeking patterns in the developing world. This is reflected in the new word 'MoSoSo' (Mobile Social Software) [52]. Researchers have also experimented a messaging application for camera phones to collectively create stories called Media Stories with the idea of collectively created albums [53]. Using this kind of system, people can exchange information more quickly. A new fully-distributed mobile portfolio improves the flexibility to conduct interactions or share portfolio resources among the members of a community [54]. This distributed portfolio is expected to help team members interact, and exchange resources and experiences.

4.4 Other Emerging Applications

There is a notable phenomenon that the female are underrepresented in the areas of STEM (Science, Technology, Engineering, and Mathematics). Researchers from Germany created an online community and e-mentoring program called CyberMentor for German high school girls and women to encourage them to engage in STEM vocational fields [55].

Online community is viewed as a more conducive organizational form to human-centric computing than traditional business organizations [56]. In previous online communities, the discovery of people with the same interests or a similar context was accomplished manually by the users themselves. Recently, researchers have designed a framework for user-centric community platforms to support flexible and adaptive management of user communities [57].

5 Conclusion and Future Work

Along with the wide adoption of Web 2.0 technologies, online communities are expected to continue to grow. As a research area, online communities are increasingly attracting attention of researchers from both social sciences and computing technologies. From a sociological point of view, social theories are widely used to analyze the characteristics in online communities. On the other hand, some online phenomena can not be explained using current social theories, and thus new creative theories are needed. From a computing technologies viewpoint, researchers are focusing on techniques for mining the community and computing community characteristics. Community mining aims to find hidden communities and synthesize various computing techniques. Computing community characteristics help users understand a community and accomplish their goals.

As is shown in this paper, online communities are flourishing both in research and in practice. However, much work still remains for social computing researchers and developers. Some future directions that deserve greater attention include but not limited to the follows:

- Utilize fruitful results in social studies to improve the community experience;
- Propose new social theories or modify existed ones to explain the emergent phenomena in online communities;
- Explore the structure and formation mechanism of online communities;

- Develop novel algorithms for mining community and computing community characteristics;
- Enhance self-organization and self-management of community, and the adaptability of community building platforms to other devices especially mobile devices;
- Automate information and knowledge sharing, and expand the scope of online community applications.

Acknowledgements

This work is supported in part by NNSFC #60621001 and #60573078, MOST #2006AA010106, #2006CB705500 and #2004CB318103, CAS #2F05N01 and #2F07C01, and NSF #IIS-0527563 and #IIS-0428241.

References

1. Schuler, D.: Community networks: building a new participatory medium. Communications of the ACM 37, 38–51 (1994)
2. Hiltz, S.R.: Online Communities: A Case Study of the Office of the Future. Interactive computer systems (1984)
3. Preece, J.: Online communities: Designing usability, supporting sociability. John Wiley & Sons, Chichester, England (2000)
4. Whittaker, S., Isaacs, E., O'Day, V.: Widening the net: workshop report on the theory and practice of physical and network communities. SIGCHI Bull. 29, 27–30 (1997)
5. Schuler, D.: Social computing. Communications of the ACM 47, 29 (1994)
6. Wang, F.Y., et al.: Social computing: From social informatics to social intelligence. IEEE Intelligent Systems 22, 79–83 (2007)
7. Brown, J.R., et al.: Human-centered computing, online communities, and virtual environments. IEEE Computer Graphics and Applications 19, 70–74 (1999)
8. Wilson, S.M., Peterson, L.C.: The anthropology of online communities. Annual review of anthropology 31, 449–467 (2002)
9. Jang, C.Y.: Managing fairness: Reward distribution in a self-organized online game player community. In: HCI 2007:Online Communities and Social Computing, pp. 375–384 (2007)
10. Brignall, T.W., et al.: An online community as a new tribalism: the world of warcraft. In: Proceedings of the 40th Hawaii International Conference on System Sciences (2007)
11. Preece, J., Ghozati, K.: Observations and explorations of empathy online. In: The Internet and Health Communication: Experience and Expectations, pp. 237–260. Sage Publications Inc., Thousand Oaks (2001)
12. Ludford, P.J., et al.: Think different: Increasing online community participation using uniqueness and group dissimilarity. In: CHI 2004, vol. 6, pp. 631–638 (2004)
13. Koh, J., et al.: Encouraging participation in virtual communities. Communications of the ACM 50, 68–73 (2007)
14. Wellman, B.: For a social network analysis of computer networks: a sociological perspective on collaborative work and virtual community. In: Proceedings of the 1996 ACM SIGCPR/SIGMIS conference on Computer personnel research, ACM Press, New York (1996)

15. Staab, S., et al.: Social networks applied. IEEE Intelligent Systems, 80–93 (January/ February 2005)
16. Bordia, P.: Face-to-face versus computer-mediated communication: A synthesis of the experimental literature. Journal of Business Communication 34, 99–118 (1997)
17. Herring, S.C.: Slouching toward the ordinary: current trends in computer-mediated communication. New media & society 6, 26–36 (2004)
18. Cassell, J., Tversky, D.: The language of online intercultural community formation. Journal of Computer Mediated Communication 10 (2005)
19. Gefen, D., Ridings, C.M.: If you spoke as she does, sir, instead of the way you do: a sociolinguistics perspective of gender differences in virtual communities. SIGMIS Database 36, 78–92 (2005)
20. http://www.wikipedia.org
21. Nonnecke, B., Preece, J.: Shedding light on lurkers in online communities. In: Ethnographic Studies in Real and Virtual Environments: Inhabited Information Spaces and Connected Communities, Edinburgh (1999)
22. Nonnecke, B., Preece, J.: Lurker demographics: counting the silent. In: CHI 2000: Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 73–80. ACM, New York (2000)
23. Nonnecke, B., Preece, J.: Why lurkers lurk. In: Americas Conference on Information Systems 2001 (2001)
24. Preece, J., et al.: The top 5 reasons for lurking: improving community experiences for everyone. Computers in Human Behavior, 2 (2004)
25. Lee, Y.W., et al.: Lurking as participation? A community perspective on lurkers' identity and negotiability. In: ICLS 2006, pp. 404–410 (2006)
26. Lampe, C., et al.: A familiar face (book): Profile elements as signals in an online social network. In: CHI 2007 Proceedings Online Representation of Self, pp. 435–444 (2007)
27. Pfeil, U., Zaphiris, P.: Patterns of empathy in online communication. In: CHI 2007: proceedings of the SIGCHI conference on Human factors in computing systems, pp. 919–928. ACM, New York (2007)
28. Cosley, D., et al.: Is Seeing Believing? How Recommender Interfaces Affect Users' Opinions. In: Recommender Systems and Social Computing, CHI 2003, pp. 585–592 (2003)
29. Kilner, P.G., Hoadley, C.M.: Anonymity options and professional participation in an online community of practice. In: CSCL 2005: Proceedings of th 2005 conference on Computer support for collaborative learning, International Society of the Learning Sciences, pp. 272–280 (2005)
30. Wellman, B.: Computer networks as social networks. Science 293, 2031–2034 (2001)
31. Kumar, R., et al.: Trawling the web for emerging cyber-communities. In: WWW 1999: proceeding of the eighth international conference on World Wide Web, pp. 1481–1493. Elsevier, New York (1999)
32. Flake, G.W., Lawrence, S., Giles, C.L.: Efficient identification of web communities. In: KDD 2000: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 150–160. ACM, New York (2000)
33. Flake, F., et al.: Self-organization and identification of web communities. Computer 35, 66–70 (2002)
34. Cai, D., et al.: Mining hidden community in heterogeneous social networks. In: LinkKDD 2005: Proceedings of the 3rd international workshop on Link discovery, pp. 58–65. ACM, New York (2005)
35. Yang, B., et al.: Community mining from signed social networks. IEEE transactions on knowledge and data engineering 19, 1333–1348 (2007)

36. Blood, R.: How blogging software reshapes the online community. Communications of the ACM 47, 53–55 (2004)
37. Zhang, J., et al.: Expertise networks in online communities: Structure and algorithms. In: WWW 2007, Banff, Alberta, Canada, pp. 221–230 (2007)
38. Nolker, R.D., Zhou, L.: Social computing and weighting to identify member roles in online communities. In: Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2005), pp. 1–7 (2005)
39. Golbeck, J., Parsia, B., Hendler, J.: Trust networks on the semantic web. In: WWW 2003 (2003)
40. Golbeck, J.A.: Computing and applying trust in web-based social networks. PhD thesis, University of Maryland, College Park in partial fulfillment (2005)
41. Kanawati, R., Malek, M.: Computing social networks for information sharing: a case-based approach. In: Schuler, D. (ed.) HCII 2007 and OCSC 2007. LNCS, vol. 4564, pp. 86–95. Springer, Heidelberg (2007)
42. Lim, E.P., et al.: Measuring qualities of articles contributed by online communities. In: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence (2006)
43. Hiltz, S.R., Wellman, B.: Asynchronous learning networks as a virtual classroom. Communications of the ACM 40, 44–49 (1997)
44. Andrews, D.C.: Audience-specific online community design. Communications of the ACM 45, 64 (2002)
45. Ahn, Y.Y., et al.: Analysis of topological characteristics of huge online social networking services. In: WWW 2007 / Track: Semantic Web, pp. 835–844 (2007)
46. Fu, F., et al.: Empirical analysis of online social networks in the age of web 2.0, Physica A (2007)
47. Ankolekar, A., et al.: Supporting online problem solving communities with the semantic web. In: WWW 2006, Edinburgh, Scotland (2006)
48. Mika, P.: Ontologies are us: A unified model of social networks and semantics. Web Semant 5, 5–15 (2007)
49. Armstrong, A., Hagel, J.: The real value of on-line communities. Harvard Business Review 74, 134–141 (1999)
50. Schubert, P., Ginsburg, M.: Virtual communities of transaction: the role of personalization in electronic commerce. Electronic Markets 10, 45–55 (2000)
51. Balasubramanian, S., Mahajan, V.: The economic leverage of the virtual community. Int. J. Electron. Commerce 5, 103–138 (2001)
52. Kolko, B.E., et al.: Mobile social software for the developing world. In: HCI 2007:Online Communities and Social Computing (2007)
53. Salovaara, A., et al.: Collective creation and sense-making of mobile media. In: CHI 2006 Proceedings Social Computing 3 (2006)
54. Herrera, O., et al.: A mobile portfolio to support communities of practice in science education. In: HCI 2007: Online Communities and Social Computing, pp. 416–425 (2007)
55. Schimke, D., et al.: The relationship between social presence and group identification within online communities and its impact on the success of online communities. In: HCI 2007: Online Communities and Social Computing, pp. 160–168 (2007)
56. Zhao, D., et al.: The future of work: What does online community have to do with it? In: Proceedings of the 40th Hawaii International Conference on System Sciences, pp. 1530–1605 (2007)
57. Foell, S., et al.: Self-organizing pervasive online communities. In: Eighth International Symposium on Autonomous Decentralized Systems (ISADS 2007) (2007)

User-Centered Interface Design of Social Websites

Yang-Cheng Lin[1] and Chung-Hsing Yeh[2]

[1] Department of Arts and Design, National Hualien University of Education, Hualien, 970, Taiwan
lyc0914@cm1.hinet.net
[2] Clayton School of Information Technology, Faculty of Information Technology, Monash University, Clayton, Victoria 3800, Australia
ChungHsing.Yeh@infotech.monash.edu.au

Abstract. This paper develops neural network (NN) models to examine how key design elements of a social website will affect users' feelings or perceptions. An experimental study of 96 university websites is conducted based on a user-centered approach. The study identifies seven website design elements and 33 representative websites as experimental samples for training and testing four NN models. These four NN models are built to formulate the relationship between seven website design elements and three users' feelings of websites. The result of the study shows that the combined NN model has an accuracy rate of 83.93% for predicting the values of three users' feelings of websites. This suggests that the combined NN model is a promising approach for modeling users' specific expectations of websites, thus providing an effective mechanism for facilitating user-centered interface design of social websites.

1 Introduction

The Internet or websites play increasingly significant role in our daily life, due to the rapid development of social software and Web 2.0 technologies [4], such as blogs, collective intelligence, mash-up, peer to peer network (P2P), podcasts, really simple syndication (RSS), wikis, and social network service [1, 2, 5, 16]. These new technologies have changed the interaction of social behaviors. For example, we can send emails to others to say "Happy New Year", instead of meeting face to face. In addition, we can express our feelings and comments against a certain electronic product or maker brand on our own blogs. There is a tendency towards user generated content (UGC) or user created content (UCC) [5] based on the successes of YouTube, WiKi, or Cyworld (the biggest website of social network service in Korea, having 18 million members about one-third of the Korean population [19]).

According to the BIGresearch survey in June 2006 [18], there were 87 % users (or consumers) who searched and read the relevant comments or recommendations to products on the Internet, to decide for themselves what products they should buy, whose information they should consume, what marketing they want, and finally go to shopping and buy. Thus, the key factor for an effective social website is the users, particularly the younger generations who spend so much of their time on the Internet.

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 366–376, 2008.

To get users' attention and to give visual focus on the website function, an effective (social) website need to use adequate graphics and have a good usability (or human factors) [15]. Effective use of graphics will enhance the website appearance and make it visually appealing, particularly affecting how the users "feel" about the website [17]. On the other hand, the usability (or human factors) is applied to create a highly usable system through a process that involves getting information from users who actually use the website, and makes it easy to learn, easy to use, easy to remember, error tolerant, and subjectively pleasing [14].

In this paper, we aim to explore the relationship between the users' feelings (or perceptions) and design elements of a website, in order to comprehend what are the key design elements for an effective website? How to use the adequate graphics to enhance a website appearance? What characters should a website with good usability has? Is there an optimal combination of website design elements that best matches desirable feelings of the users? For example, if web designers want to design "easy-to-use" or "clear-to-follow" websites, are there guidelines of the website design to carry out? To illustrate how the approach used in this paper can answer these questions, we conduct an experimental study on university websites using a number of modeling and analytical techniques including user-centered methods [8, 10], morphological analysis [7, 9], and neural networks [13].

In subsequent sections, we first present the user-centered experiments for extracting representative website samples and for identifying key design elements of websites using morphological analysis. We then describe an evaluation process for assessing the website samples with respect to users' feelings characterized by three image word pairs. Next we construct and evaluate NN models based on the experimental data. Finally we discuss how the NN models can be used as a design database for supporting the website design process.

2 User-Centered Experiments

In the experiment study, we first collected 96 homepages of university websites as experimental samples. We asked 111 subjects in total to estimate these experimental samples. Except for the 6 expert subjects of the web designers, we investigated the views of young people as they usually paid more attention to websites than other age groups. In what follows, we present the experimental study and its results in the context of the four primary phases of the user-centered experiment.

2.1 Experimental Subjects

The experimental study involved 111 subjects, divided into 5 groups. Each of the fist two groups had 30 subjects individually to extract the representative image word pairs based on four primary steps (please refer to our previous study [6]). The third group had 6 expert subjects who were the web designers with more than 5 years of web design experience to perform the morphological analysis. There were 30 subjects in the fourth group for evaluating the website image of the experimental samples, whose result is to be used as a numerical data source for constructing the NN models (presented in Section 3). The average age of the 30 subjects in the fourth group was 20.5 and each

had more than 6 years' experience of using websites. The fifth group had 15 subjects (the average age being 23.6 with more than 10 years' experience of using websites) for estimating the experimental samples, whose result is to be used as a basis for evaluating the performance of the NN models (presented in Section 4).

2.2 Morphological Analysis of Website Design Elements

The morphological analysis was conducted in two steps [7, 9]. In the first step, 6 design experts of the third group with at least 5 years of website design experience were asked to write down the key design elements of websites that affect users' feelings of websites, according to their knowledge and experience. In the second step, the 6 experts formed a focus group [3] to discuss the results and combine similar opinions or components. Table 1 shows the result of the morphological analysis. There are 7 key influential design elements extracted from the 96 website samples, including "Ratio of Graphics to Text", "Blank Ratio", "Layout Style", "Frame Style", "Hyperlink Style", "Number of Colors", and "Background Color". In the morphological analysis, the 6 experts discarded the minor design elements, such as the spacing between paragraphs, and the size of the margins around the text.

Table 1. Extracted design elements of websites

Elements		Type 1	Type 2	Type 3	Type 4	Type 5
X_1	Ratio of Graphics to Text	Above 3	Between 3-1	1	Between 1-1/3	Below 1/3
X_2	Blank Ratio	0% - 20 %	20% - 40%	40% - 60%		
X_3	Layout Style	2 columns	3 columns	Multiple columns		
X_4	Frame Style	Up and down	Left and right	Compound style		
X_5	Hyperlink Style	Only text	Text and symbol	Text and icon		
X_6	Number of Colors	Below 4 colors	4-7 colors	Above 7 colors		
X_7	Background Color	Cold color	Warm color	Neutral color		

Each design element has different types of its own, ranging from 3 to 5, as indicated by the type number 1, 2, 3, 4 or 5 in Table 1. For example, "Ratio of Graphics to Text (X_1)" means that the ratio between the area used by graphics and the area used by text on a website. This ratio takes 5 values, corresponding to 5 different types, including Above 3, Between 3-1, 1, Between 1-1/3, and Below 1/3. The design element "Hyperlink Style (X_5)" has 3 types, including Only text, Text and symbol, and Text and

icon, meaning that the hyperlink style is “only” text, or “adding” some symbols (e.g. small circles or square shapes without a specific meaning), or “adding” some icons (e.g. photographs or drawings with a specific meaning).

2.3 Experimental Website Samples

According to the result of morphological analysis, there are the 23 types of the 7 design elements in Table 1. If the website has a particular design element type, the value of the corresponding input is 1; otherwise, the value is 0. We then perform a cluster analysis on the 96 website samples in order to facilitate the assessment process to be made by the fourth group of subjects in the following section. With the generation of a cluster tree diagram, we extract 33 representative website samples, including 27 training samples and 6 test samples for training and testing the NN model to be developed, as shown in Fig. 1. To collect numerical data for performing numerical analysis, the degree to which the 33 website samples match a given set of user feelings has to be assessed.

2.4 Users’ Feelings of Website Samples

Kansei Engineering [11] has been successfully applied in the design field to explore the relationship between users’ feeling (Kansei) of an object (product or system) and the

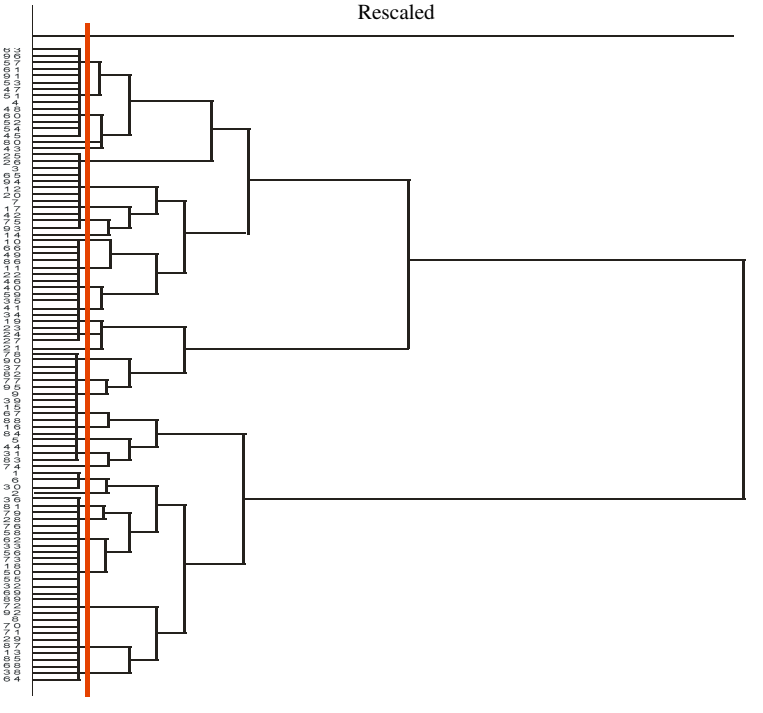


Fig. 1. Cluster tree diagram of 96 homepages of university websites

Table 2. The evaluation result for 33 representative website samples

Web Page No.	X_1	X_2	X_3	X_4	X_5	X_6	X_7	E-D value	C-C value	C-M value
1	2	2	2	3	2	2	2	2.28	1.70	2.90
2	3	3	3	3	1	1	2	2.11	2.16	2.52
3	3	3	2	1	2	3	1	2.63	2.59	3.06
4	5	3	2	3	2	1	3	3.26	3.48	2.52
5	3	2	3	1	3	2	1	2.84	2.57	3.06
6	1	1	2	1	3	3	3	2.44	2.48	3.78
7	3	3	1	2	1	1	1	2.47	2.42	2.00
8	5	2	2	3	2	1	3	2.78	3.16	2.64
9	1	1	1	1	3	3	3	3.27	3.11	4.42
10	1	2	1	2	1	2	3	2.60	2.28	3.52
11	3	3	2	1	2	1	1	2.77	2.33	2.68
12	4	1	2	2	1	2	2	2.64	2.63	2.99
13	2	2	3	1	1	3	3	2.27	2.33	2.68
14	2	1	3	2	1	2	3	2.73	2.31	2.21
15	4	1	2	1	3	3	2	2.89	2.63	2.33
16	4	2	3	3	3	3	1	2.67	2.78	2.84
17	2	3	2	3	2	1	3	2.26	2.00	2.91
18	4	1	2	3	2	3	2	2.68	2.59	2.12
19	5	2	3	2	1	2	1	2.96	2.90	2.31
20	5	3	1	2	1	1	3	2.34	2.28	1.87
21	4	2	2	3	3	3	2	2.37	2.26	3.67
22	1	1	2	3	3	2	3	2.70	2.49	3.42
23	3	2	3	1	1	2	3	2.73	3.00	2.90
24	4	3	3	1	1	2	2	3.15	2.89	2.47
25	5	3	3	3	3	1	2	4.11	3.36	2.36
26	2	2	3	3	2	1	1	3.02	3.07	2.31
27	1	1	3	3	2	3	1	3.06	3.26	3.25
28*	2	2	2	1	3	2	2	2.80	2.20	3.27
29*	2	3	1	2	2	1	3	2.27	2.27	3.00
30*	1	2	2	2	3	2	3	2.80	3.07	4.00
31*	3	3	1	1	1	3	1	2.53	2.47	3.20
32*	2	1	3	3	3	3	1	3.60	3.40	3.00
33*	2	1	3	3	2	3	2	2.80	2.87	3.27

design elements of the object [7, 8, 10]. A pair of feeling (or image) words is often used to describe users' feeling or perception about a specific image of an object. Based on our previous study [6], the first two groups of subjects extract 3 feeling word pairs for describing the feelings of websites, which were easy-difficult (E-D), clear-confusing (C-C), and classic-modern (C-M).

In this section, 30 subjects of the fourth group were asked to assess the degree to which each of the 33 website samples matches the E-D, C-C and C-M feelings respectively using 5 scales (1-5) of the semantic differential (SD) method [7, 8, 10]. For example, each subject assessed each website sample for the E-D feeling on a scale of 1 to 5, where 1 is very easy to use and 5 is very difficult to use. For the C-C feeling, the value of 1 represents the website is very clear to follow, and 5 represents the website is very confusing. For the C-M feeling, the value of 1 represents the website looks very classic, and 5 represents the website looks very modern.

The last three columns of Table 2 show the assessment results for the 33 experimental samples, including 6 test samples (asterisked). The E-D, C-C, or C-M feeling value shown is the average of the values assessed by the 30 subjects. For each representative website sample in Table 2, the first column shows the website number and Columns 2-8 show the corresponding type number for each of its 7 design elements, as given in Table 1. For example, No. 1 website is the clearest to follow (with a C-C value of 1.70), while No. 4 website is the most confusing (with a C-C value of 3.48). Table 2 provides a numerical data source for training and testing neural network models to be presented in the following section.

3 NN Models for Evaluating the Optimal Combination of Website Feelings

In this section, we develop neural network (NN) models in order to answer the research questions. With their effective learning ability, NN models are used to examine the complex relationship between input variables (design elements of websites) and output variables (feelings of users). In this study, we use the multilayered feedforward NNs trained with the backpropagation learning algorithm, as it is an effective and popular supervised learning algorithm [12, 13]. In addition, the learning rule used is Delta-Rule and the transfer function used is Sigmoid [13]. In order to examine whether the NN model is an effective technique for determining the optimal combination on website design for matching a desirable users' feelings, we develop four NN models for the training and test process. These four models are called the combined NN model, and three single NN models, i.e. the E-D NN model, the C-C NN model, and the C-M NN model, respectively, in our experimental study.

3.1 The Combined NN Model

To examine the relationship between the 7 key design elements of websites and the 3 users' feelings of websites, we develop the combined NN model with a single hidden layer. The 23 types of the 7 key design elements in Table 1 are used as the 23 input variables (neurons) for the combined NN model. If the website has a particular design element type, the value of the corresponding input neuron is 1; otherwise, the value is 0. The combined NN model is developed by combining all 3 users' feelings as 3 output neurons of the model, using the assessed average values of the E-D, C-C, and C-M feelings. In this paper, we apply the following four most widely used rules [13] to determine the number of hidden neurons in the single hidden layer for each model. Each model is associated with the rule used, such as -HN1, -HN2, -HN3, or -HN4.

(The numbers of input neurons + the numbers of output neurons) / 2 (1)
(The numbers of input neurons * the numbers of output neurons) ^ 0.5 (2)
(The numbers of input neurons + the numbers of output neurons) (3)
(The numbers of input neurons + the numbers of output neurons) * 2 (4)

Table 3 lists the neurons of the combined NN model and three single NN models (i.e. the E-D, C-C, and C-M NN models), including the input layer, hidden layer, and output layer. The 27 website samples in the training set, given in Table 2, were used to train the NN models. Each model was trained 5,000 epochs at each run. When the cumulative training epochs were over 100,000, the training process was completed. Table 4 shows the training epochs of each model run and their corresponding root of mean square (RMS) errors.

Table 3. Neurons of the combined NN model and the single models

The combined NN model	Input layer : 23 neurons for 23 types of 7 design elements. Output layer: 3 neurons for the E-D, C-C, and C-M feeling values.
-HN1 model	Hidden layer: 13 neurons, (23+3)/2=13.
-HN2 model	Hidden layer: 8 neurons, (23*3)^0.5=8.31≒8.
-HN3 model	Hidden layer: 26 neurons, (23+3)=26.
-HN4 model	Hidden layer: 52 neurons, (23+3)*2=52.
The single NN models	Input layer : 23 neurons for 23 types of 7 design elements. Output layer: 1 neuron for the E-D, C-C, or C-M feeling value individually.
-HN1 model	Hidden layer: 12 neurons, (23+1)/2=12.
-HN2 model	Hidden layer: 5 neurons, (23*1)^0.5=4.80≒5.
-HN3 model	Hidden layer: 24 neurons, (23+1)=24.
-HN4 model	Hidden layer: 48 neurons, (23+1)*2=48.

3.2 The Single NN Models

The single NN models are developed using each of 3 users' feelings (i.e. the E-D feeling, the C-C feeling , and the C-M feeling) as the output neuron individually. Like the combined NN model, there are 23 types of the 7 key design elements used as the 23 input neurons for each of three single NN models. The last 6 rows of Table 3 show the neurons of the single NN models, and Table 4 give the training epochs and their corresponding RMS errors.

As shown in Table 4, the lowest RMS error of the combined NN model (0.0104) is smaller than three single NN models (0.0542, 0.0591 and 0.0323, respectively). The result indicates that the combined NN model has the highest training consistency on the given users' feelings. The result suggests that the combined NN model is the better model for simulating users' feelings in this study. In addition, the RMS error of the combined NN model using the HN1 rule in (1) is the lowest (0.0104), as compared to the other three rules. The E-D NN model has the same result. However, the lowest RMS

Table 4. RMS errors of NN models for the training set

RMS errors	Learning epochs									
	10000	20000	30000	40000	50000	60000	70000	80000	90000	100000
The combined NN model										
-HN1	0.0355	0.0206	0.0140	0.0127	0.0119	0.0112	0.0106	0.0105	0.0104	0.0104*
-HN2	0.0372	0.0243	0.0170	0.0157	0.0147	0.0139	0.0132	0.0131	0.0130	0.0130
-HN3	0.0360	0.0222	0.0159	0.0147	0.0139	0.0132	0.0126	0.0125	0.0124	0.0124
-HN4	0.0457	0.0271	0.0168	0.0147	0.0133	0.0121	0.0111	0.0109	0.0109	0.0108
The E-D NN model										
-HN1	0.0634	0.0601	0.0591	0.0571	0.0565	0.0557	0.0548	0.0543	0.0542	0.0542*
-HN2	0.0634	0.0599	0.0586	0.0570	0.0565	0.0559	0.0553	0.0548	0.0548	0.0548
-HN3	0.0658	0.0607	0.0601	0.0579	0.0576	0.0572	0.0567	0.0562	0.0561	0.0561
-HN4	0.0671	0.0611	0.0599	0.0589	0.0588	0.0587	0.0586	0.0578	0.0578	0.0578
The C-C NN model										
-HN1	0.0787	0.0747	0.0732	0.0703	0.0692	0.0678	0.0660	0.0654	0.0652	0.0651
-HN2	0.0781	0.0732	0.0690	0.0659	0.0643	0.0623	0.0600	0.0594	0.0592	0.0591*
-HN3	0.0815	0.0755	0.0749	0.0726	0.0724	0.0721	0.0718	0.0711	0.0711	0.0711
-HN4	0.0842	0.0750	0.0740	0.0733	0.0731	0.0729	0.0726	0.0717	0.0717	0.0717
The C-M NN model										
-HN1	0.0671	0.0645	0.0642	0.0625	0.0623	0.0621	0.0619	0.0614	0.0614	0.0614
-HN2	0.0653	0.0577	0.0447	0.0409	0.0382	0.0356	0.0330	0.0326	0.0325	0.0323*
-HN3	0.0695	0.0647	0.0644	0.0626	0.0625	0.0623	0.0620	0.0614	0.0614	0.0614
-HN4	0.0667	0.0658	0.0652	0.0615	0.0612	0.0607	0.0601	0.0593	0.0593	0.0593

errors of the C-C NN model (0.0591) and the C-M NN model (0.0323) are those using the HN2 rule in (2). This result is in line with the notion that there is no best rule for determining the number of neurons in the hidden layer and it largely depends on the nature of the problem [7, 8, 9, 17]. To examine the performance of these NN models, we perform the test on all the models in the following section.

4 Performance Evaluation and Discussion

In order to examine if these four NN models can be applied to predict the users' feelings on a new website with a given set of design elements, 15 subjects of the fifth group were asked to evaluate 6 test website sample listed in the last 6 rows of Table 2. Rows 2-4 of Table 5 show the average values of the three uses' feelings on the 6 test samples evaluated by the 15 subjects, which are used as a comparison base for the performance evaluation. With the 6 test samples as the input, Table 5 shows the corresponding three feelings' values predicted by using the four NN models respectively. The last column of Table 5 shows the (lowest) RMS errors of the four models in comparison with the evaluated feelings' values. If there is no difference or error between the predicted values and the expected values, the RMS error is 0.

Table 5. Predicted feeling values and RMS errors of NN models for the test set

Website sample no.			28	29	30	31	32	33	RMS errors
Evaluated users' feelings	E-D value		2.80	2.27	2.80	2.53	3.60	2.80	
	C-C value		2.20	2.27	3.07	2.47	3.40	2.87	
	C-M value		3.27	3.00	4.00	3.20	3.00	3.27	
The combined NN model	-HN1	E-D	2.43	3.00	2.36	2.69	2.79	2.82	
		C-C	1.71	2.27	1.96	3.03	2.36	2.42	0.1868
		C-M	3.15	2.43	3.85	2.94	2.17	2.08	
	-HN2	E-D	2.59	2.86	2.39	2.59	2.89	2.80	
		C-C	1.72	2.18	1.90	2.74	2.30	2.37	0.1888
		C-M	3.44	2.09	3.92	2.82	2.09	2.00	
	-HN3	E-D	2.63	3.01	2.59	2.64	3.06	3.11	
		C-C	1.76	2.29	2.06	2.84	2.58	2.64	0.1607 *
		C-M	3.19	2.51	3.94	2.80	2.17	2.10	
	-HN4	E-D	2.59	2.54	2.60	2.72	2.79	2.89	
		C-C	1.67	1.80	1.94	2.98	2.54	2.50	0.1791
		C-M	3.38	2.52	4.20	2.46	1.98	2.21	
The E-D NN model	-HN1		2.59	3.29	2.88	2.00	2.54	3.09	0.1960
	-HN2		2.59	3.29	2.89	2.02	2.51	3.12	0.1991
	-HN3		2.60	3.28	2.86	2.00	2.55	3.10	0.1950 *
	-HN4		2.60	3.28	2.85	1.99	2.54	3.10	0.1968
The C-C NN model	-HN1		1.74	2.21	2.03	2.35	2.43	2.57	0.2102
	-HN2		1.75	2.17	2.01	2.44	2.42	2.63	0.2111
	-HN3		1.76	2.19	2.01	2.27	2.45	2.62	0.2100 *
	-HN4		1.75	2.19	2.00	2.26	2.45	2.62	0.2110
The C-M NN model	-HN1		3.25	2.78	4.51	2.29	2.17	2.38	0.1549
	-HN2		3.29	2.31	4.55	2.46	2.25	2.25	0.1642
	-HN3		3.26	2.75	4.50	2.30	2.20	2.39	0.1529 *
	-HN4		3.28	2.76	4.49	2.27	2.18	2.38	0.1551

As shown in Table 5, the C-M NN model using the HN3 rule in (3) has the lowest RMS error (0.1529). This suggests that the C-M NN model has the highest prediction ability in determining the best design combination of website elements. In addition, the RMS error of the combined NN model using the HN3 rule in (3) is 0.1607. This result indicates that the combined NN model has an accuracy rate of 83.93% (100%-16.07%) for predicting the values of the E-D, C-C, and C-M feelings of websites. This suggests that the combined NN model is a promising approach for modeling users' feelings of websites.

The results obtained from the NN models can be used to help web designers to work out the best combination of design elements for a particular design concept represented by a set of users' feelings. The web designer can focus on the determination of the desirable website feelings, and the NN models can help determine what combination of design elements can best match the desirable website feelings. To illustrate, Table 6 shows the ranking of the optimal design combinations for a set of website feelings represented by the E-D, C-C, and C-M values of 2. The best combination of design

Table 6. The optimal combination of design elements for the E-D, C-C, and C-M feelings of 2

Ranking	X_1 Ratio of Graphics and Text	X_2 Blank Ratio	X_3 Layout Style	X_4 Frame Style	X_5 Hyperlink Style	X_6 Number of Colors	X_7 Background Color	**Total value of E-D, C-C and C-M feelings**
1 (the best)	Between 3-1 (2)	0% - 20% (1)	3 columns (2)	Left and right (2)	Only text (1)	Below 4 colors (1)	Cold color (1)	2.06, 1.98, 2.07
2 (better)	Between 3-1 (2)	20% - 40% (2)	Multiple columns (3)	Left and right (2)	Only text (1)	Below 4 colors (1)	Neutral color (3)	2.09, 1.92, 1.98
3 (better)	Between 3-1 (2)	0% - 20% (1)	2 columns (1)	Compound style (3)	Only text (1)	Below 4 colors (1)	Neutral color (3)	2.08, 2.09, 1.97

elements given in Table 6 has its E-D value being 2.06, C-C value being 1.98, and C-M value being 2.07, which is the closest among all design combinations.

As shown in Table 6, there is a useful result that the optimal combinations of design elements have the same type on X_1 element (Ratio of Graphics and Text), X_5 element (Hyperlink Style), and X_6 element (Number of Colors), whose correspondent type are Between 3-1 (type 2 of X_1), Only text (type 1 of X_5), and Below 4 colors (type 1 of X_6) respectively. This result indicates that if website designers want to design a website with a little easy to use (the value of 2 using a 5-point scale), a little clear to follow, and a little classic for looking, they can consider the "Between 3-1" type of the "Ratio of Graphics and Text" element, the "Only text" of the "Hyperlink Style", and the "Below 4 colors" of the "Number of Colors". In other words, a website with the "Between 3-1", "Only text", and "Below 4 colors" characters will be perceived as a little easy to use, a little clear to follow, and a little classic by users. These design-supporting results provide useful insights in designing design elements of a website for reflecting the users' feelings of the website.

5 Conclusion

In this paper, we have presented an experiment study on university homepages to address the issue of the key design elements for an effective social website, based on user-centered interface design. The result of the study shows that the NN models can be used to help web designers work out the best combination of design elements for a particular design concept represented by a set of users' feelings. The web designer can focus on the determination of the desirable website feelings, and the NN models can help determine what combination of design elements can best match the desirable website feelings. Although the university homepages are chosen as an illustration, the approach is applicable to other kinds of websites with various design elements and other users' feelings. These design-supporting results provide useful insights in facilitating the user-centered web interface design process.

References

1. Ankolekar, A., Krötzsch, M., Tran, T., Vrandečić, D.: The Two Cultures: Mashing up Web 2.0 and the Semantic Web. Web Semantics: Science, Services and Agents on the World Wide Web 6, 70–75 (2008)
2. Buffa, M., Gandon, F., Ereteo, G., Sander, P., Faron, C.: Sweet Wiki: A Semantic Wiki. Web Semantics: Science, Services and Agents on the World Wide Web 6, 70–75 (2008)
3. Caplan, S.: Using Focus Groups Methodology for Ergonomic Design. Ergonomic 33, 527–533 (1990)
4. Fu, F., Liu, L., Wang, L.: Empirical Analysis of Online Social Networks in the Age of Web 2.0. Physica A: Statistical Mechanics and its Applications 387, 675–684 (2008)
5. Greaves, M., Mika, P.: Semantic Web and Web 2.0. Web Semantics: Science, Services and Agents on the World Wide Web 6, 1–3 (2008)
6. Guan, S.-S., Lin, Y.-C.: Building the Web Design System Using the Kansei Engineering. Journal of Design 17, 59–74 (2002)
7. Lai, H.-H., Lin, Y.-C., Yeh, C.-H.: Form Design of Product Image Using Grey Relational Analysis and Neural Network Models. Computers and Operations Research 32, 2689–2711 (2005)
8. Lai, H.-H., Lin, Y.-C., Yeh, C.-H., Wei, C.-H.: User-Oriented Design for the Optimal Combination on Product Design. International Journal of Production Economics 100, 253–267 (2006)
9. Lin, Y.-C., Lai, H.-H., Yeh, C.-H.: Neural Network Models for Product Image Design. In: Negoita, M.G., Howlett, R.J., Jain, L.C. (eds.) KES 2004. LNCS (LNAI), vol. 3215, pp. 618–624. Springer, Heidelberg (2004)
10. Lin, Y.-C., Lai, H.-H., Yeh, C.-H.: Consumer-Oriented Product Form Design Based on Fuzzy Logic: A Case Study of Mobile Phones. International Journal of Industrial Ergonomics 37, 531–543 (2007)
11. Nagamachi, M.: Kansei Engineering: A New Ergonomics Consumer-Oriented Technology for Product Development. International Journal of Industrial Ergonomics 15, 3–10 (1995)
12. Negnevitsky, M.: Artificial intelligence. Addison-Wesley, New York (2002)
13. Nelson, M.: Illingworth WT. A Practical Guide to Neural Nets. Addison-Wesley, New York (1991)
14. Nielsen, J.: Usability Engineering. Academic Press, United Kingdom (1993)
15. Nielsen, J.: Designing Web Usability: The Practice of Simplicity, New Riders, USA (1999)
16. Stephens, M., Collins, M.: Web 2.0, Library 2.0, and the Hyperlinked Library. Serials Review 33, 253–256 (2007)
17. Yeh, C.-H., Lin, Y.-C., Chang, Y.-H.: A Neural Network Approach to Website Design. Dynamics of Continuous, Discrete and Impulsive Systems: Series B, Applications and Algorithms (DCDIS Series B) 14(S2), 1598–1601 (2007)
18. http://www.bigresaerch.com/
19. http://www.cyworld.com/common/main.asp

Discovering Trends in Collaborative Tagging Systems

Aaron Sun[1], Daniel Zeng[1], Huiqian Li[2], and Xiaolong Zheng[2]

[1] Department of Management Information Systems, University of Arizona, Tucson, Arizona
[2] The Key Lab of Complex Systems and Intelligence Science, Institute of Automation, Chinese Academy of Sciences, China
{asun,zeng}@email.arizona.edu, {xiaolongzheng82,pluvious1981}@gmail.com

Abstract. Collaborative tagging systems (CTS) offer an interesting social computing application context for topic detection and tracking research. In this paper, we apply a statistical approach for discovering topic-specific bursts from a popular CTS - del.icio.us. This approach allows trend discovery from different components of the system such as users, tags, and resources. Based on the detected topic bursts, we perform a preliminary analysis of related burst formation patterns. Our findings indicate that users and resources contributing to the bursts can be classified into two categories: old and new, based on their past usage histories. This classification scheme leads to interesting empirical findings.

Keywords: burst, trend discovery, collaborative tagging.

1 Introduction

Being able to identify "hot" topics and emerging trends is in critical need in many application contexts (e.g., research, business, policy making). . In recent years, Collaborative Tagging Systems (CTS) [1] , as part of social computing and in paritcular application of social software, have gained significant popularity for their revolutionary ways of re-organizing information and helping form online communities. In CTS, conceptual descriptions in the form of collections of "tags" are assigned by registered users to some Web resources they have visited. In this paper, we analyze the emergence of topic bursts using data collected from a popular CTS - Del.icio.us. We first use a widely-adopted statistical technique to discover the topic-specific trends. Given the identified topic bursts, we then study their formation patterns by examining how different types of users have contributed to the dynamic formation of the trends.

This paper is organized as follows. In the next section, we briefly review previous studies on topic burst detection. In Section 3, we describe our data collection procedure. We present our topic bursts detection method in Section 4 as well as major empirical findings. In Section 5, we mainly focus on the formation patterns of the identified bursts. Section 6 concludes with a summary of our work and possible future research directions.

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 377–383, 2008.

2 Related Work

Analysis of temporal data has been an active topic of research for the last few years. Among various streams of related research activities, the area of Topic Detection and Tracking (TDT) [2] is concerned with discovering topically related material in textual materials. R. Swan and J. Allan proposed a χ^2 approach for extracting significant time varying features from news articles. The rapid development of Web technologies have presented many new challenges and opportunities for TDT studies Vlachos et. al. studied MSN search engine queries that arrive over time and identify bursts and semantically similar queries. E. Amitay et. al. [3] discussed using timestamps extracted from Web pages to approximate the age of the content with the primary goal of detecting significant events and trends. The underlying value of CTS as to TDT has also been noted in recent studies. S. Golder and B. Huberman [1] performed a systematic study on the structure of Del.icio.us as well as its dynamical aspects. They discussed the feasibility of discovering bursts of popularity in bookmarks and gave a simple example. A. Hotho et. al. [4] presented a PageRank-like algorithm to discover and rank the popular topics discovered in the user-tag-resource network environment of Del.icio.us.

3 Dataset

We collected data from Del.icio.us between Nov. 10 and Nov. 15, 2007 following the steps described below. We first chose a variety of topic keywords to narrow down the focus of interest. These keywords include: “game”, “movie”, “music”, and “book”, among others. For each keyword, we downloaded a complete list of Web resources that have been tagged by this keyword. We subsequently collected their individual tagging histories. Every time when a Web resource was bookmarked by someone, we are interested in such information as the user ID of the annotator, co-occurring tags, and date of tagging. At the end, we obtained a data set covering 20 categories, with 62,263,783 tagging activities captured in total. Each tagging activity is a vector consists of user/annotator ID, bookmarked URL, the tag assigned by the user, and the date the bookmark was created. Tags may be arbitrary strings. The tagging history of any resource r is recorded on a monthly basis from year 2003 to 2007, which provides sufficient data for us to capture the overall monthly trends.

4 Tag Burst Detection

4.1 Analytical Method

We characterize a tagging activity as a vector $a = \{(u, t, r, d)|u \in U, t \in T, r \in R\}$, where U represents the entire set of users, T is the set of all relevant tags (vocabulary), R is the entire collection of Web resources being annotated, and timestamp d records the date when the tagging activity occurs. By organizing

the entire set of tagging activities by month, we can readilyconstruct a data stream $ds(t) = (d_1, d_2, ..., d_i, ...d_n)$ for any tag t of interest, with data points d_i corresponding to different tagging dates.

In the next step, we use a simple statistical model, the χ^2 model model, inspired by previous work of R. Swan and J. Allan [2], to determine if the appearance of tag t in date d is significant. This model considers tag t to be generated by a random process with an unknown stationary distribution. (We are not concerned with the actual distribution here for simplicity.) In order to verify the validity of this stationary property, one can first build a 2×2 contingency table to characterize the presence and the absence of tag t. Specifically, let N denote the number of tagging activities that include tag t in month d, and $\overline{N}$ be the number of tagging activities without tag t in month d. The 2×2 contingency table then includes both measurements in month $d = d_0$ and $d < d_0$, respectively. Given this table, we can perform a χ^2 test with one degree of freedom to measure if the stationary assumption is violated. Statistically, for a χ^2 value of 7.879, there is a 0.005 probability that a feature from a stationary process would be identified as not being stationary. We thus adopt a threshold-based strategy: for any tag t under test having a χ^2 value higher than 7.879, we conclude that the hidden tag generation process has varied and therefore classify tag t as a "burst" tag of month d.

4.2 Tag Bursts

We began with an examination of tag popularities using one of the data categories - "game". The "game" category contains 1,395,453 tagging activities, in which 45,536 unique tags have been used. Among these tags, more than half of them have been only used once, and it's not surprising to observe that the tag occurrence frequency follows a long-tail distribution (Figure 1 (a)).

In order to efficiently find the burst tags, it is necessary for us to reduce the sample size before performing the χ^2 test. We preprocessed our samples based

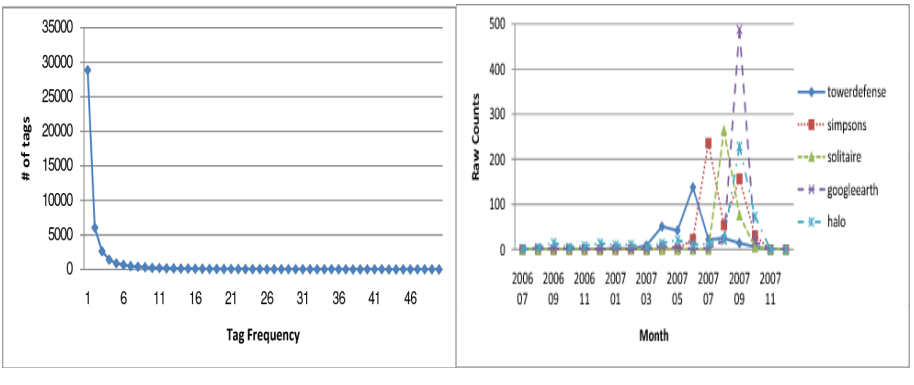


Fig. 1. (a)The distribution of tag frequency (b)Raw counts of the burst tags

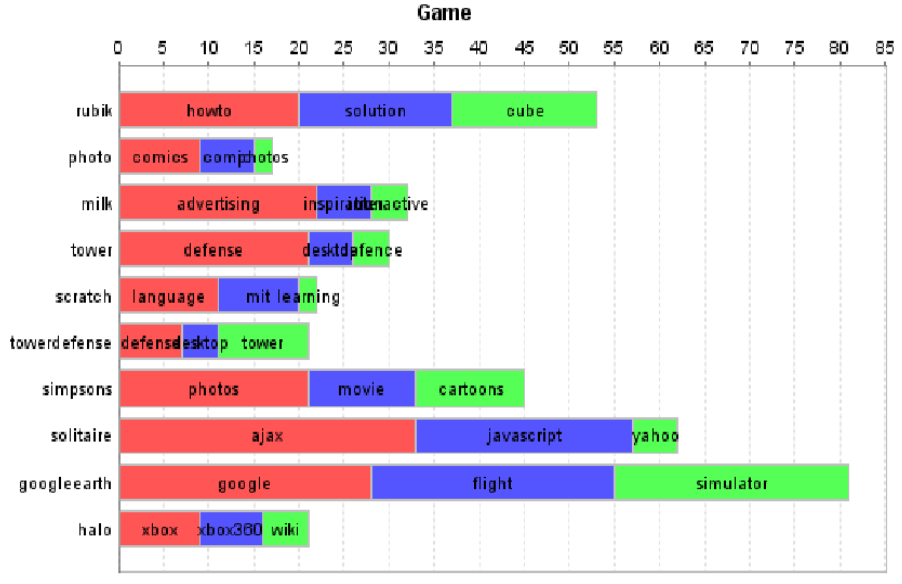


Fig. 2. The most significant tag in "game" category between Jan. to Oct., 2007

on previous studies [1] by first removing those infrequent tags whose occurrence frequencies are less than 10 per year. On the other hand, in practice, we found that some frequently occurred tags might also be reported as burst tags. For instance, in the "game" category, we have such examples as "game", "fun", and "cool", which are all top-ranked high-frequency tags. We believed that these tags had little value for us to understand the underlying tag formation and usage mechanisms. As such, the top 30 most frequently occurred tags in this category were discarded in our study.

We finally obtained a list of 1,472 candidate tags. Then starting from year 2006, we calculate the χ^2 value for each tag t in each month. If the value is above $7.879 (p < 0.005)$ we conclude that the appearance of t in that month is significant. Tag t can thus be considered as a burst tag of the month, and its χ^2 value indicates the intensity level of the burst. In Figure 1 (b), we plot the usage patterns of some detected burst tags. From the figure, it is evident that the burst periods are alike each other, having uniformly spike-shaped usage counts.

To understand the underlying events behind the identified burst tags, we performed the co-occurrence analysis aiming to study the relationship between burst formation and co-occurrence. We use a stacked bar chart in Figure 2 to illustrate our approach. The Y-axis of the chart shows the most significant tag (measured by χ^2 value) of each month from Jan. to Oct. 2007. The significance of tag t is represented by the length of the entire horizontal bar, which is equivalent to t's normalized χ^2 value. Each layer of the bar represents one of t's co-occurring pair, whose length also corresponds to its occurrence frequency. To save space,

we only draw three layers here - standing for the three most frequently co-occurred pairs. These co-occurred pairs can help us understand why some of the tags become popular. For instance, in the last week of Aug. 2007, Google released a new version of Google Earth, and this update included a fascinating hidden feature - a secret Flight Simulator - the reason that we saw Google's name extended to the area of gaming. Another example is the burst tag "halo" in Oct. 2007, which corresponded to the release of Halo 3 on Sep. 25, 2007, which is a popular Xbox 360 based game.

In the last step, we listed all the major URLs that have been annotated by the burst tags. They provide further clue about the meanings and usage of these (potentially ambiguous) tags. (Due to the page limit, we omit the URL details.)

5 Patterns of Burst Formation

Having identified a set of topic bursts, we now turn to the question of how these bursts are formed. The fundamental question we are concerned with is: when a certain tag is receiving increasing attention from users, how do these users contribute to the formation of the burst by various means? More specifically, do they create the trends by simultaneously introducing diverse information sources centered on a similar topic, or do they simply play the role of "trend-chasers" without bringing in new topics with them? In this section, we developed a simple model in an attempt to describe the browsing patterns of Del.icio.us users by making quantitative observations. We classify users and resources related to a certain tag t into two categories: *old* and *new*, based on their past usage history. For instance, *new* users U_{new} pertaining to tag t are defined as users who have not used t before date d. Likewise, *new* resources R_{new} pertain to tag t are resources having their first-time exposure to the public in d. New users turn into *old users* U_{old} when they keep on using tag t in the ensuing months. Similarly, if the *new* resources are mentioned repeatedly after d, they become *old* resources R_{old} as well. Note that the process from *new* to R_{old} is not reversible. Each user or resource can be set as *new* only once before it becomes old. We expect such a classification scheme could help us to answer the question posed above. Intuitively, if the majority of users tend to revisit their favorite topics, we will probably observe that the user population of the given burst tag contains more U_{old} than U_{new}. From the resource perspective, if we observe that the tag bursts result in higher R_{new} than R_{old}, we can conclude that it is more likely that users create the trend rather than follow it.

The classification scheme proposed above leads to stable patterns in which that bursts are dominantly contributed by new users. Empirically, we found that usually the old users only account for a small proportion of burst population (Figure 3 (a)). This stable pattern could be explained by the traditional social theory of fads/fashion [5]. In Del.icio.us, users have very little, almost zero, cost to bookmark other people's collections. The low cost of acquisition drives them to follow the preceding user's behavior blindly. For example, they are glad to bookmark those popular Web pages which are recommended by the system. As

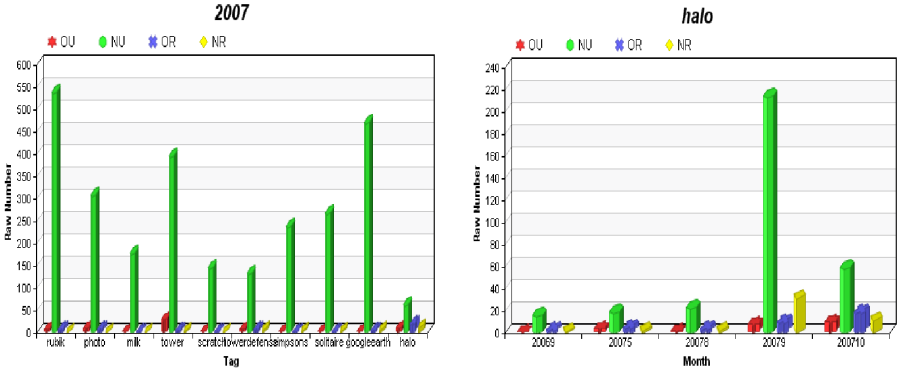


Fig. 3. User contribution to the tag bursts

is pointed out in [6], since these users do not make choices with regard to their own information, the caused mass behavior is often fragile in essence.

Another interesting finding is about the constitution of resource. As we have discussed before, the blindness of the bookmarking behavior determines that users tend to follow the trend. That can be testified by observing more R_{old} than R_{new} in the burst period (Figure 3 (a)). However, some exceptions exist. For instance, in Figure 3 (b), when Halo 3 was introduced to the public for the first time, multiple new resources were bookmarked as obviously they did not even exist before the release time. However, in the next month, the proportion of new resources shrinks to a level lower than that of the old resources.

6 Conclusions and Future Directions

CTS provide an interesting application domain for TDT study due to their large and active user base and frequent use of diverse tags. In this study, we propose to use a statistical approach to identify the topic bursts from CTS. Our approach has some methodological advantages: (1) It is not limited to web-pages, i.e., it is independent of the type of content that is tagged. (2) It is easy to implement and extend. Given the topic bursts identified, we examine the formation patterns of the bursts.

Our future work will involve more fine-grained analysis of tag bursts (e.g., on a daily basis) In addition, we also plan to explore the impact of the community structure on the burst dynamics.

Acknowledgements

This work is supported in part by NNSFC #60621001 and #60573078, MOST #2006CB705500, #2004CB318103, and #2006AA010106, CAS #2F05N01 and #2F07C01, and NSF #IIS-0527563 and #IIS-0428241.

References

1. Golder, S.A., Huberman, B.A.: Usage patterns of collaborative tagging systems. Journal of Information Science 32(2), 198–208 (2006)
2. Swan, R., Allan, J.: Extracting significant time varying features from text. In: Proc. of the 8th Intl. Conf. on Information and knowledge management, pp. 38–45. ACM Press, New York (1999)
3. Vlachos, M., Meek, C., Vagena, Z., Gunopulos, D.: Identifying similarities, periodicities and bursts for online search queries. In: Proc. of the 2004 ACM SIGMOD Intl. Conf. on Management of Data, pp. 131–142. ACM Press, New York (2004)
4. Hotho, A., Jaschke, R., Schmitz, C., Stumme, G.: Trend detection in folksonomies. In: Proc. First International Conference on Semantics And Digital Media Technology (2006)
5. Bikhchandani, S., Hirshleifer, D., Welch, I.: Learning from the behavior of others: Conformity, fads, and informational cascades. Journal of Economic Perspectives 12(3), 151–170 (1998)
6. Bikhchandani, S., Hirshleifer, D., Welch, I.: A theory of fads, fashion, custom, and cultural change as informational cascades. The Journal of Political Economy 100(5), 992–1026 (1992)

Socio-contextual Filters for Discovering Similar Knowledge-Gathering Tasks in Generic Information Systems

Balaji Rajendran

Centre for Development of Advanced Computing
68, Electronics City, Bangalore – 560 100, India
balaji@ncb.ernet.in

Abstract. The task of knowledge-gathering through an Information System has become increasingly challenging, due to the multitude of activities that are being facilitated through the system. A user-centric approach is presented in this paper where various activities executed by a user, for a knowledge-gathering task, are mapped to certain cognitive states in our model and the transitions between those states are used to indicate the progress made by the user. We propose a socio-contextual filtering algorithm for discovering similar tasks that were executed by other users and claim that such a socio-contextually related task would help in reducing the cognitive load, efforts and the time required for a user, naïve to a given knowledge gathering task. We demonstrate this through the fewer number of state transitions that occur in our model for a guided user.

Keywords: Socio-contextual Filters, Knowledge-gathering tasks, Social Information Systems, Socio-contextual Search, Task similarity, Social-spaces, Cognitive-load reduction.

1 Introduction

The process of knowledge-gathering through Information Systems has become a cognitively complex process, because of the numerous activities that are facilitated through the system and due to the diverse requirements of various users. This has demanded specialized searching and refining skills within the virtual spaces of an information system for effective accomplishment of the tasks by its users.

We present a user-centric approach to the problem, based on a social strategy to unlock the dormant knowledge of an underlying information system. We aim to utilize the concept of Social Capital [1] by discovering similar knowledge-gathering tasks executed by other users through the same information system, to help a user naïve to the current task. This approach would also help in bridging the structural holes [2] that would exist in a social information system catering to diverse users [3].

The rest of the paper is organized as follows: Section 2 highlights some of the work relevant to social and contextual techniques for knowledge-gathering. Section 3

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 384–389, 2008.

illustrates our model, the efficiency of our approach and the socio-contextual filtering algorithm. Section 4 mentions the details of implementation and the results obtained.

2 Related Work

The work of Wu et al. [4] states that the information flow in a social system happens only between those individuals who have the least social distance between them. This brings out the need to explicate the hidden knowledge in the system to other users. Recommendation systems have effectively utilized the socially networked systems to recommend a service or a product based on the collective actions of the users. Upendra et al., [5], describe a technique for personalized recommendations based on user-profiling and also state an information filtering mechanism in their work.

The importance of context in knowledge sharing is emphasized in the work of Zhu et al., [6], who present a model for representing the contexts and a method for measuring the similarity between the contexts. The need for contextualizing the search performed by users through global search engines like Google [11] and personalizing the search results based on the contexts of the user has been attempted by Paul et al. [7] in their work. A personalized contextual information retrieval model based on an extension of Bayesian Network is presented by Zemirli et al., [8].

A large-scale evaluation of the collaborative web search techniques, by Barry Smyth et al. [9] indicate that the exploitation of repetition and regularity in query spaces could provide significant benefits to users. We aim to exploit such repetition in knowledge-gathering tasks executed by many users within the Information System.

3 Socio-contextual Filtering: Model and Algorithm

A user naïve to a concept or a domain, to accomplish a knowledge-gathering task in that domain, has to forage for the right sources within the accessible information spaces of a system. This is a cognitively complex process, as the user has to choose the right resources among heaps of resources in an information system. Our aim is to guide such users by indicating the most probable sources, derived from the activities executed by other users involved in similar tasks of knowledge gathering.

First, we define the terms related to our work, and describe a model that captures the various states of a user, involved in a knowledge-gathering task and examine the effectiveness of our approach through the model. Second, we describe the socio-contextual filtering algorithm for discovering similar knowledge-gathering tasks.

3.1 Definitions

Task: A task T is defined by a set of attribute-value pairs $\{a_1{:}v_1, a_2{:}v_2 \ldots a_n{:}v_n\}$. A task T is accomplished by a user through a set of activities $\{A_1, A_2 \ldots A_n\}$ that can be performed through the information system. The attributes $a_1, a_2 \ldots a_n$ define a task, while the values $v_1, v_2 \ldots v_n$ reflect the preferences of the user performing the task. The knowledge and approach of the user determine the activities and their sequence followed for the accomplishment of the task.

$$T = \{a_1{:}v_1, a_2{:}v_2 \ldots a_n{:}v_n\};\ T \Rightarrow \{A_1, A_2 \ldots A_n\}$$

Similarity of Tasks: We look for similarity between two tasks T_i and T_j, only if they are conceptually similar, i.e., only if there is a one-to-one correspondence between the attributes of T_i and T_j. An ideal case of similarity would then be: each of the attribute-value pairs of T_i exactly matching with one of the attribute-value pairs of T_j. If a subset of the attribute-value pairs of T_i match with either the subset or the full set of T_j then, we assume them to be partially similar.

Knowledge-gathering Task: A task that requires execution of search-and-find kind of activities for its accomplishment. The typical activities involved in a knowledge-gathering task are: searching for sources of information, filtering and choosing the sources, and gathering information from the chosen source(s). Hence, all such activities can be classified under: Searching, Filtering, and Gathering.

Social Space: The Social Space SS, of an information system is defined as the collection of tasks accomplished by its various users. It can be represented as:

$$SS = \{U_1 \rightarrow T_1 \ldots. U_n \rightarrow T_n\} \text{ with each task accomplished as } T_i \Rightarrow \{A_1 \ldots A_n\}$$

Socio-contextual Filters: These filters are meant for discovering highly relevant tasks from the heap of tasks in the social space SS. We use Time and Space, as the top-level filters along with other Socio-contextual filters.

Time Filter: The dimension of time is used to limit and select only those tasks that were accomplished recently from the given social space SS.

Space Filter: The dimension of space, in the form of regions is used to select those tasks that were accomplished by users from a specific geography.

Other Filters: In real-life situations, humans implicitly apply socio-contextual filters, while gathering information. Such filters are in the form of occupation, educational qualification etc… that can be used for getting to the more relevant tasks.

3.2 Model and Approach

Figure 1 represents the typical cognitive states of the user, involved in a knowledge-gathering task. These states are obtained by mapping the various activities performed by the user, during the accomplishment of a task. The transitions indicate the progress of the user in the task.

A user, to accomplish a new knowledge-gathering task, starts with the Q0 State. As the user begins to forage for sources to gather knowledge, he enters the search state (S). The user enters the Filter state (F), as he selects the appropriate sources from a list of sources. A user then utilizes a source to gather the required information for accomplishing the task, in the Gather state (G). This can include activities such as raising questions in forums, or in a chat room, or reading through a page. After gathering the required knowledge, the user reaches the Decision state (D), the end state of the knowledge-gathering task. We assume that the activities performed by the user correspond to only one given task at a time.

As it can be observed, there are many possible transitions from a given state, and a user might not proceed linearly as described above. Typically, a user oscillates between the Search – Filter – Gather states, until he is able to gather the required information. However, transitions can happen from Q_0 to G, when a user has prior

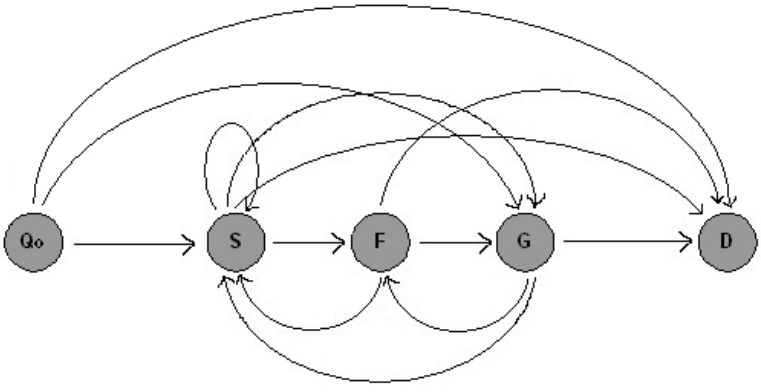


Fig. 1. Cognitive States of the user involved in a knowledge-gathering task

knowledge of a source to gather the information for his task. A transition from Q_0 to D indicates that the user has complete knowledge about the task. A user might also decide to exit the task, without successfully accomplishing it.

Approach: We use the conceptual and socio-contextual filters to discover relevant tasks from the Social Space SS, and determine the similarity of each task with the task to be executed by the current user. The activities performed by other users in those similar discovered tasks are analyzed using our model. Among the transitions to the Gather state, G, we examine those activities that were performed by the user, in the longest state G`, as this would be the key to accomplish the task. The sources that are discovered through this approach are presented to the user, involved in a similar task.

Effectiveness of the Approach: Consider a knowledge-gathering Task T_i, to be executed by user U_1, who has little or no knowledge about the task. Assume a task T_j executed by user U_2, that is conceptually and contextually similar to Ti discovered through the algorithm (discussed in the next section) from the social space SS.

Let's assume that U_2 accomplished T_j as: T_j => {A_1, A_2, A_3, A_2, A_3, A_1, A_2, A_3, A_4}, which can be mapped to the cognitive states of the user as: S => {A_1}, F => {A_2}, G => {A_3, A_4} and their transitions during the accomplishment of T_j would then be: U_2 T_j => {Q_0, S, F, G_1, F, G_2, S, F, G_3, G_4, D}. From the gather states G_1 to G_4, let us assume that the transition to G_3 is where the user spent most of his time, during the accomplishment of the task. Then the activities performed in the state G_3 by the user U_2 becomes the guide for User U_1.

Without prior knowledge, the execution of task by U_1 would be: U_1T_i => {Q_0, S, F, G……}. With guided inputs, the execution of task by U_1 would be: U_1T_i => {Q_0, G…} thereby resulting in fewer transitions. However, there may be gaps in the knowledge space of User U_1, who then might enter the Search states as U_1T_i => {Q_0, G_1, S, F, G_2, S, S ….}. The activities executed in the subsequent Gather state will also be learned and put up to the next new user of a similar task. As the learning improves, better guidance can be given to the user, resulting in fewer transitions for the user.

The best-case scenario for our approach would be the guidance that transitions the user U_1 from the Q_0 state to the Gather states G' (G_1... G_n) directly and to the last state

D as: U_1T_i => {Q_0, G', D}. The worst-case scenario occurs, if the user has large gaps in his knowledge required for the task, or when our algorithm gives a dissimilar task, leading to a large number of transitions such as: U_1T_i => {Q_0, G_1, S, S, F, G_2, F, G_3, S, F, G_4, S, F, G_5 ….. D}.

3.3 Socio-contextual Filtering Algorithm

Input: Given a knowledge-gathering task T with the social and contextual attributes and value ranges of interest to be executed by a user U, along with a set of conceptually similar tasks extracted from the Social Space SS.

Output: A list of tasks that is socio-contextually relevant and similar to T.

Assumptions: a) Every task in SS is associated with a user and all the activities performed by the user for its accomplishment is logged;

b) A user would not be involved in more than one task at the same time.

```
Task[] Socio-Contextual_Filters(ConcepFilteredTasks[], T) {
  for each task Tj in ConcepFilteredTasks[] do {
          /* Time Filter */
    if the time of accomplishment of Tj is within the
threshold-level t {
          /* Region Filter */
    Get the full user profile of Ux, who had accomplished Tj;
    if the region of Ux, is within the range of interest to T,
then {
          /* Other Social Filters */
    if the values of the relevant social parameters of Ux is
within the range of interest to T, then {
             add Tj to SociallyFilteredTasks[];
} } } }
return SociallyFilteredTasks[];
}
```

4 Implementation and Results

A Social Information System [10] that facilitated many activities for its users was used for implementing our approach. Each possible activity facilitated through the system was categorized into one of the states in our model. For example, an activity of a user such as reading a news item or viewing a multimedia clip was categorized to be in Gather state. The activities of each user in a session, from login to logout are mapped to the corresponding states. As the tasks emerge, we add those tasks containing activities in their search and gather states to the Social Space SS, of knowledge-gathering tasks. The Socio-contextual Filtering algorithm then chooses the relevant similar tasks, and the sources used in the activities performed in the longest Gather state, of each user are collected and organized to guide the current user.

We studied the benefit of our approach, by measuring the number of state transitions that the users went through. The results showed that most of the guided users were saved at least 25% of transitions that unguided users went through. Also, the Socio-contextual filtering algorithm when applied in conjunction with conceptual filtering was effective to 70% of the users of the system.

5 Conclusion

We used the social capital [1] to unlock the dormant knowledge of an information system to assist the users in their knowledge-gathering tasks. The user-centric approach to the problem brought out the cognitive states that a user went through, for executing a knowledge-gathering task. The Socio-contextual filtering algorithm along with conceptual filtering helped in discovering highly relevant and similar tasks that were executed by other users in order to guide the current user.

Acknowledgment. I thank Dr. K. Iyakutti for his support and guidance. I also thank Mr. N. Subramanian, C-DAC and my team-members who supported me during this work.

References

1. Coleman, J.S.: Social Capital in the Creation of Human Capital. The American Journal of Sociology, Supplement: Organizations and Institutions: Sociological and economic Approaches to the Analysis of Social Structure 94, S95–S120 (1988)
2. Burt, R.S.: Structural holes and good ideas. The American Journal of Sociology, 110 (in press)
3. Cummings, J.N.: Work groups, structural diversity, and knowledge sharing in a global organization. Management Science 50(3), 352–364 (2004)
4. Wu., F., Huberman, B.A., Adamic, L.A., Tyler, J.R.: Information flow in social groups. Physica: Statistical and Theoretical Physics 337(1-2), 327 (2004)
5. Shardanand, U., Maes, P.: Social information filtering: algorithms for automating "word of mouth". In: Proceedings of the SIGCHI conference on Human factors in computing systems (1995)
6. Zhu, X., Pan, X., Wang, S.: Approaches to Context-based Knowledge Share and Reuse. In: Proceedings of the Fourth International Conference on Fuzzy systems and Knowledge Discovery, vol. 1, pp. 741–746 (2007)
7. Chirita, P.A., Firan, C.S., Nejdl, W.: Summarizing Local Context to Personalize Global Web Search. In: CIKM 2006, pp. 287–296 (2006)
8. Nesrine Zemirli, W., Tamine-Lechani, L., Boughanem, M.: A Personalized Retrieval Model based on Influence Diagrams. In: Proceedings of the 2nd International Workshop on Context-Based Information Retrieval (2007)
9. Smyth, B., Balfe, E., Boydell, O., Bradley, K., Briggs, P., Coyle, M., Freyne, J.: A live-user Evaluation of Collaborative Web Search. In: Proceedings of International Joint Conferences on Artificial Intelligence, pp. 1419–1424 (2005)
10. Rajendran, B., Venkataraman, N., Murugesan, P., Tandel, K.: Establishment of Community Information Network in a Developing Nation. In: Proceedings of the IEEE Tencon 2005 (2005), http://ieeexplore.ieee.org/iel5/4084859/4084860/04085345.pdf?isnumber=4084860&prod=CNF&arnumber=4085345&arSt=1&ared=6&arAuthor=Rajendran%2C+B.%3B+Venkataraman%2C+N.N.%3B+Murugesan%2C+P.%3B+Tandel%2C+K.
11. Google: http://www.google.com

Exploring Social Dynamics in Online Bookmarking Systems

Xiaolong Zheng[1], Huiqian Li[1], and Aaron Sun[2]

[1] The Key Lab of Complex Systems and Intelligence Science, Institute of Automation, Chinese Academy of Sciences, China
[2] Department of Management Information Systems, The University of Arizona, USA

1 Introduction

Web 2.0 technologies have spawned different types of information sharing systems, including online bookmarking systems. These information sharing systems have facilitated collaboration among their users with similar interests. They also provide a powerful means of sharing, organizing, and finding contents and contacts [1]. In this paper we focus on evaluating social interaction among users on *Del.icio.us*, which is one of the most popular and paradigmatic online bookmarking systems.

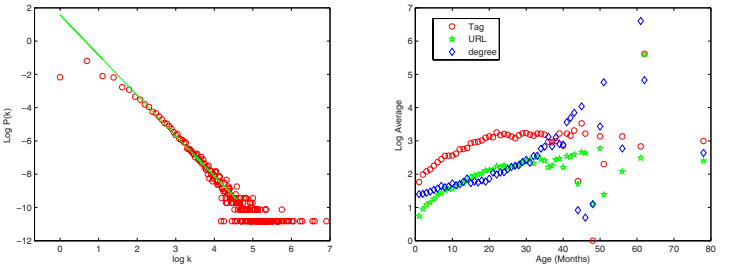
2 Degree Distribution

In *del.icio.us*, the collaborative tagging is globally visible among users. The process of tagging develops genuine social aspects, and the tagging system demonstrates social dynamics of user activity. We have designed website crawlers to collect the data and divided the dataset into several smaller datasets such as blog, finance, book, etc. Our analysis in this section was performed on the blog dataset which consists of 50,190 users. Fig. 1(a) shows the degree distribution of the user network, whose nodes are users and edges represent their social relationships. From this Figure, we can find that the network generally exhibits a power-law degree distribution: most of the nodes have small degree, and a few nodes have significantly higher degree. To test how well the degree distribution is modeled by the power-law, we used the least square method to fit the power-law behavior. We conclude that the distribution function $P(k)$ and degree k have the following approximate relation, $P(k) \sim k^{-2.4209}$.

3 User Activity Evolution

To analyze social dynamics, we use statistical approaches to identify the global characteristics of users in the tagging process. Fig. 1(b) displays the evolution of three variables: the average number of tags, URLs, and the average degree. We describe these variables as functions of age, measured by the final time we collected the data minus the time users first participated in the *Del.icio.us*. We can

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 390–391, 2008.



(a) Degree distribution (b) User activity evolution

Fig. 1. Degree distribution and user activity evolution

characterize the evolution into two processes. In the first process, while users' age ranges from 1 to about 40 months, the average number of tags (*circle*) and the average number of URLs (*pentacle*) increase sublinearly with users' age. The average degree (*diamond*) increases similarly but superlinearly. However, in the second process when users' age is larger than 40 months, these relations disappear. This suggests that there exists a multi-scaling behavior in the tagging activity of users. When new users participate in the *Del.icio.us*, they will post some interesting web pages and use many tags to categorize and describe them. In the meanwhile, they will search some friends with common interests on *Del.icio.us*. As time goes by, some users exhibit stable interests in a consistent way while others not. This causes the numbers of their tags, URLs, and friends show different patterns.

4 Conclusions

This paper investigates the degree distribution of *Del.icio.us* user network and explores user activity evolution. We observe that the degree distribution exhibits the power-law property and the tagging activity shows an interesting multi-scaling behavior. Our current work is concerned with in-depth empirical analysis of these phenomena and related modeling work.

Acknowledgments

This work is supported in part by NNSFC #60621001 and #60573078, MOST #2006AA010106, #2006CB705500 and #2004CB318103, CAS #2F05N01 and #2F07C01, and NSF #IIS-0527563 and #IIS-0428241.

References

1. Mislove, A., Marcon, M., Gummadi, K.P., Druschel, P., Bhattacharjee, B.: Measurement and Analysis of Online Social Networks. In: IMC 2007, San Diego, California, USA (2007)

Dispositional Factors in the Use of Social Networking Sites: Findings and Implications for Social Computing Research

Peter A. Bibby

School of Psychology
University of Nottingham,
University Park,
Nottingham, NG7 2RD, UK
pal@psychology.nottingham.ac.uk

Abstract. The paper presents findings of a study that relates dispositional factors such as extroversion, stability, self-esteem and narcissism to the use of social networking sites (SNSs). Each of these dispositional factors is shown to be related to different types of usage of SNSs. It is argued that attempts to model the use of SNSs and thereby target particular information to particular users would benefit greatly from using modeling techniques that can parameterize such dispositional factors.

1 Introduction

Social computing takes computational approaches to studying social interactions on the web, using mobile phones, texting and other technological interventions that facilitate social interactions either proximally or distally. It embraces insights from outside computer science including engineering, physics, human computer interaction and the social sciences such as sociology, economics and psychology. The current paper presents a case for incorporating a particular aspect of psychological research, that is, personality research. It has been argued that people are not passive users of technology, but rather actively shape the technology [1], [2]. How people shape the technology may well be at least in part determined by their personalities.

Take, for example, two personality characteristics: extroversion and neuroticism. Extroverts are outgoing friendly people who seek excitement, are impulsive and take risks whereas an introvert is reflective, is happy in their own company, and avoids large social events [3]. At the same time neurotics are emotionally unstable anxious and worrying. It would be unusual if these characteristics did not influence the use of social networking sites such as MySpace, Friendster or Facebook. Hamburger and Ben-Artzi found that both extroversion and neuroticism are related to different patterns of internet use for both men and women. For men, the use of internet-based leisure services was positively related to extroversion but neuroticism predicted their level of information service use. On the other hand, for women, social site usage was predicted positively by neuroticism and negatively by extroversion [4].

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 392–400, 2008.

McElroy, Hendrickson, Townsend, & DeMarie [5] recently examined the relationship between the Big Five [6] personality constructs of agreeableness, conscientiousness, extroversion, stability and openness to experience. Agreeable people tend to help other people, to be sympathetic, are good natured and tolerant of other people. Conscientiousness is related to the propensity to be strong-willed, reliable, organized and self-disciplined. Extroversion in this personality theory relates to sociability, cheerfulness and seeking excitement. Stability is essentially the opposite of neuroticism and represents appropriate psychological adjustment, emotional stability, reduced anxiety and worry. People who are open to experience are curious and willing to explore new ideas. McElroy et al [6] found, for example, that neuroticism is positively related to the frequency of e-selling. This is consonant with the work of Amiel and Sargant [7], and may reflect the need to avoid the anxiety associated with face-to-face communications often associated with neuroticism. The findings also suggest that openness to experience is related to information seeking, chatroom and bulletin board usage. This is not dissimilar to the findings of Tuten and Bosnjak [8].

A related dispositional factor is self-esteem. Harter [9] has argued that acceptance by ones peers and feedback on oneself are important determinants of self-esteem. They are also major features of social networking sites. Studies have found a negative relationship between internet usage and self-esteem. Kraut, Patterson, Lundmark, Kiesler, Mukopadhyay and Scherlis [10], in a longitudinal study of internet use, found that greater use predicted a decline in the size of the user's social circle and self-esteem and an increase in loneliness and depression. There is also research that demonstrates a positive relationship. Kraut, Kiesler, Boneva, Cummings, Helgeson and Crawford [11] report an increase in social well being for extraverts implying a "rich get richer" model of internet usage.

A third aspect of personality that may well affect the use of social networking sites is narcissism. Raskin and Terry [12] further developed a measure of narcissism; the Narcissistic Personality Inventory (NPI). Kubarych, Deary & Austin [13] have argued that the subscales of the NPI are best understood as representing global factors of power/control and exhibitionism. People who score high in the power/control domain tend to feel that they are in charge of their lives, that they can manipulate other people, and that they do not need to rely on other people. Those who score high on exhibitionism are vain, like being the center of attention and feel that they are entitled to other people's attention. Given the remarkable ability to self disclose that is made available to user's of social networking sites, it would be surprising if the exhibitionism component of the narcissism did not relate to usage of such sites. Furthermore, as many of these sites allow users to interact with other users pages, it may well be that the power/control dimension of the scale is also related to the frequency and type of usage of such sites.

A final factor, which is not directly a measure of personality, but may reflect aspects of personality that are not captured by the Big Five, Self-Esteem and Narcissism, is the quality of the user's social support network. Bargh and McKenna [14] point out that the internet can provide productive ground for the development of friendships or close relationships through the common interests and values of the users. Furthermore, they argue that given there are often no equivalent offline groups so virtual groups can become central to one's social support systems and networks. In

some cases, it is possible that such virtual support systems can be more extensive and effective than offline social support networks. The extent to which these social support networks are primarily online or offline is therefore a likely determining factor in the use of SNSs.

Apart from the obvious implication that a user's personality may influence a user's interaction with social computing systems, why should the social computing community be interested? At least in part, the answer can be drawn from a recent paper by Sun [15]. He argued that in order to successfully model social systems it is important to use psychologically oriented cognitive architectures rather than purely engineering based architectures. There are three primary reasons: first they are more realistic and may behave in more humanlike ways; second they further our understanding of human cognition; and third and most importantly in this context, they may provide a base for understanding collective human activity. Personality variables can be incorporated into such systems through the appropriate parameterization of different aspects of the architecture's behavior.

For example, Sun [15] presents a model, CLARION, that is comprised of four components: the action-centered subsystem (ACS), the non-action centered subsystem (NACS), the motivational subsystem (MS) and the metacognitive subsystem (MCS). For the purposes of this paper, neither the ACS or NACS are directly relevant, but both the MS and MCS are. The MS provides the goals and drives for the cognitive architecture and the MCS is responsible for the reinforcement, goal setting and regulation of the system. These are aspects of a human that are directly reflected in the personality of that human. For example, extraverts seek out social company whereas introverts tend to avoid it. The conscientious may have particular information seeking goals whereas those who score high in openness to experience may seek out new information or leisure pursuits. Thus, the goals and drives of the user vary with respect to personality. Not only is personality important at the motivational level but self regulation and reinforcement are similarly influenced by personality. Those who are high in self-esteem are less likely to be affected by a poor experience of a SNS than those who are low in self-esteem. Negative reinforcement of social activity is more detrimental to people not high in self-esteem. Furthermore, narcissists can simply disregard anything that does not fit with their self-aggrandised view of themselves.

The remainder of this paper presents the results of a survey of the use of SNSs alongside the measurement of different personality constructs including the Big Five, Self-Esteem, Narcissism and Social Support. How these findings can be used to further social computing research will be discussed later.

2 Method

2.1 Participants

A convenience sample of 174 undergraduate and postgraduate students in a British university participated (mean age 20.53, sd=1.95). Participants were volunteers who were not paid for completing the questionnaire. Participants were asked about their current relationship status: 80 were single or dating and 94 were in a relationship.

There were no statistically significant differences between males and females on any of the SNS questions. There was only one statistically significant difference between those who were single or dating, with those who were single reporting use of SNSs for romantic purposes more often.

2.2 Measures

The questionnaire given to the participants included the following aspects of their use of SNSs and their personality.

Using SNSs: A set of 22 statements concerning use of SNSs was developed. Example items include "To keep in touch with people I have met on line", "to pass time when bored", "to keep in touch with friends and family who live far away", and "to communicate with people who I am romantically interested in". Seven responses were available varying from "more than 2 or 3 times per day" to "less than once a month". A higher score is associated with more frequent use of the social network site. The items were factor analyzed and three clear factors comprising 11 items emerged. The first factor included 4 items such as "To pass time when bored" and "To entertain myself". It concerns occupying a person's free time (Cronbach's alpha=.83). The second factor included 5 items such as "To post videos that I have created", "To find others with the same interests," and "To learn about new music". They all involve a participant's interests (Cronbach's alpha=.74). The final factor included 2 items; "To keep in touch with people you don't have time to see in person" and "To keep in touch with friends or relatives who live far away". These items refer to keeping in touch. (Cronbach's alpha=.72).

Rosenberg's Self-Esteem Scale[16]: This is a 10 item Likert scale with items answered on a four point scale (from strongly agree to strongly disagree). The higher the score the higher the self esteem. Cronbach's alpha for the current sample is .89.

Functional Social Support: This is an adaptation of the Duke-UNC functional social support questionnaire [17]. It is a 7 item scale with 5 possible responses ranging from "As much as I would like" to "Much less that I would like". A high score indicates strong social support. Participants were asked to complete the scale twice, first with respect to online friends and second with respect to offline friends. Cronbach's alpha for both the online and offline friends was .88. A difference score was calculated by subtracting the online score from the offline score. A negative score indicates stronger social support online and a positive score indicates stronger social support offline.

Narcissistic Personality Inventory: The 40 item version of the NPI was used [12]. Statements such as "I have a natural talent for influencing people" or "I think I am a special person" were responded to on a 5 point scale varying from strongly disagree to strongly agree. Following, Kubarych, Deary & Austin [13] two measures were derived from the scale; a measure of the power component of the scale (Cronbach's alpha=.85) and another for the exhibitionism component (Cronbach's alpha=.86). A high score indicates greater degree of the narcissism personality component.

The Big Five: Goldberg's [18] bipolar adjective checklist was used to measure five personality traits: extroversion, agreeableness, conscientiousness, stability and openness to experience. Participants indicated on a 9 point scale the extent to which they were best described as belonging to one end of the adjective pair or the other. Cronbach's alphas were extroversion=.85, agreeableness=.87, conscientiousness=.83, stability=.85, openness=.72. The higher the score the stronger the trait.

3 Results

The average responses to the 22 item social network questionnaire are shown in Table 1. For the three factors the most common type of activity is simply to keep oneself occupied. Respondents, on average, reporting using a social network site two or three times per week. On the other hand, the maintaining interests factor shows that this kind of activity only occurs on average only once per month. In terms of keeping in touch with family and friends at a distance this occurs a little more than once per week. For the items that did not load onto clear factors, keeping in touch with friends is the most frequently reported activity, occurring on average two or three times per week. The least frequent activity is keeping a blog, with this occurring infrequently, as little as less than once per month.

The next question is to what extent dispositional factors predict the self-rated frequency of social network site use. To do this a series of hierarchical regressions were calculated. The order of entry of the dispositional factors into the regressions was decided a priori to reflect the generally accepted importance of these factors in the psychology literature. First, the Big Five factors, extroversion, agreeableness, stability, conscientiousness and openness to experience were entered. At the next step, Rosenberg's self-esteem measure was entered. The two narcissism components, power and exhibitionism, were next entered. Finally the social support score was entered.

For the occupy oneself (when bored) SNS factor 8.4% of the variability in the frequency of doing activities on a SNS was accounted for by the Big Five factors with stability being the primary predictor. A negative relationship was found with stability (β=-.241) suggesting that as a user's neuroticism increases then they are more inclined to use the SNS to occupy their time. A further 2% of the variability was accounted for by the Narcissism factors with Exhibitionism showing a significant positive relationship (β=.161). The higher the Narcissism score the more likely the user is to use the SNS when bored or having nothing better to do. Finally, an additional 3.5% of the variability was accounted for by the social support score. This was a negative relationship (β=-.199) suggesting that the more satisfied the user is with their offline social support network the less likely they are to turn to the SNS when they have nothing else to do.

For the leisure interests factor, 15.2% of the variability was explained by the Big Five factors with extroversion and conscientiousness factors being the main predictors. Conscientiousness showed a stronger relationship, which was negative (β=-.450), implying that more conscientious users spend less time seeking information about their leisure interests. The extroversion factor showed a positive relationship

Table 1. The mean frequency ratings (and standard deviations) for the three SNS factors and the remaining unfactored SNS items

Social Networking Site Factors	*Mean**	*Sd*
Occupying time	5.08	1.51
Maintaining interests	1.77	.97
Keeping in touch over a distance	4.47	1.42
Other Social Networking Site Items		
To keep in touch with friends	5.35	1.53
To look at photographs others have taken	4.79	1.74
To give or receive information with people you know	4.74	1.74
To post pictures that I have taken	3.28	1.71
To communicate with others in whom I interested romantically	2.75	1.95
To watch uploaded videos	2.74	1.74
To share ideas and opinions	2.72	1.78
To publicize events that I am holding	2.60	1.72
To help others	2.58	1.68
To keep in touch with people I've met online	1.66	1.44
To keep a blog	1.48	1.19

* 1=less than once per month, 2=once per month, 3=two or three times per month, 4= once per week, 5=two or three times per week, 6=once a day, 7=more than once per day.

(β=.202) suggesting that more time is spent maintaining leisure pursuits on the SNS by the more extroverted individuals. Again the narcissism factors explained a significant additional proportion of the variability, 3.7%, with the exhibitionism component showing a positive relationship (β=.223). This implies that the more exhibitionist a user is the more they seek out their leisure pursuits on line.

There were no significant relationships between any of the dispositional factors and the keeping in touch factor of the SNS questionnaire. Only three of the remaining eleven SNS usage questions achieved statistical significance: 'To communicate with people who I am interested in romantically', 'To show others encouragement' and 'To publicize events that I am holding'. For the first of these, 12.6% of the variability was accounted for by the Big Five factors with stability and extroversion being the main predictors. For stability there was a negative relationship (β= -.288) implying that more neurotic users are more likely to communicate with would be romantic attachments online than more stable users. The relationship was positive for extroversion (β=.234). The more extrovert are more likely to use the SNS to make contact romantically with other users. A further 2% of the variability was explained by self-esteem. This was a negative relationship (β=-.172) implying that those lower in self-esteem use SNS for romantic purposes more than those with higher self-esteem. Seven per cent more was explained by the Narcissism factors though only the exhibitionism factor showed a significant relationship. This was positive (β=.280) implying again that higher the level of narcissism the more frequent the usage of the SNSs for romantic purposes. Finally, an additional 2.4% was accounted for by the social support factor. The coefficient was negative (β=-.164) indicating that users who

are more satisfied with their offline support network are less likely to use the SNSs to achieve their romantic goals.

For the showing encouragement item only the Big Five factors accounted for a significant proportion of the variability in SNS usage (15.6%). In this case extroversion showed a positive relationship (β=.223) and conscientiousness (β=-.406) and stability (β=-.207) negative relationships. Extroverts offer help more frequently and those who are conscientious and stable less so. For the publicizing events item only the Big Five variables accounted for a significant proportion of the variability (7.1%) with extroversion being the primary predictor (β=.293). Extroverts are more likely to publicize their own events.

4 Discussion

Overall, it is clear from these results that personality makes a substantial difference to the ways that people interact in SNSs. Perhaps given the extensive psychological research on these dispositional factors the results are unsurprising. Extroversion predicts activities that are associated with outgoing people who enjoy social interactions. Conscientiousness predicts avoiding behaviors not associated with hard work. Stability, or rather it's converse neuroticism, predicts behaviors that allow the user to avoid the anxieties associated with face-to-face communication. Self-Esteem is associated not using SNS for romantic purposes. Narcissists, are essentially exhibitionists, who enjoy the attention seeking and disclosing nature of SNSs. Finally, those who are satisfied with their offline social network presumably use those relationships to occupy their time and to seek out romantic relationships.

What is important, however, is that personality is a good predictor of SNS usage. The fine details of which behaviors are moderated by which personality factors are for the most part unknown for many of the different activities not considered here that users indulge in when interacting with social computing technologies. However, paying close attention to psychological theories of personality or behavioral dispositions, would allow designers, modelers and researchers to draw useful conclusions about what is likely to happen.

These findings are indicative of the ways in which different users use SNSs for different purposes. Any attempt to model such interactions in order to predict who is going to do what may well be destined to fail given that somewhere between 10% and 20% of the variability in the frequency of different internet activities can be predicted on the basis of intrinsic properties of the users and not properties of the SNSs themselves. This may well explain why word of mouth or viral marketing on these websites is not as successful as might be expected. If the wrong users are being targeted because dispositional aspects of the users are not being taken into account then perhaps we should be unsurprised when the targets are missed.

What lessons should be drawn from these findings. First, any modeling, done either for the purposes of understanding the human mind and human social interactions as advocated by Sun (2007) or for any other purposes, would probably do well to take account of individual differences in behavioral dispositions. The precise details of how these factors can be parameterized is unclear, but since the distribution and frequency of many of these personality factors and their associated behaviors are

well known in the Psychology literature, perhaps any modeling enterprise would benefit from having different cognitive architectures that represent those different distributions of behaviors. Given that Sun (2007) has an architecture that already incorporates motivational and metacognitive subsystems adaptations to his model may well be a good place to start. Whether such systems scale up to hundreds of thousands of users is of course an issue. But if it does, then scaling up to the right kinds of users would save a lot of time and effort. If nothing else such an enterprise would provide a fascinating insight into the development of social networks both online and offline.

Second, and more pragmatically, establishing means of identifying different personality types, either through the computational modeling of the actual use of SNSs or through the direct or surreptitious collection of personality data, may well benefit both the users and the owners of SNSs. The user's may benefit through the adaptation of the SNSs to their own personal needs and drives and the owners may benefit through direct, and importantly, appropriate marketing of products. Given the remarkable amounts of money being paid for internet sites such as MySpace and Facebook, and the huge financial investments in their development and maintenance, it may well be that building personality or at least user's behavioral dispositions into the underlying structures and models will be economically efficacious.

Finally, as was noted earlier, some users may well be harmed by their interactions with social technologies such as the internet. If designers, modelers and researchers in the area of social computing were able to predict who will gain and who might lose on the basis of incorporating theories of personality some of that harm may well be avoided. It would bring socially responsible thinking to the forefront. That would surely be no bad thing.

Acknowledgements

The author would like to thank Eamonn Ferguson, Claire Lawrence and Wei Chuan Yin for their time and effort when writing this paper.

References

1. Fischer, C.: America Calling: A Social History of the Telephone to 1940. University of California Press, Berkeley (1992)
2. Hughes Jr., R., Hans, J.D.: Computers, the internet, and families: a review of the role new technology plays in family life. Journal of Family Issues 22, 778–792 (2001)
3. Eysenck, H.J., Eysenck, S.E.G.: Manual: Eysenck Personality Inventory. Educational and Industrial Testing Service, San Diego, CA (1975)
4. Hamburger, Y.A., Ben-Artzi, E.: The relationship between extraversion and neuroticism and the different uses of the Internet. Computers in Human Behaviour 16, 441–449 (2000)
5. McElroy, J.C., Hendrickson, A.R., Townsend, A.M., DeMarie, S.M.: Dispositional Factors in Internet Use: Personality Versus Cognitive Style. MIS Quarterly 31, 809–820 (2007)
6. Costa, P.T., McCrae, R.R.: Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) Professional Manual, Psychological Assessment Resources, Odessa, FL (1992)

7. Amiel, T., Sargent, S.L.: Individual Differences in Internet Usage Motives. Computers in Human Behavior 20, 711–726 (2004)
8. Tuten, T., Bosnjak, M.: Understanding Differences in Web Usage: The Role of Need for Cognition and the Five Factor Model of Personality. Social Behavior and Personality 29, 391–398 (2001)
9. Harter, S.: The construction of the self: a developmental perspective. Guilford Press, New York (1999)
10. Kraut, P., Patterson, M., Lundmark, V., Kiesler, S., Mukopadhyay, T., Scherlis, W.: Internet paradox: a social technology that reduces social involvement and psychological well-being? American Psychologist 53, 65–77 (1998)
11. Kraut, R., Kiesler, S., Boneva, B., Cummings, J.N., Helgeson, V., Crawford, A.M.: Internet paradox revisited. Journal of Social Issues 58, 49–74 (2002)
12. Raskin, R., Terry, H.: A principal components analysis of the Narcissistic Personality Inventory and further evidence of its construct validity. Journal of Personality and Social Psychology 54, 890–902 (1988)
13. Kubarych, T.S., Deary, I.J., Austin, E.J.: The Narcissistic Personality Inventory: factor structure in a non-clinical sample. Personality and Individual Differences 36, 857–872 (2004)
14. Bargh, J.A., McKenna, K.Y.A.: The internet and social life. Annual Review of Psychology 55, 573–590 (2004)
15. Sun, R.: Cognitive Social Simulation Incorporating Cognitive Architectures. IEEE Intelligent Systems 22, 33–39 (2007)
16. Rosenberg, M.: Society and the adolescent self-image. Princeton University Press, Princeton (1965)
17. Broadhead, W.E., Gehlbach, S.H., DeGruy, F.V., Kaplan, B.H.: Functional versus structural social support and health care utilization in a family medicine outpatient practice. Medical Care 27, 221–233 (1989)
18. Goldberg, L.R.: The development of markers for the Big-Five factor structure. Psychological Assessment 4, 26–42 (1992)

Agent-Based Social Simulation and Modeling in Social Computing

Xiaochen Li[1], Wenji Mao[1], Daniel Zeng[1,2], and Fei-Yue Wang[1]

[1] Key Laboratory of Complex Systems and Intelligence Sciences, Institute of Automation, Chinese Academy of Sciences
[2] Department of Management Information Systems, University of Arizona

Abstract. Agent-based social simulation (ABSS) as a main computational approach to social simulation has attracted increasing attention in the field of social computing. With the development of computer and information technologies, many new ABSS approaches have been proposed with wide application.. In this paper, we aim at reviewing research and applications of agent-based social simulation and modeling in recent years from a social computing perspective. We identify the underlying social theories for ABSS, its simulation and modeling techniques, and computational frameworks from both individual agent and multi-agent system perspective. We finally address some future research issues in agent-based social simulation and modeling.

Keywords: social simulation and modeling, agent-based approach, computational framework, social computing.

1 Introduction

With the growing popularity of social software and Web 2.0, increased academic interest in social network analysis, and the rise of open source as a viable method of production, social computing has attracted increasing attention. Social computing takes a computational approach to the study and modeling of social interactions and communications [1, 2]. As a research area of social computing, computer simulation of social phenomenon is a promising field of research at the intersection of social and mathematical science [3]. Social simulation is the modeling or simulation of social phenomena or objects (society, organizations, markets, human beings) which normally performed by a computer.

One major computational modeling approach for social simulation is agent-based social simulation (ABSS). Schelling [4, 5] is credited for developing the first social agent-based simulation in which agents represent people and agent interactions represent a socially relevant process. Schelling applied notions of cellular automata to study housing segregation patterns [6]. Epstein and Axtell [7] developed the first large scale agent model, the Sugarscape, which extended from modeling people to modeling the entire society. A lot of social processes were observed in their model including death, disease, trade, wealth, sex and reproduction, culture, conflict and war, and externalities such as pollution [6].

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 401–412, 2008.

Societies, in particular, human societies, are often complex adaptive systems. There are a lot of non-linear interactions between their members or between people. Traditional computational and mathematical models can hardly represent these kinds of complex systems since complex social processes can't be represented as an equation. In agent-based models, the agents can have a one-to-one correspondence with the individuals (or organizations, or other agents) that exists in the real social world being modeled, while the interactions between the agents can likewise correspond to the interactions between the real world individuals [8]. ABSS represents a methodological approach that could contribute to two aspects: (1) the rigorous testing, refinement, and extension of existing theories that have proved to be difficult to formulate and evaluate using standard statistical and mathematical tools; and (2) a deeper understanding of fundamental causal mechanisms in multi-agent systems whose study is currently separated by artificial disciplinary boundaries [9].

ABSS is a cross disciplinary research and application field. As shown in Figure 1, Davidsson [10] defines and differentiates the research areas that are combination of agent-based computing, computer simulation, and social science such as Agent-Based Social Simulation (ABSS), Social Aspects of Agent Systems (SAAS), Multi Agent Based Simulation (MABS), and Social Simulation (SocSim). Other than ABSS which is defined above, SAAS consists of social science and agent-based computing and includes the study of norms, institutions, organizations, co-operation, competition, etc. MABS mainly uses agent technology for simulating any phenomena on a computer. Finally SocSim is lying on the intersection between social sciences and computer simulation and corresponds to the simulation of social phenomena on a computer using typically simple models such as cellular automata.

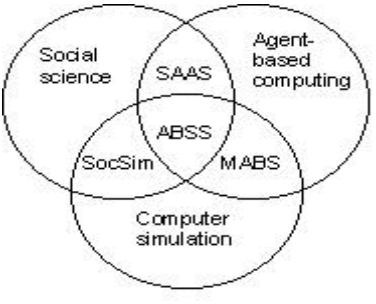


Fig. 1. Intersections of the research areas of social science, agent-based computing, and computer simulation

Although agent-based social simulation has been proposed since 1970s, it is getting more popular in recent decades with the development of artificial intelligence and computational theory. A lot of new approaches to ABSS have been proposed by researchers and the application domain is expanding rapidly. It thus becomes rather confusing for the ABSS researchers to adopt proper agent-based simulation approaches for their specific modeling and simulation problem. However, few researchers have tried to summarize the research and development of ABSS in recent years. To contribute to this important research field, this paper aims at reviewing both

social and computing theories for ABSS, agent-based frameworks as well as main applications for social simulation and modeling. We emphasize the relationship of social theory (e.g., social psychology theory, organizational theory) with agent-based model. The computational frameworks, such as agent-based architectures, social norms and organizations, are also discussed in the following sections. Our work contributes to the identification of the current-state-of-the-art in ABSS research and development. We also raise some future issues of agent-based social simulation.

The layout of this paper is as follows. In Section 2, we discuss the interrelation of social theory and computational model. In Section 3, we review various computational frameworks of ABSS. In Section 4, we describe the main application areas for agent-based social simulation and modeling. Section 5 presents some free toolkits and discusses the ABSS validation methodologies. In Section 6, we conclude with some future research issues of agent-based social simulation and modeling.

2 Social Theories in ABSS

Just as shown in Figure 1, social science provides theoretical foundation for the studies of social simulation and modeling. Social simulation studies provide an ideal opportunity for filling the gap between empirical research and theoretical work. In particular, social simulation provides not only a methodology for testing hypotheses, but also an observatory of social process [11]. One of the most attractive features of social simulation is to examine the feasibility of a new social theory. With social model, we can test social theories in computational models. On the other hand, simulating a social theory is not an easy task.

A concrete algorithmic framework for Social Comparison Theory (SCT) [12] was proposed by [13], and evaluated in several crowd behavior scenarios. Social Comparison Theory is a popular social psychology theory that has been continuously evolving since the 1950s. The key idea in this theory is that humans, lacking objective means to evaluate their state, compare themselves to others that are similar. To be usable by computerized models, SCT's axioms must be transformed into an algorithm that, when executed by an agent, will prescribe social comparison behavior.

Besides social psychology theory which has mentioned above, organizational theory and anthropology have inspired many researchers and been widely used in social simulation and modeling. In social modeling and simulation, organizational theory attempts to explain the organizational behavior of agents and virtual organizations, and helps investigate issues such as organization structure and social pattern. For instance, Kolp *et al* [14] adopted organizational theory as guidance for describing model structure and organization in multi-agent systems. The theories they adopted in their work contain Organization Theory (where the aim is to understand the structure and design of an organization, [15, 16, 17]) and Strategic Alliances (that model the strategic collaborations of independent organizational stakeholders who have agreed to pursue a set of shared business objectives, [18, 19, 20, 21]). Moreover, researchers (e.g., [22, 23]) claim that organizational theory plays a predominant role in their model design.

Besides SCT and organizational theory, anthropology is the study of human beings over time and space. It seeks to understand humans by exploring the differences and

similarities between people in terms of cultures, societies and biology all over the world and throughout their existence. Bordini *et al* [24] proposed the use of cognitive anthropology [25] as a theoretical foundation and suggested that fieldwork practice in social anthropology [26] can provide useful techniques for an agent's adaptation to a strange society. They argued that further research linking social anthropology and multi-agent systems should provide useful techniques for migrant agents to gain a more thorough understanding of the target societies.

3 Computational Frameworks for Agent-Based Social Simulation and Modeling

We review some main frameworks for both single agent and multi-agent systems (MAS). First, in Sections 3.1 and 3.2, we identify some key developments of agent frameworks, from individual agent and multi-agent perspectives, respectively. Then in Section 3.3, we discuss cognitive architectures involved in social simulation and modeling.

3.1 Agent Frameworks

3.1.1 BDI Architecture and Its Extensions

Social phenomenon emerges from numerous individuals' behavior. In modeling and simulating social behavior, agents can be viewed as autonomous entities that can observe the environment, talk with other entities, and make their own decisions according to their judgments. Agent architecture is designed to help agents make decisions and interact with other agents. The most popular one perhaps is the BDI architecture. BDI describes the agent's metal state that is composed of three parts: belief, desire and intention. The agent can make decisions according to the states of its beliefs, desires, and intentions like humans. Until now, a number of the BDI architecture and its variances have been implemented in human behavior modeling [27].

Although the BDI architecture has been successfully applied to many problems, it ignores a fact that human decision making not only depends on their rational judgment, but also sometimes relates to their emotion states. Pereira *et al* [28] argued that emotion can be essential to make human behavior simulation more natural and effective. They presented a conceptual Emotional-BDI architecture, an extension to the original BDI architecture. However, their work did not clearly represent the difference between emotional agents and normal rational agents [29]. Jiang and Vidal [29] focused on how emotions influence an agent's decision making and proposed a generic Emotional BDI architecture which takes both primary emotions and secondary emotions into account. They claimed that multi-agent systems built with this kind of emotional agents will be able to achieve higher social welfare than those built with traditional selfish agents.

3.1.2 Layered Architecture

Layered architecture is another important agent architecture used in social simulation and modeling. Layered architectures realize decision making via various software

layers, each of which explicitly reasons about the environment based on different levels of abstraction [30]. For example, COGNITIVA [31] is a cognitive, multilayered architecture for the design of intelligent virtual agents. The architecture covers three layers, including reactive, deliberative and social layers, and several kinds of behavior associated with different layers in the architecture. It also provides agents with emotion-influenced behavior and has been extended to model their social interactions. This layered architecture is tested in a predation simulation environment. The results show that COGNITIVA's management of individual behavior and its capability to model social interactions are a good example for the design and control of intelligent virtual agents.

3.2 Multi-agent Frameworks

3.2.1 Modeling Organizations

Modeling virtual organizations is an active research area in social simulation. In recent years, various virtual organizations are built by the researchers, such as institution, group, firm, community. A hot research topic in virtual organizations is the modeling and simulation of organizational structure. The organizational structure for multi-agent systems usually involves two fundamental concepts: agent roles and their relations in terms of which the collective behavior of individual agents is specified and the overall behavior of the MAS is determined [32]. The specification of the overall behavior of MAS concerns the optimization of agents' activities, the management and security of the information flow among agents, as well as the enforcement of certain outcomes [32].

In modeling multi-agent organizations, Kolp *et al* [14] described several organizational styles (mainly the structure-in-5 and the joint-venture) as the meta-class of organizational structures for multi-agent simulations, which adopted concepts from organizational theories. They argued that organizational structure can offer a set of design parameters to coordinate the assignments of organizational objectives and processes in multi-agent systems, thereby affecting how an organization functions. Grossi *et al* [32] defined a formal relation between institutions and organizational structures. They aimed to show how norms can be refined to construct organizational structures which are closer to an implemented system.

For organization design itself, it is not always easy to achieve organizational behavior via agent-based models. There are several reasons for this. Organizations are composed of individual agents, and individual agents are autonomy entities pursuing their own goals based on their beliefs and capabilities. However, their behavior and actions are sometimes irregular, unpredictable or deviate from expected behavior. So the question arises as how organizational behavior can emerge from individual agents' interactions. Other approaches have been proposed, for example, based on social norms.

3.2.2 Modeling Social Norms

In order to regulate agents' behavior for the emergence of systematic, organizational behavior, social norms are widely used for organization formation. It can be considered as a top-bottom methodology. For example, social law (a set of restrictions defined on the agents' activities which allow them to do something aimed at the same

time to constrain their behavior [33]), social reputation (a norm which the satisfied agent will get reward, but the violated one gets no punishment [34]), etc. As social norm is an abstract concept in sociology, normative frameworks are designed to implement norms in electronic institutions.

One such normative framework was proposed in [35]. Their work introduced a normative language which is expressive enough to represent the familiar types of MAS-inspired normative frameworks. Another example is OMNI (Organizational Model for Normative Institutions), proposed in [36]. Their framework specified global goals of the system independently from those of the specific agents that populated the system. They presented the norms in the environment that should govern the emergent behavior of the agent society as a whole and the actions of individuals. Both the norms that regulate interactions between agents and the contextual meaning of those interactions are important aspects when modeling multi-agent behavior.

3.2.3 Modeling Emergent Social Phenomena

In recent years, agent-based models become increasingly popular in capturing emergent social phenomenon. However, as emergent phenomenon emerges from intelligent agents' interactions, it can be rather complex and difficult to coordinate agents' behavior so as to produce some desired social pattern or phenomenon, such as social rumor and social order. As a result, model frameworks are often built to resolve this problem. In contrast to social norms which pursue regulating agents' behavior, this kind of frameworks is aimed at producing emergent behavior or phenomena.

Xia and Huang [37] designed the notion of social rumor in a standard game-theoretic framework, and introduced a simple and natural strategy-select rule called behavior update rule to restrict agents' behavior to one particular strategy and lead to emergence of social rumor or anti-rumor. Social order is another kind of emergent behavior which can be obtained by using social norms and social control. Grizard *et al* [38] presented a normative P2P architecture to obtain social order in multi-agent systems. They proposed the use of two types of norms that coexist: rules and conventions. Social control is obtained by providing a non-intrusive control infrastructure that helps the agents build reputation values based on their respect of norms. The experiments show interesting results, notably the fact that agents with good reputations are rapidly identified and at the center of communities of agents with similar behavior.

3.3 Cognitive Architecture

Cognitive architecture integrates cognitive science with agent-based simulation to help describe, explain, and predict social phenomena, through capturing the cognition of individual agents. Sun [39] argued that agent-based social simulation with multi-agent systems can benefit from incorporating cognitive architectures. Sun's work elaborated the CLARION cognitive architecture and discussed three case studies of applying CLARION to social simulation. The results suggest that agents can capture a variety of cognitive processes in a psychologically realistic way from the CLARION architecture. CLARION may help more accurately capture organizational performance data and formulate deeper explanations of the observed results.

Other prestigious cognitive architectures include SOAR [40] and ACT-R [41], among others. They have their theoretical basis on human cognition, and are originally focused only on the internal information processing of an intelligent agent, including tasks such as reasoning, planning, problem-solving, and learning. For example, SOAR has been used in modeling social agents [42] as well as evaluating social theory [13]. ACT-R is common used for the design of social robots in social simulation.

4 Main Applications

The main applications of agent-based social simulation are social robots, virtual markets, education and training, terrorism, and so on. Below we shall discuss them.

4.1 Social Robots

Social robot is a kind of robot different from traditional industrial robot. It is designed to communicate with humans and to participate in human society. Trafton *et al* [43] presented a social robot which is able to mimic the child's hiding behavior. They sought to understand how children learned to play hide and seek, and thus created a robot that can understand how to play hide and seek. ACT-R architecture is used in the design of their robot model for integrating and organizing psychological data based on the observation of children. The case study results show that a 3 1/2 year old child who did not have perspective taking skills was able to learn how to play a credible game with robot. Social robot is regarded as a special approach to understand the nature of society and human behavior in social computing.

4.2 Online Markets

Virtual markets are a hot application topic of social simulation. It is attractive in investigating market mechanism and consumer behavior. Multi-agent framework is often designed in virtual markets to model the market dynamics. Karacapilidis and Moraïtis [44] described an agent-based artificial market system whose underlying interaction protocols provide several advanced features. The development of the software agents proposed in their system is based on a generic and reusable multi-agent framework. Using the system, actors (i.e., customers and merchants) can delegate a variety of tasks to intelligent agents that act as their artificial employees. Zhang and Zhang [45] investigated the irrational investors' survival status in artificial stock markets. They designed a typical artificial stock market framework to simulate virtual market. Their framework has four basic categories of actors: issuers, investors, and the exchange. The results reveal that rational investors cannot "eliminate" irrational investors, even in the long run. Besides the paradigms above, some featured virtual markets (see [46, 47]) are also built to explore markets from different aspects.

4.3 Education and Training

Education in the 21st century is faced with increasing demands of using intelligent systems. The learning environments of today and tomorrow must handle distributed

and dynamically changing contents, a geographical dispersion of students and teachers, and generations of learners who spend hours a day interacting with multimillion-dollar multimedia environments [48]. The new learning environments are moving beyond kindergarten through college classrooms to distance learning, lifelong education, and on-the-job training [48]. Swartout *et al* [49] described the virtual humans developed as part of the Mission Rehearsal Exercise project, a virtual reality-based training system. Virtual humans are software artifacts that look like, act like and interact with humans but exist in virtual environments. Almost all components of the virtual humans are implemented in SOAR, allowing each component to be implemented with production rules that read from and write to a common working memory (i.e. blackboard). In such applications, a learner can explore stressful social situations in the safety of a virtual world.

4.4 Security-Related Applications

After the 9/11 events, the security problem especially terrorism becomes critical worldwide. Many researchers seek to take advantage of agent-based model to simulate terrorist activities. Carley *et al* [50] described a scalable citywide multi-agent model named BioWar, which is capable of simulating the effects of weaponized biological and chemical attacks. Moon and Carley [51] developed a simple theoretical multi-agent simulation model for a real-world terrorist network to show how changes in the co-evolution of social and geospatial dimensions affect group behavior. Both models use social network to represent the organization's current structural characteristics. Organizational structure performs well in representing the relationship of agents (or terrorists) according to their results.

5 Toolkits and Validation

For the researchers who seek to build agent-based social model, many ABMS software environment are freely available now. These include Repast, Swarm, NetLogo and MASON, among others [6]. Macal and North [6] elaborated the capability of these four toolkits and compare them from different aspects. After we get a toolkit and run a social simulation, an important question is how to validate or verify the experiment results. However, the assurance/verification and validation of agent-based models are difficult, because of the heterogeneity of agents, and the possibility of the emergence of new patterns of macro behavior as a result of the agent interactions at the micro level [52]. Midgley *et al* [52] suggested that the central approach to validate ABM will most likely continue the scientific tradition of empirical validation and testing. Moss and Edmonds [53] presented a cross-validation approach and claimed that there are at least two stages of empirical validation of ABM, the micro and the macro.

6 Future Research

As a critical part of social computing research, social simulation is playing an increasingly important role in today's flattened and interconnected society. In

particular, agent-based social simulation has already been widely applied to many promising areas, such as training, education, social and psychological research. However, there are two major drawbacks associated with agent-based approaches. One drawback is that the patterns and the outcomes of agent interactions are inherently unpredictable [54]. In addition, predicting the behavior of the overall system based on its constituent components is extremely difficult (sometimes impossible) because of the emergence of collective behavior [54].

Despite these drawbacks, in many situations agent-based social simulation and modeling techniques provide valuable insights that cannot be obtained otherwise. Research in this area is progressing steadily. Current agent-based social simulation has focused on the design of artificial social systems and specific applications. The ABSS enabling technologies are still in great demand. Key future research directions include conflict and cooperation between agents, trust and norm formation dynamics, modeling social and organizational structure, group decision making and collective behavior, emergence and evolution of organizations, among other.

We also predict that the design of human-like cognitive agents will become a hot research direction in agent-based social simulation. So does virtual organization. However, the validation methodology is not yet mature. Immediate attention is called for to resolve some of these methodological challenges. From the point of view of technical research, we expect application of advanced computing theories and methods may lead to reduction in model design complexity.

Acknowledgements

This work is supported in part by NNSFC #60621001 and #60573078, MOST #2006AA010106, #2006CB705500 and #2004CB318103, CAS #2F05N01 and #2F07C01, and NSF #IIS-0527563 and #IIS-0428241.

References

1. Wang, F.-Y., Carley, K.M., Zeng, D., Mao, W.: Social Computing: From Social Informatics to Social Intelligence. Intelligent Systems 22, 79–83 (2007)
2. Zeng, D., Wang, F.-Y., Carley, K.M.: Guest Editors' Introduction: Social Computing. IEEE Educational Activities Department 22, 20–22 (2007)
3. Conte, R., Gilbert, N., Sichman, J.S.: MAS and Social Simulation: A Suitable Commitment. In: Sichman, J.S., Conte, R., Gilbert, N. (eds.) MABS 1998. LNCS (LNAI), vol. 1534, pp. 1–9. Springer, Heidelberg (1998)
4. Schelling, T.C.: Dynamic Models of Segregation. Journal of Mathematical Sociology 1, 143–186 (1971)
5. Schelling, T.C.: Micromotives and Macrobehavior. Norton, New York (1978)
6. Macal, C.M., North, M.J.: Tutorial on Agent-Based Modeling and Simulation. In: Proceedings of the 37th Conference on Winter Simulation. Winter Simulation Conference, Orlando, Florida (2005)
7. Epstein, J.M., Axtell, R.: Growing Artificial Societies: Social Science from the Bottom Up. The Brookings Institution (1996)
8. Gilbert, N.: Agent-Based Social Simulation: Dealing with Complexity (2005)

9. Axelrod, R., Tesfatsion, L.S.: A Guide for Newcomers to Agent-Based Modeling in the Social Sciences. Iowa State University, Department of Economics (2006)
10. Davidsson, P.: Agent Based Social Simulation: a Computer Science View. Journal of Artificial Societies and Social Simulation 11 (2002)
11. Conte, R., Gilbert, N.: Computer Simulation for Social Theory. In: Conte, N.G.R. (ed.) Artificial Societies: The Computer Simulation of Social Life, pp. 1–18. UCL Press, London (1995)
12. Festinger, L.: A Theory of Social Comparison Processes. Human Relations, 117–140 (1954)
13. Fridman, N., Kaminka, G.A.: Towards a Cognitive Model of Crowd Behavior Based on Social Comparison Theory. AAAI, Stanford, California, USA (2007)
14. Kolp, M., Giorgini, P., Mylopoulos, J.: Multi-Agent Architectures as Organizational Structures. Autonomous Agents and Multi-Agent Systems 13, 3–25 (2006)
15. Mintzberg, H.: Structure in Fives: Designing Effective Organizations. Prentice-Hall, Englewood Cliffs (1992)
16. Morabito, J., Sack, I., Bhate, A.: Organization Modeling: Innovative Architectures for the 21st Century. Prentice-Hall, Englewood Cliffs (1999)
17. Scott, W.R.: Organizations: Rational Natural and Open Systems. Prentice-Hall, Englewood Cliffs (1998)
18. Dussauge, P., Garrette, B.: Cooperative Strategy: Competing Successfully through Strategic Alliances. Wiley and Sons, Chichester (1999)
19. Gomes-Casseres, B.: The Alliance Revolution: The New Shape of Business Rivalry. Harvard University Press (1996)
20. Segil, L.: Intelligent Business Alliances: How to Profit Using Today's Most Important Strategic Tool. Times Business (1996)
21. Yoshino, M.Y., Rangan, U.S.: Strategic Alliances: An Entrepreneurial Approach to Globalization. Harvard Business School Press (1995)
22. Conte, R., Paolucci, M.: Responsibility for Societies of Agents. Journal of Artificial Societies and Social Simulation 7 (2004)
23. Carley, K.M., Prietula, M.J., Lin, Z.: Design Versus Cognition: the Interaction of Agent Cognition and Organizational Design on Organizational Performance. Journal of Artificial Societies and Social Simulation 1 (1998)
24. Bordini, R.H., Campbell, J.A., Vieira, R.: Extending Ascribed Intensional Ontologies with Taxonomical Relations in Anthropological Descriptions of Multi-Agent Systems. Journal of Artificial Societies and Social Simulation 11 (1998)
25. Tyler, S.A.: Cognitive Anthropology. Holt, Rinehart and Winston Inc., New York (1969)
26. Bernard, H.R.: Research Methods in Anthropology: Qualitative and Quantitative Approaches. Sage Publications, Thousand Oaks, California (1994)
27. Georgeff, M.P., Pell, B., Pollack, M.E., Tambe, M., Wooldridge, M.: The Belief-Desire-Intention Model of Agency. In: Rao, A.S., Singh, M.P., Müller, J.P. (eds.) ATAL 1998. LNCS (LNAI), vol. 1555, pp. 1–10. Springer, Heidelberg (1999)
28. Pereira, D., Oliveira, E., Moreira, N., Sarmento, L.: Towards an Architecture for Emotional BDI Agents. In: Oliveira, E. (ed.) Portuguese Conference on Artificial intelligence, 2005. EPIA 2005, pp. 40–46 (2005)
29. Jiang, H., Vidal, J.e.M.: From Rational to Emotional Agents. In: Proceedings of the AAAI Workshop on Cognitive Modeling and Agent-based Social Simulation (2006)
30. Muller, J.P.: The Design of Intelligent Agents: A Layered Approach. Springer, New York (1997)

31. Freitas, J.S.d., Imbert, R., Queiroz, J.: Modeling Emotion-Influenced Social Behavior for Intelligent Virtual Agents. In: Gelbukh, A., Kuri Morales, Á.F. (eds.) MICAI 2007. LNCS (LNAI), vol. 4827, pp. 370–380. Springer, Heidelberg (2007)
32. Grossi, D., Dignum, F., Dastani, M., Royakkers, L.: Foundations of Organizational Structures in Multiagent Systems. In: Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems, ACM, The Netherlands (2005)
33. Boella, G., Torre, L.v.d.: Enforceable Social Laws. In: Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems, ACM, The Netherlands (2005)
34. Hahn, C., Fley, B., Florian, M., Spresny, D., Fischer, K.: Social Reputation: a Mechanism for Flexible Self-Regulation of Multiagent Systems. Journal of Artificial Societies and Social Simulation 10 (2007)
35. Garcia-Camino, A., Noriega, P., Rodriguez-Aguilar, J.A.: Implementing Norms in Electronic Institutions. In: Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems, ACM, The Netherlands (2005)
36. Vázquez-Salceda, J., Dignum, V., Dignum, F.: Organizing Multiagent Systems, vol. 11, pp. 307–360. Kluwer Academic Publishers, Dordrecht (2005)
37. Xia, Z., Huang, L.: Emergence of Social Rumor: Modeling, Analysis, and Simulations. In: Computational Science – ICCS 2007, pp. 90–97 (2007)
38. Grizard, A., Vercouter, L., Stratulat, T., Muller, G.: A Peer-to-Peer Normative System to Achieve Social Order. In: Coordination, Organizations, Institutions, and Norms in Agent Systems II, pp. 274–289 (2007)
39. Sun, R.: Cognitive Social Simulation Incorporating Cognitive Architectures. IEEE Educational Activities Department 22, 33–39 (2007)
40. Newell, A.: Unified Theories of Cognition. Harvard University Press (1990)
41. Anderson, J.R., Lebiere, C.: The Atomic Components of Thought. Lawrence Erlbaum Associates, Mahwah (1998)
42. Gratch, J., Mao, W., Marsella, S.: Modeling Social Emotions and Social Attributions. In: Sun, R. (ed.) Cognition and Multi-Agent Interaction, Cambridge University Press, Cambridge (2006)
43. Trafton, J.G., Alan, C.S., Dennis, P., Magdalena, D.B., William, A., Nicholas, L.C., Derek, P.B.: Children and Robots Learning to Play Hide and Seek. In: Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction, Salt Lake City Utah, USA, ACM Press, New York (2006)
44. Karacapilidis, N., Moratis, P.: Intelligent Agents for an Artificial Market System. In: Proceedings of the Fifth International Conference on Autonomous Agents, Montreal Quebec, Canada, ACM Press, New York (2001)
45. Zhang, Y., Zhang, W.: Can Irrational Investors Survive? A Social-Computing Perspective. Intelligent Systems, IEEE 22, 58–64 (2007)
46. Tay, N.S.P., Lusch, R.F.: Agent-Based Modeling of Ambidextrous Organizations: Virtualizing Competitive Strategy. Intelligent Systems, IEEE 22, 50–57 (2007)
47. Hoffmann, A.O.I., Wander, J., Eije, J.H.V.: Social Simulation of Stock Markets: Taking It to the Next Level. Journal of Artificial Societies and Social Simulation 11 (2007)
48. Aroyo, L., Graesser, A., Johnson, L.: Guest Editors' Introduction: Intelligent Educational Systems of the Present and Future. Intelligent Systems, IEEE 22, 20–21 (2007)
49. Swartout, W., Gratch, J., Hill, R.W., Hovy, E., Marsella, S., Rickel, J., Traum, D.: Toward Virtual Humans. American Association for Artificial Intelligence 27, 96–108 (2006)

50. Carley, K.M., Fridsma, D.B., Casman, E., Yahja, A., Altman, N., Chen, L.-C., Kaminsky, B., Nave, D.: BioWar: Scalable Agent-Based Model of Bioattacks. IEEE Transactions on Systems, Man and Cybernetics, Part A 36, 252–265 (2006)
51. Moon, I.-C., Carley, K.M.: Modeling and Simulating Terrorist Networks in Social and Geospatial Dimensions. Intelligent Systems, IEEE 22, 40–49 (2007)
52. Midgley, D., Marks, R., Kunchamwar, D.: Building and Assurance of Agent-Based Models: An Example and Challenge to the Field. Journal of Business Research 60, 884–893 (2007)
53. Moss, S., Edmonds, B.: Sociology and Simulation: Statistical and Qualitative Cross-Validation. American Journal of Sociology 110, 1095–1131 (2005)
54. Jennings, N.R.: On Agent-Based Software Engineering, vol. 117, pp. 277–296. Elsevier Science Publishers Ltd, Amsterdam (2000)

Transforming Raw-Email Data into Social-Network Information

Terrill L. Frantz and Kathleen M. Carley

School of Computer Science, Carnegie Mellon University,
Pittsburgh, PA 15213, USA
{terrill,kathleen.carley}@cs.cmu.edu

Abstract. We describe the technical background necessary for working with real-world email data as well as the process of how to transform it into a social-network representation for research purposes. We demonstrate how to use the CEMAP feature in ORA social-network analysis software, which operationalizes the entire process for researchers with, or without, in-depth computer expertise. A detailed example is provided to show the steps involved in the process and to highlight the various add-on features of CEMAP, such as anonymization and removing duplicate emails.

Keywords: email, social networks, social network analysis, socio-technical networks, ORA.

1 Introduction

Email has become ubiquitous and the routine computer-disk storage of it is an invaluable artifact of real-world human-communication networks, for an individual, a group, an organization, or an entire society. The data captured in the distributed network of email systems is rich in information that extends well beyond its initial functional purpose as merely a data backup (see Diesner, Frantz & Carley, 2005). Original and copies of email data persists in computer files all over the planet, which are physically located on our laptops, desktops, and organizational file servers, and which represents a rich and precise history of who we communicated with, when we communicated, and what was communicated, across tens of trillions of real-world human exchanges. Whether or not the user has access to specific email data will be dependent on the legal and organizational environment in which they are investigating.

This massive data repository forms a planet-level, intertwined data-set that social computing researchers widely recognize as being rich in social information—if the data can be harvested efficiently and effectively analyzed. In this paper, we describe the practical, technical background that a social computing researcher must understand to begin to data mine and study this email data; we provide instructions and present an example on how to operationalize real-world email data research by making use of the CEMAP feature in the ORA social-network analysis software program (see the Appendix Section for more information on ORA.).

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 413–420, 2008.

2 Email Technology and Terminology

The email exchange that two people have, is actually an exchange between two computers, often with several other computers serving as intermediaries in the delivery process. Ultimately, a specific message sent by one person is received by the target person, in its electronic form, via a client software program that picks up the email from its host email server. The email message rests on a designated email server until the receiver "picks up" the message from the server; this is practically identical to the process of picking up hardcopy mail from the post, but in electronic form. The electronic email can be delivered to the post office repository for you to physically pick up, or directly to your personal mail box for you to pick up. Once you pick up the message, via your email client software, you can read it, print it, store it, or follow up on its contents, if you desire.

In the world-wide electronic email architecture, the message can be delivered to the equivalent of the post office lobby-box, to what is called an IMAP server. Alternatively, the email can be delivered to your personal mailbox, what can be called a Postoffice Protocol (POP) server. There are many differences between these two email servers, and your email account is most certainly one or the other of these. The primary difference between an IMAP and a POP email server is the storage feature of the server. An IMAP server will allow you (or your email client software) to persist, or store, your email physically on that server. A POP server only serves as a temporary holding station for a message that is removed once it has been retrieved by your email client software. An IMAP server is enabled to store the message even after the email has been initially retrieved. It should be noted that the POP protocol calls for an email to be removed from the incoming mail box once it has been retrieved, however some software extensions allow for a read-only access to the POP inbox, resulting in the message remaining in the inbox when retrieved and is therefore managed by the client software level.

When an email is retrieved from either an IMAP or a POP server, which is specified by a globally unique name, the email message is formatted into the standard MBOX format, which is a world-wide standard that allows for different email client software programs to access the email from the server, IMAP or POP, without confusion. The MBOX format specifies that email is represented as a text file where

```
From - Wed Feb 27 09:10:44 2008
From: "tom Thumb" <tthumb@msn.com>
To: "terrill frantz" <terrill@org-sim.com>
Subject: Journal Article – For review
Date: Fri, 22 Feb 2008 22:01:26 -0600

Terrill,
Please submit the article review by next
weekend.
Cheers,
Tom.
```

Fig. 1. Example email message in MBOX Email Format

each email begins with a "From "-labeled field and ends with a blank line. Between these delimiters the email is formed into two sections: the Header Section and the Body Section. The Header Section holds the envelop-level data such as the To:, From:, Subject:, Date:, et cetera. The Body Section holds the message text and file attachments, if present. Figure 1 shows an example email message in MBOX format. The header section consists of four header fields: "From", "To", "Subject" and "Date." And the Body Section is below the Header Section, delimited by a blank line.

3 Transformation of Email into Meta-networks

The data making up an email is rich in information about the particular instance of the communication network. The Header Section contains transactional data such as who emailed who, when, and about what (the subject line) that can be represented as a network. The email's Body Section contains text that can be processed into a network by applying a natural language processor to it to also construct a network configuration (the process for doing this is beyond the scope of this paper, but interested readers should look into Automap software from CASOS to perform this NLP transformation; see the Appendix.).

Focusing on the header data, there are several bi-modal networks that can be constructed: From-task, task-to, task-cc, task-bcc. From these four principal networks, various secondary networks can be created by folding them together in various ways. For example, a From-To network can be constructed by folding the From-Task and the Task-To networks together. Most often, an analyst will construct a person-to-person social network, which can be constructed by folding the From-task, task-to, task-cc, and task-bcc networks into a single network representation that captures the person-person network. Several other socio-technical networks can be constructed from the header data, but we limit ourselves to a discussion on the actor to actor networks in this paper.

4 Using ORA to Transform Email Data

The Organizational Risk Analyzer (ORA) is a software program that operates on complex network data and is used for analyzing complex socio-technical network data. ORA produces reports that support the social network analysis of the data, provides visualization capabilities, grouping algorithms, and has an extensive number of tools for manipulating the network data as well. One feature of ORA is the ability to import raw email data; this feature is called CEMAP. CEMAP is an easy to use tool that can copy raw email data directly from an email server into ORA in social network form. This allows quick access to real-world email data in a form conducive to network analysis, reporting and visualization. To use the CEMAP tool, first the ORA software is loaded and installed onto a computer. ORA has a simple installation script for easy installation onto Windows-based computers. Below we provide instructions on how to run the CEMAP processing via ORA once the software has been loaded on a users computer.

After starting ORA, to enter the CEMAP sub-system, the user looks to ORA's primary menu and selects "File." Next, select the "CEMAP parser" menu item. A

CEMAP dialogue box will appear (see Fig. 2.). This is the primary input window for CEMAP. There are a number of data fields and options, which are explained in detail below. Once the relevant data fields are completed and the various options are selected, the user presses the "Start extraction" button (bottom button as shown in Fig. 2).

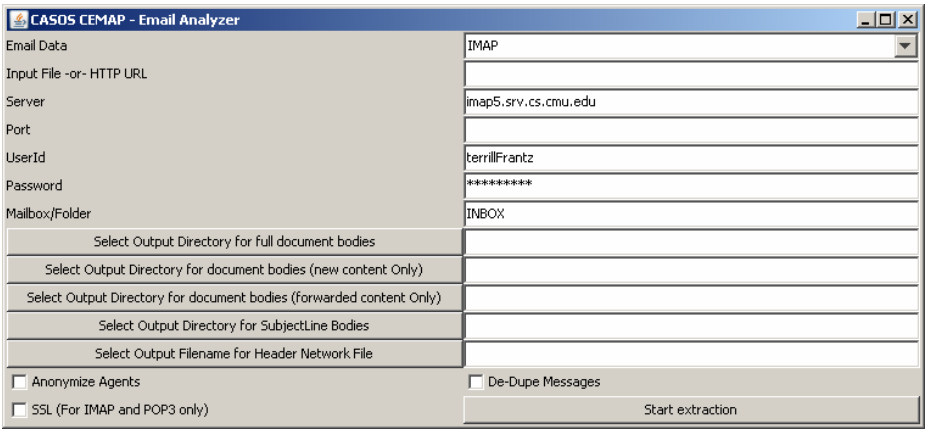


Fig. 2. The main input screen of CEMAP (sample input fields entered)

An ORA progress window (Fig. 3.) will appear that indicates the number of email records processed as it obtains and processes the emails from the specified input source. Upon completing the extract process, a completion message box (Fig. 4.) will appear. The user can press "OK" in the completion message box to indicate the completion of the extract process. As a result, the user will find loaded into ORA the email header data in network form. Figure 5 shows the ORA desktop with an email dataset loaded. Notice the nodesets and networks that are created from the CEMAP process. The network most often of immediate interest is the eAgent2eAgent network; this is the folded network derived from the from-task, task-to, task-cc, and task-bcc networks. From this point, the user can use all of the various reporting and visualization features of ORA to analyze the email data. Figure 5. shows a visualization of the eAgent2eAgent network, as an example.

Fig. 3. The progress box

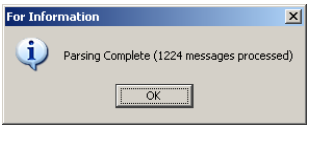


Fig. 4. The completion indicator

4.1 Input Data Fields

There are three types of input fields that should be completed prior to starting the execution of the CEMAP processing. There are fields that specific the input source, the processing features, and the output destinations. These are described below in Table 1.

Table 1. User input fields in the CEMAP processing screen

INPUT FIELDS		
Email Data		A drop-down selection box that indicates the source of the email data and the format to expect the data in. This can be an email server type, (IMAP, or POP3), or a disk file (TXT or MBOX). The two options, CASOSDB and DYNETML are options for internal use only.
Input File –or –HTTP	For TXT or MBOX only	If the source data is a data file this indicates where the data file is located, either locally on a computer disk, or accessible via the Internet using a URL.
Server	For IMAP or POP3 only	the world-unique name of the email server that has the email data desired to be processed.
.	For IMAP or POP3 only	The port number of the email server service. This can usually be left blank as defaults will be applied according to the server type.
UserId	For IMAP or POP3 only	This is the email user name associated with the email account on the mail server.
Password	For IMAP or POP3 only	This is the password for the user name specified above.
Mailbox/Folder	For IMAP or POP3 only	This is the folder in which to read the email data from – most often, set this value to "INBOX"
SSL	For IMAP or POP3 only	Check this box if your email servers requires SSL security.
PROCESSING FIELDS		
Anonymize Agents		Check this box if the names in the email networks are to be masked for privacy.
De-Dupe Messages		Check this box if duplicate emails are to be removed form the network dataset.
OUTPUT FIELDS		
Full document bodies	Expects a directory name	Stores the entire Body Section for each email message into a separate data file.
New content	Expects a directory name	Stores only the new content of the Body Section for each email message into a separate data file.
Forwarded content	Expects a directory name	Stores only the forwarded content of the Body Section for each email message into a separate data file.
Subject Line	Expects a directory name	Stores the Subject line of the each email message into a separate data file.
Network File	Expects a filename	Stores the network representation constructed for all of the emails in a DynetML data file. By default, this same file will be automatically loaded into ORA workspace.

OUTPUT FIELDS: Each of these fields indicate where CEMAP should place various parts of the Body Section text for Natural Language Processing, or where to place the network data file. Automatically, CEMAP will place the network file into ORA workspace. Leaving any or all of these fields blank will no affect processing, merely it will not save the respective portions of the Body Section unless a destination is explicitly indicated.

4.2 Processing TXT Data Files

To accommodate the Microsoft Exchange email software, the TXT format for email can be processed by CEMAP. This format is a comma-delimited rendition of email data as exported by Microsoft Outlook. Table 2 shows the ordering of the TXT file field names. This file should have the fieldnames in the first row of the file.

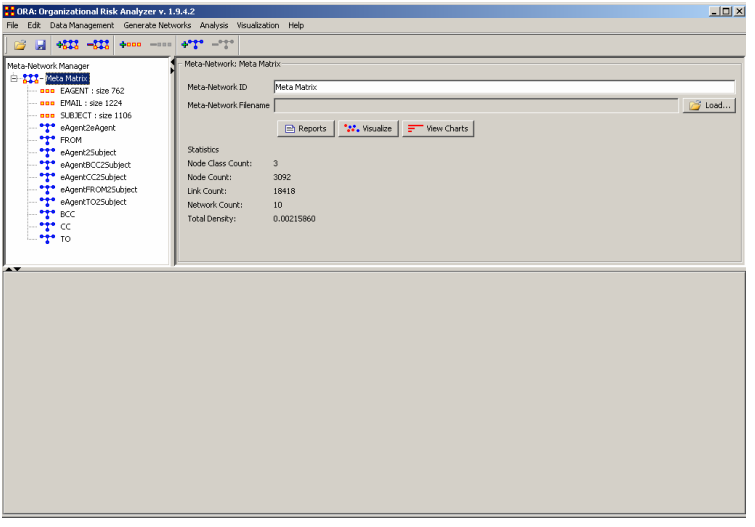


Fig. 5. ORA desktop showing a processed and loaded email meta-network

Table 2. TXT file field names (in strict order)

Subject
Body
From: (Name)
From: (Address)
From: (Type)
To: (Name)
To: (Address)
To: (Type)
CC: (Name)
CC: (Address)
CC: (Type)
BCC: (Name)
BCC: (Address)
BCC: (Type)

5 Using ORA to Explore Email Data

Above we describe the mechanics on how to import and transform email data into network form for analysis using CEMAP in the ORA software program. Figure 6 is a visualization of the network that is produced from a sample email server. The methods for analyzing this data is beyond the scope of this short paper, but suffice to say that a great deal can be learned about the communication network that is manifested in raw email data once it is harvested using the techniques we describe in this paper.

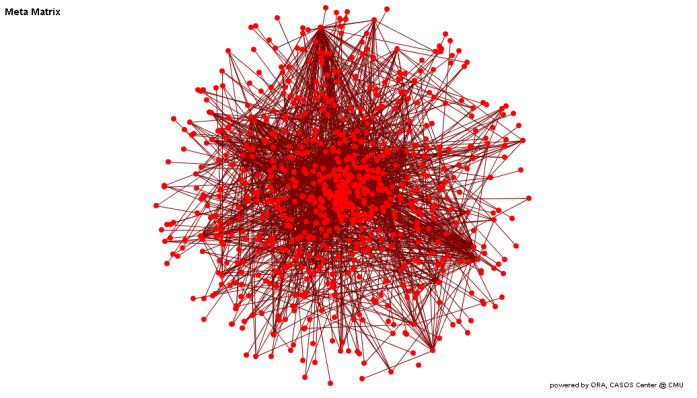


Fig. 6. ORA desktop showing a processed and loaded email meta-network

Acknowledgements

This work is part of the Dynamic Networks project at the center for Computational Analysis of Social and Organizational Systems (CASOS) of the School of Computer Science (SCS) at Carnegie Mellon University (CMU). This work is supported in part by the Office of Naval Research (ONR), United States Navy. Additional support was provided by National Science Foundation (NSF) Integrative Graduate Education and Research Traineeship (IGERT) program, NSF 045 2598, NSF 045 2487, and CASOS. MURI: Air Force Office of Scientific Research, FA9550-05-1-0388 for Computational Modeling of Cultural Dimensions in Adversary Organizations. SPAWAR, Network Analysis and Computational Modeling for Combating Terrorist Threats, MOAT Phase II, DNA application to counter-narcotic investigations related to marijuana. The views and proposal contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of the Office of Naval Research, the National Science Foundation, or the U.S. government.

References

Carley, K. M., Reminga, J.: ORA: Organization Risk Analyzer. Carnegie Mellon University, School of Computer Science, Institute for Software Research International, Technical Report CMU-ISRI-04-106. (2004)

Carley, K. M., Columbus, D., DeReno, M., Reminga, J., Moon, I.:ORA User's Guide 2007. Carnegie Mellon University, School of Computer Science, Institute for Software Research, Technical Report, CMU-ISRI-07-115. (2007)

Carley, K. M., Diesner, J., DeReno, M.: AutoMap User's Guide. Carnegie Mellon University, School of Computer Science, Institute for Software Research, Technical Report, CMU-ISRI-06-114. (2006)

Diesner, J., Carley, K. M.: AutoMap 1.2 – Extract, analyze, represent, and compare mental models from texts. Carnegie Mellon University, School of Computer Science, Institute for Software Research, Technical Report, CMU-ISRI-04-100. (2004)

Diesner, J., Frantz, T., Carley, K. M.: Communication Networks from the Enron Email Corpus "It's Always About the People. Enron is no Different," Computational and Mathematical Organization Theory, 11, 201 – 228. (2005)

Hirshman, B. R., Carley, K. M., Kowalchuck, M. J.: Specifying Agents in Construct. Carnegie Mellon University, School of Computer Science, Institute for Software Research, Technical Report, CMU-ISRI-07-107. (2007)

Schreiber, C., Carley, K. M.: Construct - A Multi-agent Network Model for the Co-evolution of Agents and Socio-cultural Environments. Carnegie Mellon University, School of Computer Science, Institute for Software Research International, Technical Report CMU-ISRI-04-109. (2004)

Appendix: ORA and Automap

The Organization Risk Analyzer (ORA) (www.casos.cs.cmu.edu/projects/ora; Carley and Reminga, 2004; Carley, Columbus et al., 2007) is a software program that computes traditional social network, dynamic network, and link analysis metrics on single and meta-network data. It also allows for traditional node-link and advanced visualizations and user-editing of the meta-network data and well as providing several other aids for advanced analysis, including error detection and what-if analysis using simulation tools. Most measures contained in standard social network tools are in ORA as well as metrics for assessing key nodes in a more complex meta-network or assessing change over time. ORA can be used to identify key nodes, assess change, identify vulnerabilities in networks, suggest missing data, and assess the impact of node and link extraction. Specialized reports for dealing with semantic networks, email data, standardized social networks and general organizational meta-networks exist. ORA can import and export data in a wide variety of forms, including DyNetML. ORA can computes over 100 measures and perform over-time analysis on a series of associated meta-networks. ORA generates formatted reports viewable on the computer screen or inside an Internet browser. It can interoperate with other social network software and is tightly integrated with Automap and the Construct simulation (Schreiber & Carley, 2004; Hirshman, Carley & Kowalchuck, 2007). ORA uses a Java interface for ease of use and a C++ computational backend for speed. A SOAP-based API is also available.

Automap (www.casos.cs.cmu.edu/projects/automap; Diesner & Carley, 2004; Carley Diesner & DeReno, 2006) is a software tool for extracting semantic networks and meta-networks from raw, free-flowing, text-based documents, such as email bodies. From one or more emails, Automap constructs semantic networks by applying advanced information extraction techniques to the collection of words and sentences in the document(s). It constructs the semantic network by encoding links among words of a sentence or paragraph based on both proximity and parts of speech analysis. The resultant network of concepts, a.k.a., word-entities, is the semantic network for the document. Then these concepts are cross-classified into an ontology based on the user chosen entity classes resulting in a meta-network representation of the text. For example, concepts representing people's names are classified as people, those representing locations as locations, and so on. The semantic network can be thought of as a sample of the email author's mental model of the subject at the time the document was written.

Using Social Networks to Organize Researcher Community

Xian-Ming Xu[1], Justin Zhan[1], and Hai-tao Zhu[2]

[1] SION Lab, Carnegie Mellon University
{xianminx,justinzh}@andrew.cmu.edu
[2] AsiaInfo, Beijing, China
zhu198422@yahoo.com.cn

Abstract. Social Network is being a popular word in WEB 2.0 era. Various social network websites connect us and our friends together. A buzzword which describes the idea is "*communicate with anyone anywhere*". By using social network websites like *Facebook* or *blog*, we can find out the communities which share the same interests with us. In the academic area, people sharing the similar research interest can also form community which is one type of social network. In this paper, we are trying to identify the people with similar research interest and help them build up such kind of social network. The information sources mainly come from the research papers and articles which can be searched freely from some research organizations' websites, such as *ACM web Portal* and *IEEE web Portal*. In addition, we would like to find out more personal information about experts in particular research field by using common search tools such as *Google* and *Live Search*.

Keywords: Social Networks, Citation Network, Researcher Community.

1 Introduction

Social network would be one of the most successful Internet applications in Web 2.0 Era [1]. *Facebook*, *MySpace* and many other social network websites keep us with our family and friends closer than ever before. Furthermore, we can easily find new friends and related resources through current friends. We can borrow ideas from the model of these social networks to help improve the communication among research communities. On one hand, their models are valuable for us to refer to when solving similar problems, such as how to identify and build up communities. On the other hand, if researchers want to find useful information such as important papers in the field that they are not familiar, there are few good mechanisms or systems to help them. Based on these two facts, we are going to apply the architecture of social network to build up a researcher network to help novice researchers to locate useful information.

In fact, the focal point is not how to get information for users, but how to show the information to them. By exploring the websites, you can find out the information you want, but most of the information is irrelevant to their interest. Search engines receive our keywords then fetch related information from their database and rank each piece of information to show a better view of the results to users. But actually, even though

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 421–427, 2008.

search engines provide us the information we need most of time, it returns too many options. For research community, we need definite information of particular person, such as his name, email address, background. It would be very convenient if we can provide exact information about the particular person for our research users and help them identify the experts in their research interest.

In the following part, we are going to show how to reuse current available resources to build up such a system. By reusing these search tools, we would like to show the users definite information about the particular researcher and the network of the researcher community in which individuals share the same interest. The basic steps of this process are:

1. Build up relationship network of authors of the papers extracted from paper source websites by searching keywords.
2. Extract information of each author using general search engines like *Google*, *Live Search*.
3. Re-organize the information and integrate the information into the relationship network and add new relations from the information of each author.

For the whole process, we mainly use search engines to extract information, apply webpage parser tool to analyze webpage information and use social network model to build up the architecture.

The outline of the rest of the paper is as follows. We first discuss related work in the fields of research paper recommendation and the principle of search engines in second part. We then talk about the process of building up the system, what information we are going to collect about the researchers, how to organize the information and how to build up the researcher network.

2 Related Work

Actually, there are many tools and services to find the scientific literatures on the web. What we use most is search engines like *Google*, especially *Google Scholar* search. By using search engines, we can easily locate a specified paper. Another common media is through the commercial scientific literature database websites, like *ACM* portal and *IEEE* portal. There are also some free portals for searching the citation index, like CiteSeer. There are many papers on how to build up a system to recommend research papers. In [2], the authors apply collaborative filtering to recommend research papers by using citation web between papers to create the ratings matrix. Lawrence, Bollacker, and Giles have a set of papers [3, 4, 5, 6] about *CiteSeer* on how to build a digital library for research papers. He et al. proposed a method based on citations to retrieve scholarly publications [8]. In this paper, we also would like to introduce some basic background information about social network and principle of search engines for better understanding.

2.1 Social Network

A social network is a graph, G = (V, E), where V is a set of nodes representing persons and E is a set of edges (V * V) representing the relationships between the

corresponding persons, such as values, visions, idea, financial exchange, friends, kinship, dislike, conflict, trade, web links, sexual relations, disease transmission (epidemiology), airline routes, etc. The resulting structures are often very complex. Social network analysis views social relationships in terms of nodes and ties. Nodes are the individual actors within the networks, and ties are the relationships between the actors. There can be many kinds of ties between the nodes. Research in a number of academic fields has shown that social networks operate on many levels, from families up to the level of nations, and play a critical role in determining the way problems are solved, organizations are run, and the degree to which individuals succeed in achieving their goals. In its simplest form, a social network is a map of all of the relevant ties between the nodes being studied. The network can also be used to determine the social capital of individual actors. These concepts are often displayed in a social network diagram, where nodes are the points and ties are the lines. Nodes are connected together because of relations. Relation between two nodes can be abstracted as a connection. This relation can be anything which can link two individuals together, but most of time, we build up such diagram because of our interest, so in fact, we just put the relations that we are interested into the diagram, and it also make the diagram understandable. In the case of our researcher network, each node would stand for one author in the field that users interested in, and the relation that link two researchers can be co-authors of the same paper, or from the same organization or institute, the same nationality or even the same age.

2.2 Principle of Search Engine

Since we are going to use similar algorithms to explore the interested authors of papers in a particular field, we would like to give a brief introduction to the principle how search engines work. One type of search engines uses the technology called *spider* or *crawler* to expand their network [9]. This *spider robot* visits a webpage, reads it, and then analyzes the links inside of the webpage and later follows these links to other web pages. In this way, the spider can expand to almost anywhere on Internet since we believe that all the websites are connected together.

In the case of researcher network, we use a similar way. We are going to fetch papers by particular keywords, and then use the returned paper list to extend the range of the papers.

1. Send keywords to paper sources web portal.
2. Analyze the paper list returned by these portals and extract the reference part of each paper.
3. Follow the references to find out other papers.
4. Repeat step 2 and 3 for several times to reach an extension level.
5. Combine all of the return lists
6. Find out the authors of these papers
7. Use portals to find out how many times each paper has been cited or referenced.
8. Find out the importance (weights) of each author by adding the number of times of reference of all his paper.

9. Rank the authors by the weight.
10. Find out if the authors in the list have any co-authors, if yes, then link them together.

3 Author Information Reorganization

For a novice to a new research field, it would be very helpful if he or she can quickly find out who are the famous researchers in this field. After knowing the researcher's information, new comers can communicate with them via email or other media and thus can benefit quickly through discussion with experts. In addition, even for veteran, they need a platform to know other researchers in their field to get more effective communication and share creative ideas and information with each other. Actually, it would be quite constructive if we can help to build up the research community in specified area automatically. If we want to build up such community identified by users' interest, we need at least to collect some information about the main authors in the research field. In the following part, we are going to discuss what information of authors we are going to collect and how we organize this information to form a community.

3.1 Information Collection

The information about the famous experts in a field is more interesting for novice than that of less famous expert. Based on this assumption, we are going to collect information of limited number of people who are relatively more important in the field. But we do not eliminate the function for users to search information of a specified particular person.

Table 1. Information About the Author

Profile information	Identity information	Name
		Age
		Gender
		Phone
		Email
		Organization/ company/ institute
		Title/ Position
		Address
	Accessory information	Biography
		Personal Homepage
		Instant Messaging
		Blog
		Social network websites. Facebook, Hi, Orkut,
		Interest
		Blog article
		Newsroup, BBS, Google group
Professional information		Papers
		Projects

The information about author can be categorized into two classes: one is the personal profile information, which includes identity information for the person, and another one is the professional information in the field, such as papers and projects. For professional information, we focus on the information related to certain field author. Examples are the paper author published the projects and research interest, relationship with other professionals. The information we are going to collect is illustrated in Table 1.

3.2 How to Collect Information

Before colleting the information for experts, we need to know who the experts are, in other words, we need to locate the important authors in the field specified by the user. By using paper resources portal like *ACM Portal*, *IEEE Xplore*, *CiteSeer*, etc, we can get a list of most relevant articles in the field. Since the number of articles returned by search would be probably very large, it is better to get part of the list. One good strategy is to get top 20 articles returned by the search engines ordered by relevance. From this candidate list, we use spider algorithms used by search engines to extend the range by using the reference list of each paper. The reason to use extended paper list but the return list directly from the search is due to the assumption that the papers from reference list have stronger relation than papers from search result list. However, if we just use reference list to extend the range, we could only find the papers that these candidate list cited, which means that the papers published after the candidate list would be excluded. In this sense, it would be better to use *CiteSeer* to find out those papers that cite the candidate list papers. Through this forward and backward extension, we can get a relatively complete list of important papers.

If each paper has 10 citations on average, there would come out much less than 200 new articles because we can anticipate that the articles in the candidate list would be highly related in the sense that they are in the same field as they are returned by the same search keywords. After we get this articles list, we can extract the authors from the articles, and then we get the authors list. We can then just use the author name to search in the database to find out how many papers have been published by each author and how many times have been cited for each paper. After that, we calculate the weight for each author by summing the number of citations of all his papers.

Eventually, we obtain n_{ij} which is the number of citations for paper *j* by author *i,* and w_i is the weight for each authors. Then $w_i = \sum_{j=1} n_{ij}$.

After we calculate out the weight for each author, we rank the authors by their weights, and get the ranking of importance for each author.

The next step is to find out the information specified in Table 1 for each author in our ranking list. We can use Google and other search engines to help finish this task. The process is described as follows.

From the paper we searched through the paper resources website, we extract the author related information and build up the author list. By applying some strategy indicated in the article [10, 11], we can improve the quality of results dramatically. We can find information of these researchers quite easily by analyzing the results returned by Google or other search engines.

3.3 Organize the Information

After gathering all the information, the next step would be to demonstrate the information to the users. Before showing the messy information to users, we have to re-organize the information so that it has a good format and thus make users easy to find out what they care.

The way we utilize to organize the information is to put the authors into a chart, and for users that we find if they have any relationship with each other, we would draw a line between them. This relation would be: the co-authors of particular paper, from the same organization, or even graduate from the same universities.

From the final network diagram, we give a rank of authors in the field specified by users, (actually specified by keywords), and authors are linked together by line. When click the authors, we can easily find out all his/ her information in Table 1. When click the connection, we can find out the relationship between two authors. Furthermore, we can get a macro-view of the whole community, which is supposed to be separated to several groups sharing some geological or other characteristics.

4 Conclusion and Future Work

In this paper, we discussed about how to efficiently reuse current available resources to build up a system that help to form a research community in a particular field. These resources are all from free search engines services and paper search portals. We believe that such kind of system can help people especially those novice quickly identify the experts in the field that they are interested in, and thus give them a direction where they can turn for help when they encounter problems. It is also a good platform for people to have better academic communication and exchange great ideas.

We are currently developing such the system and discussing about how to provide a better interface for users to quickly locate their wanted information. In addition, we would also like to add model to recommend important papers in particular fields, plus the process of development of the particular field or topic, combining with which to provide more convenient service for users.

References

1. Oreilly, T.: What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. Communications & Strategies (1), 17 (first quarter, 2007)
2. McNee, S.M., Albert, I., Cosley, D., Gopalkrishnan, P., Lam, S.K., Rashid, A.M., Konstan, J.A., Riedl, J.: On the recommending of citations for research papers. In: Proceedings of the CSCW 2002 (2002)
3. Lawrence, S., Bollacker, K., Giles, C.L.: Indexing and retrieval of scientific literature. In: Eighth International Conference on Information and Knowledge Management, CIKM 1999, Kansas Cite, pp. 139–146 (1999)
4. Lawrence, S., Giles, C.L., Bollacker, K.: Digital libraries and autonomous citation indexing. IEEE Computer 32(6), 67–71 (1999)

5. Giles, C.L., Bollacker, K., Lawrence, S.: CiteSeer: An automatic citation indexing system. In: Witten, I., Akscyn, R., Shipman III, F.M. (eds.) Digital Libraries 1998 - The Third ACM Conference on Digital Libraries, Pittsburgh, PA, June 23-26, 1998, pp. 89–98 (1998)
6. Bollacker, K., Lawrence, S., Giles, C.L.: CiteSeer: An autonomous web agent for automatic retrieval and identification of interesting publications. In: Agents 1998 (1998)
7. Blosser, G., Zhan, J.: Privacy-preserving social networks. In: Proceedings of IEEE International Conference on Information Security and Assurance (ISA 2008), April 24-26, 2008, Busan, Korea (2008)
8. He, Y., Hui, S.C., Fong: Citation-based retrieval for scholarly publications. Intelligent Systems, IEEE 18(2), 58–65 (2003)
9. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: Bringing order to the Web (submitted for publication)
10. Boswell, W.: Google People Search (2008), http://websearch.about.com/od/peoplesearch/tp/googlepeoplesearch.htm
11. Perishable Press, Smooth Operators: Sharpen your Google Search Skills (2008), http://perishablepress.com/press/2007/04/10/sharpen-your-google-search-skills/
12. Boswell, W.: Top 10 Ways to Do a Free People Search On the Web (2008), http://websearch.about.com/od/peoplesearch/tp/peoplesearch.htm

Online Gaming Perpetrators Model

Yungchang Ku[1] and Saurabh Gupta[2]

[1] Department of Information Management, Yuan Ze University,
135, Far-East Rd., Chung-Li, 320 Taoyuan, Taiwan, R.O.C.
ycku@saturn.yzu.edu.tw
[2] Department of Communication and Computer Engineering,
The LNM Institute of Information Technology,
Rupa ki Nangal, Via-Kanota, Jaipur, (Rajasthan) India
gupta_saurabh1985@yahoo.co.in

Abstract. Online gaming carries high economic value in e-business. Online gaming players have reasonable expectation in online gaming property and associate them with values in real world markets. This phenomenon emerges as a new research issue related to the development of e-society. This study briefly digests the characteristics of online gaming perpetrators and proposes the perpetrators model in order to explain why online gaming crime happened. The perpetrators model demonstrates that the online gaming crime will happen when valued virtual property, online gaming perpetrators and absence of protection mechanism are cohesive at the same time.

Keywords: online gaming crime, virtual property, perpetrators model.

1 Introduction

With the progressive proliferation of Internet technologies, various business models have been very successful, especially in online games. Online games have become the biggest entertainment industry than any other Internet business models. According to the DFC Intelligence Report, the worldwide online game market grows from $3.4 billion in 2005 to be forecasted over $13 billion in 2011 [1]. The advantages of online gaming are not limited only in amusements, but also in the value of virtual properties. For example, a high level character in the famous online game “Lineage”, which was developed by NCSoft.com of South Korea, will be bided at least 800 USD in yahoo auction house. Same price is bided for mahjong account with 3 million mahjong currency in Acer FunTown online will trade in the market respectively. Online gaming digital items (e.g. currency, swords, armor, jewelry, etc) are not only valued in certain online game, but also value-add in real world economic markets. The transactions in online gaming marketplaces foster new business models and advantageous to the development of Internet economy. There is no doubt that online games have remarkable advantages to the financial growth worldwide. Consequently, to explore the online gaming phenomena and investigate relative issues may emerge new path to advance the development of Internet society.

Online gaming (or massively multiplayer online role-playing game, MMORPG) is a genre of online computer role-playing games (CRPGs) for entertainment with one

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 428–433, 2008.

another over the Internet [2][3]. Since more and more people join online games, many players have found real world value of the digital items in online gaming marketplaces. In the online games, the characters are developed by players over time. It is true for players that if they can obtain useful or powerful digital items for their characters to grow in strength. Moreover, lots of players will also pay much money in online gaming marketplaces to obtain specific digital items for their characters. Players take online gaming digital items as valued virtual property for granted. This phenomenon further creates high potential business advantages from bidding, selling, trading, and exchanging those online gaming virtual properties. Unfortunately, the huge amount of online gaming virtual property transactions attracts perpetrators' attention to commit online gaming crimes. Perpetrators will develop specialized crime schemes such as hacking, cheating, and scamming to increase their profits [4][5]. In order to understand online gaming crimes, the previous work investigated the characteristics of online gaming crimes from judicial documents in Taiwan [2][6]. This research extends to propose perpetrators model to explain why online gaming crimes happen.

This paper is organized as follows: First, the introductory statement of online gaming phenomenon is given. Next, the security issues of online gaming crimes and its characteristics are presented. To explain why online gaming crimes happen, perpetrators model of online gaming crimes is proposed in section 3. Finally, the conclusions and issues of future research are discussed in section 4.

2 Security Issues and Characteristics of Online Gaming Crimes

2.1 Online Gaming and Emerged Virtual Property

As we mentioned above, online gaming or online games are the specific genre of computer games that enables people to play online by LAN, Internet or by other telecommunication medium which is shown in Figure 1.

Online gaming are also known as "massive multiplayer online games", "persistent state worlds" or "massively multiplayer online role-playing games" and represented

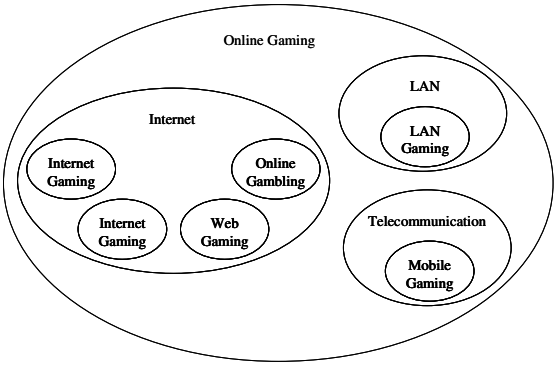


Fig. 1. The classification of online games related to communication medium. [2]

Table 1. Characteristics of online gaming

Characteristics	Description
Network Connection	Connected to host server through Internet
Player Authentication	UserID and Password
Game Objectives	Accumulate virtual property through skillful game play to reach game objectives.
Number of Players	A large number of players, playing the same instance of the game, can compete with one another.
Payment for Use	Pay for network connection time by buying a card associated with a certain amount of connection time interval via a serial number on the card.

by large, sophisticated, detailed and fantastic scenarios in evolving and narrative virtual world [7]. Five characteristics of online gaming are shown in Table 1.

Online gaming are leisure phenomenon in e-society and very popular among millions of players. For example, the number of online gaming players is expected to surpass 59 million in China by 2008 [8]. In Korea and Taiwan, there are at least 3 million players in playing Lineage. Within online gaming, players themselves become game characters to manage different activities, range from hunting, fighting, transacting, making friends and interactive communication, etc [9]. Online gaming players also develops different strategies to obtain specific digital items which can be used to enhance the abilities, skills or status of gaming character in game society. These strategies include working with friends to hunt high level monsters, exchanging with others or trading to sellers by real money. Thus, the behavior of hunting, exchanging and trading, values up the worth of digital items and emerges the importance of virtual property in online gaming.

Online gaming players treat digital items of gaming character as virtual property. Virtual property is persistent computer code that the owner has certain powers to control and manage it. The five indicia of virtual property are rivalry, persistence, interconnectivity, secondary markets and value-added by users [10][11]. Players have reasonable expectation in these digital items and associate them with values in real world markets. This phenomenon derives virtual property business model and explicitly claims that virtual property has real world value [12]. Therefore, an economy emerges in a game world with the following characteristics:

(1) Persistence: The software maintains a record of the state of the world and the resource possessions of the players, regardless of whether or not the game is "in session" for any user.
(2) Scarcity: Users must expend “real” resources such as time and money to obtain goods and/or services in the synthetic world.
(3) Specialization: Availability of the resources to players must vary. For example, a participant whose character has skills could have the ability to make swords, while other players would have to purchase them. Because this results in comparative advantage, complex trade relationships and a division of labor result.
(4) Trade: Users must be able to transfer goods and services to and from other users.

(5) Property rights: The world must record which goods and services belong to which user's identity and the code must allow that user to dispose the goods or service according to his needs.

2.2 Online Gaming Crime

Online gaming crime is one kind of computer crime. The key entities of online gaming crime include individual, a computer and a network game. Perpetrators and some online game player use illicit or immoral ways to gain advantages from online games. Examples of online gaming crime are: theft, fraud, robbery, kidnapping, threat, assault and battery, destruction of property, counterfeiting, receipt of stolen property, privacy violations, software piracy, extortion, gambling, and so on. Most of the cases can be attributed to theft and fraud. Table 2 depicts the characteristics of online gaming perpetrators in our previous works [2] [6].

Table 2. Characteristics of online gaming perpetrators

Characteristics	Description
Gender	90% in male
Age	Between 21 and 25
Profession	Main perpetrators include students, workers, and military.
Crime Scene	Internet café
Crime Type	Most crime scene attributed to theft and fraud
Loss of each case	Average loss of each case is $459 U.S. dollars
Method categories	Identity Theft, Social Engineering, Hacking Tools or System Weakness, Force or Revenge, and Unrecognized

3 Online Gaming Perpetrators Model

Online gaming virtual property has economic values in the marketplace. Most online gaming players trade the digital items with legitimate transactions. Perpetrators commit online gaming crimes to get digital items illegally in absence of protection mechanism. Online gaming perpetrators model can be proposed as Figure 2.

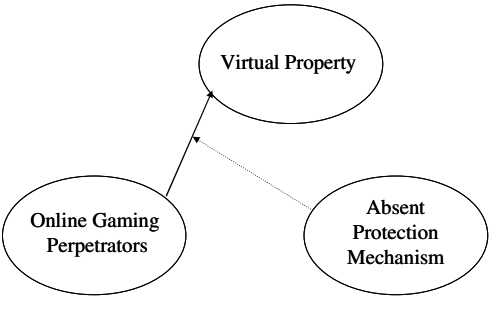


Fig. 2. Three elements of online gaming crime happened

In Figure 2, when the three elements are cohesive, then online gaming crime may be happen. These three elements are virtual property, online gaming perpetrators and absence of protection mechanism.

(1) Virtual Property (VP): In online gaming, virtual property includes the user's account and digital items with marketplace value. Because of the value, online gaming virtual property become marked target to potential perpetrators.
(2) Online Gaming Perpetrators (OGP): For usual online gaming players, they will spend much time or more money to get high level characters or digital items. However, for perpetrators, they won't spend any efforts to advance high level characters or useful digital items in online gaming instead of using illegal methods. Perpetrators will try their best to set scams to get benefits from online gaming virtual property. They will also pay close attention to search high value of online gaming digital items and the weakness of its owner for crime. Therefore, online gaming perpetrators will reveal their ambitions and skills in getting high value online gaming digital items from others.
(3) Absence of Protection Mechanism (APM): There is no doubt that virtual property set opposite side to online gaming perpetrators. The perpetrators will also show their effort in getting virtual property. If there do not exists any protection mechanism on virtual property side, the probability of being target virtual property will be raised.

The probability of being target virtual property by online gaming perpetrators is showed in (1)

$$P\,(Target\ VP) = \frac{Capable\ OGP}{APM} \tag{1}$$

In (1), P(Target VP) means the probability of being target virtual property, Capable OGP means capable online gaming perpetrators, and APM means absence of protection mechanism against online gaming perpetrators. If the force of Capable OGP is greater than APM, the probability of virtual property being target is increasing. Hence, The lower value of Target VP, even close to zero, is the better.

4 Conclusions

Online games have become the biggest entertainment industry than any other Internet business models. There is no doubt that online games have remarkable advantages to the growth of financial worldwide. This research discusses the values of online gaming virtual property to the online gaming marketplace. However, the online gaming perpetrators and crimes breach the development of online gaming environment. To understand the online gaming crime, this study briefly digests the characteristics of online gaming perpetrators and proposes the perpetrators model to explain why online gaming crime happened.

In perpetrators model, the online gaming crime will happen when virtual property, online gaming perpetrators and absence of protection mechanism are cohesive at the same time. In the future research, the qualitative case study will be held to understand the nature of perpetrators model.

References

1. DFC Intelligence Report, Analyst: Online Game Market $13 Billion by 2011 (2006), http://www.gamasutra.com
2. Chen, Y.C., Chen, P.S., Hwang, J.J., Korba, L., Song, R., Yee, G.: An Analysis of Online Gaming Crime Characteristics. Internet Research 15(3), 246–261 (2005)
3. Wikipedia, Massively Multiplayer Online Role-Playing Game, available: http://en.wikipedia.org/wiki/Massively_multiplayer_online_role-playing_game
4. Chen, Y.C., Chen, P.S., Song, R., Korba, L.: Online Gaming Crime and Security Issues – Cases and Countermeasures from Taiwan. In: Proceeding of the 2nd Annual Conference on Privacy, Security and Trust, NCR 47401 (2004)
5. Bardzell, J., Jakobsson, M., Bardzell, S., Pace, T., Odom, W., Houssian, A.: Virtual Worlds and Fraud: Approaching Cybersecurity in Massively Multiplayer Online Games. In: Proceedings of DiGRA 2007 Conference, pp. 451–742 (2007)
6. Ku, Y., Chen, Y.C., Wu, K.C., Chiu, C.: An Empirical Analysis of Online Gaming Crime Characteristics from 2002 to 2004. 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. 34–45. Springer, Heidelberg (2007)
7. Kolo, K., Baur, T.: Living a Virtual Life: Social Dynamics of Online Gaming. Game Studies 4(1) (2004) (online journal)
8. Xinhua News, China's Online Game Players to Reach 59 million in 2008, Survey Finds, http://news.xinhuanet.com
9. Whang, L.S.: Online Game Dynamics in Korean Society: Experiences and Lifestyles in the Online Game World. Korea Journal 43(3), 7–34 (2003)
10. Blazer, C.: The Five Indicia of Virtual Property. Pierce Law Review 5(1), 137–161 (2006)
11. Fairfield, J.: Virtual Property. Boston University Law Review 85, 1047–1102 (2005)
12. MacInnes, I.: Property Rights, Legal Issues, and Business Models in Virtual World Communities. Electronic Commerce Research 6, 39–56 (2006)

Proposal for a Multiagent Architecture for Self-Organizing Systems (MA-SOS)

Niriaska Perozo[1], Jose Aguilar[2], and Oswaldo Terán[3]

[1] Unidad de Investigación en Inteligencia Artificial, UCLA, Barquisimeto 3001-Venezuela
nperozo@ucla.edu.ve
[2] CEMISID, Facultad de Ingeniería. Universidad de los Andes, Mérida 5101, Venezuela
aguilar@ula.ve
[3] CEMISID, Facultad de Ingeniería. Universidad de los Andes, Mérida 5101, Venezuela
oteran@ula.ve

Abstract. This work investigates the trade-off between individual and collective behavior, to dynamically satisfy the requirements of the system through self-organization of its activities and individual (agent) adaptability. For this purpose, it is considered that each agent varies its behavioral laws (behavior-switching) dynamically, guided by its emotional state in a certain time instant.

Keywords: Emergent, SelfOrganizing Systems, Swarm Intelligence.

1 Introduction

Nowadays, to level multiagent systems (MAS) it is advisable to have a general agent architecture able to emulate human behavior, capable of modelling self-organizing systems which can adapt dynamically to their environment. For this purpose, we propose a hybrid generic architecture, where agents are able to have reactive or cognitive responses, depending on the received stimulus, and to collectively generate emerging behavior, with the goal of improving individual and social performance. Besides, considering cognition as a social phenomenon, an agent would develop individual behavior based on the collective behavior of its neighbors. Moreover, the architecture adds the characterization of the emotional state of the agents.

2 Theoretical Aspects

A cognitive architecture is a generic computational model for studying behavior and cognition at the individual level. It provides an agent with decision-making mechanisms. Among the cognitive architectures we have: **SOAR** [4]; **CLARION** [5, 8] and **ACT-R** [7, 8]. SOAR is the most complete one. It includes *working and long-term memory,* and learning mechanisms (*chunking, reinforced knowledge, etc.*). For emotional computing considering agent dynamic behavior-switching see [1, 3, and 10].

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 434–439, 2008.

3 Generalities about the Proposed Architecture

Fig. 1 shows how the learning process and acquisition of knowledge takes place in the architecture. An agent increases its knowledge through an individual learning process. It interacts (socializes) through its environment and directly with other agents using local information. A "Bottom-Up" mechanism allows emerging of collective explicit knowledge. Additionally, a feedback "Top-Down" process promotes individual learning of this collective knowledge.

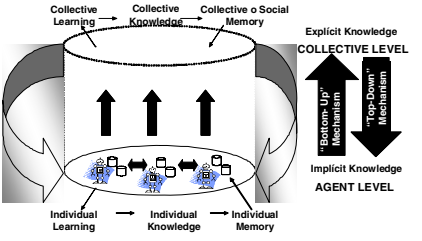


Fig. 1. Types of Knowledge and Learning in MA-SOS

It consists of the following phases involved in a circular cause-effect process of general knowledge management that reflects the process of creation, conversion, integration and diffusion of knowledge according to [6]: a) **Socialization;** consists of sharing experiences through local interactions, and requires **turning implicit knowledge into explicit transferable concepts**. b) **Aggregation;** the agent **creates trustworthy explicit knowledge** through exchange of points of view, meetings, etc. c) **Appropriation;** consists of **translating explicit knowledge into the implicit kind**.

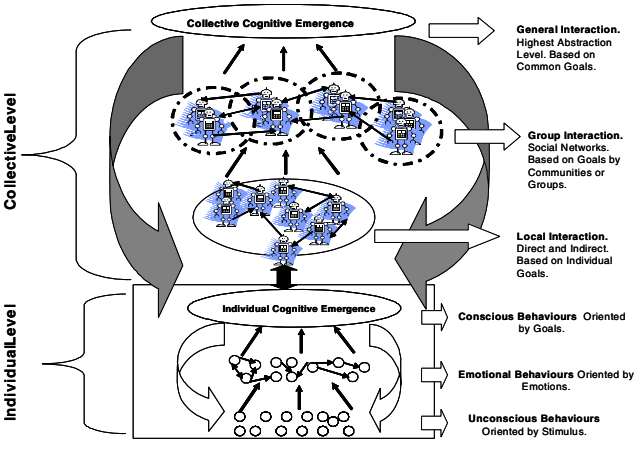


Fig. 2. Multiagent Architecture for Self-Organizing Systems

3.1 Multiagent Architecture for Self-Organizing Systems

The proposed architecture allows emerging coordination among hybrid agents. It is divided into two levels: *individual and collective* levels (see Fig. 2). C**ollective cognitive emergence** arises from three interaction levels: **Local Interaction Level,** which might be direct or indirect (via the environment); **Group Interaction Level,** involving social networks or structured groups; and, **General Interaction Level,** which includes the whole set of agents.

On the other hand, **individual cognitive emergence** consists in generating cognitive emergence imitating the way in which neurons act, when generating a range of behavior from unconscious to conscious [9]. Inspired by this, agent's behavior is modeled at three levels, each activated or inhibited depending on the agent's objective: **Unconscious or reactive Behavior**; **Emotional Behavior;** and **Conscious Behavior**.

3.1.1 Components of the Architecture at the Collective or Social Level

At this level the architecture consists of (see Fig. 3): a) **Hybrid Agent:** This agent reacts and reasons depending on perception, emotions and the state of the environment. b) **Feedback Mechanisms ("Bottom-Up" and "Top-Down" Approach):** Interaction among components could be [2]: Positive (to promote the creation of structures and changes in the system) or Negative (to offset positive feedback and help stabilize the collective pattern) Feedback. c) **Mean Field:** It represents the area circumscribed and delimited by agents within the **environment** for coordinating behavior.

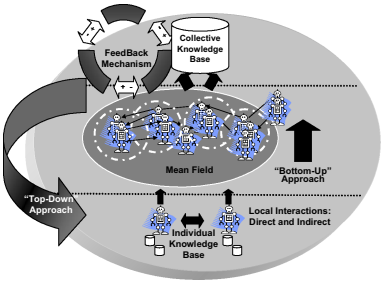


Fig. 3. Components of MA-SOS at Collective Level

3.1.2. Architecture Components at the Individual Level

The architecture at an individual level is made up of 4 components (see Fig. 4): **Reactive, Cognitive, Behavior** and **Social**. In order to exploit diversity and to favor the development of collective cognitive emergence, each agent can have hybrid behavior: reactive, emotional-reactive, and cognitive-reactive, among others.

a) Reactive Component: It produces the agent's reactive behavior. **Reactions Selector:** selects among different behavioral routines (reactive behavior) to be executed according to the agent's emotional state.

b) Behavior Component: It favors the agent's adaptation to its environment, creating an internal model of the outside world. Each agent's decision will be based on its

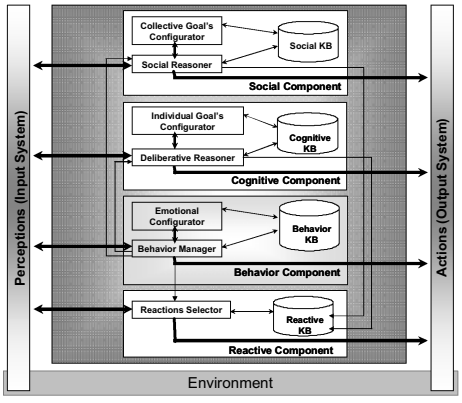


Fig. 4. Components of MA-SOS at Individual Level

individual and collective objectives, its emotional state and acquired individual and collective knowledge. The types of behavior to handle are: to imitate, to react and to reason, linked to an emotional state (positive or negative) (see Fig. 5). **Emotional Configurator:** It is the component manipulating agents' emotions. Emotions are considered as signals and evaluations that inform, modify and receive feedback from a variety of sources including reactive, cognitive processes and other agents (social processes). **Behavior Manager:** It is the component managing the behavior-switching mechanisms. Its objective is to suggest a type of behavior based on its emotional state, its goals and the situation of its neighbors (social situation) and environment in general. Knowledge associated with the agent's emotion management is stored in the **Behavior KB.**

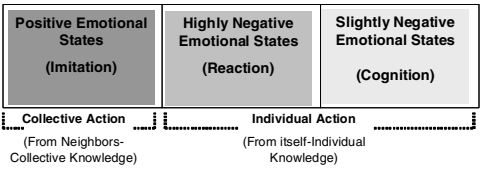


Fig. 5. Positive and Negative Emotional States with its Associated Behavior

The role of emotions in this work is for the behavior selection (e.g., which behavior is convenient according to the current emotional state) based on the classification of agent's emotions. For this classification, each agent could have emotions in three different areas [10]: goal-based emotions [e.g., *Joy, distress; Hope, Fear*], other agent's actions [*e.g., Anger, Gratitude; Gratification, Remorse; Pride, Shame; Admiration, Reproach*] and tastes/attitudes towards objects or places [*e.g., Love, Hate; Like, Dislike*]; the idea is that each agent can have an emotional memory which allows it to put "tags" to its emotions in each one of these categorizations. Moreover, each agent will have a positive, negative or neutral attitude which will affect the intensity of the emotion.

However, determining if the emotional state is positive or negative and selecting the type of behavior are both based on the following: *"Negative affection can bias problem solving strategies in humans towards local, bottom-up processing; whereas, positive affection can lead to global, top-down approaches"* [10]. Thus,

```
Rule 1: If <Emotional State> is Positive then    <Prior-
ity_Social_Behavior>

Rule 2: ElseIf <Emotional State> is Slightly Negative then
             <Priority_Cognitive_Behavior>

Rule 3: ElseIf <Emotional State> is Highly Negative then
             <Priority_Reactive_Behavior>
```

c) Cognitive Component: It is in charge of producing the agent's cognitive behavior. **Individual Goal's Configurator:** makes the configuration of the agent's individual objectives and priorities. **Deliberator:** It's responsible for cognitive mechanisms (learning, reasoning) and intentional or deliberate decision-making, among others; the knowledge generated is stored in the **Cognitive KB.**
d) Social Component: It promotes conscience in the agents about the work and experience of other agents. Specifically, it takes advantage of experiences of others (Social Learning), i.e., to avoid a long process of individual learning of things which have been learnt by its neighbors. This component connects the collaborative collective learning and the individual learning. **Collective Goal's Configurator:** makes the configuration of the agent's collective objectives and priorities. **Social Reasoner:** selects which action to imitate and from which agent based on the collective goals; the knowledge about decisions taken by its neighbors is stored in the **Social KB.**
e) Other General Elements. Input System: It provides agents with information about the world they live in. This system passes perceptions (inputs) in a parallel manner to the reactive, behavior, cognitive and social components. All components interact with that input but it is the behavior component which must establish which component has higher priority for answering. **Actions:** They are rules of condition-action (if…then) or generated from a deliberative process. **Output System:** It must choose the action of the component indicated by the Behavior Manager in case there are several answers.

4 Conclusions

Our proposed architecture is split into two levels: *an individual level and a collective one*, trying to model systems which produce or have self-organizing behavior from local interactions of agents, generating collective knowledge from those interactions. We have developed a collective architecture based on 3 phases for knowledge management. Additionally, our architecture provides each agent with an emotional state. This allows us to have a multiagent architecture whose agent's behavior might better represent a wide range of human situations, since each agent could have multiple behaviors depending on its knowledge, emotions and social situation. Therefore, it offers an alternative model to represent and better understand self-organization and emergent processes in human environments. Further work in progress includes modeling systems like *Wikipedia, free software development and collective behavior of pedestrians.* These works aim at validating our architecture from a design point of view.

Additionally, further work also involves implementing a simulation prototype, for a specific problem, in order to instantiate each mechanism and component, validating the architecture from an implementation point of view.

References

1. Gadanho, S.: Learning Behavior-Selection by Emotions and Cognition in a Multi-Goal Robot Task. Journal of Machine Learning Research 4, 385–412 (2003)
2. Camazine, S., et al.: Self-Organisation in Biological Systems. Princeton University Press, Princeton (2001)
3. Imbert, R., De Antonio, A.: Agents that Combine Emotions and Rationality: a Context Independent Cognitive Architecture. WSEAS Transactions on Computers 4(9), 1202–1209 (2005)
4. Lehman, J., et al.: A Gentle Introduction to SOAR, An Architecture for Human Cognition: 2006 Update. University of Michigan (2006)
5. Sun, R.: Desiderata for Cognitive Architectures. Philosophical Psychology 17(3), 341–373 (2004)
6. Caraballo, N.: Gestión del Conocimiento: Aprendizaje individual versus aprendizaje organizativo. Intangible Capital 2(13), 308–326 (2006)
7. Anderson, J., et al.: An Integrated Theory of the Mind. Psychological Review 111(4), 1036–1060 (2004)
8. Sun, R.: Cognition and Multiagent Interaction, From Cognitive Modeling to Social Simulation. In: Sun, R. (ed.) Rensselaer Polytechnic Institute, Cambridge U. Press, Cambridge (2005)
9. Di Marzo, G., et al.: Self-organisation and Emergence in MAS: an overview. Journal of Informatica, Ljubl 30(1), 45–54 (2006)
10. Khulood, M., Raed, Z.: Emotional Agents: A Modeling and an Application. Information and Software Technology – Elsevier (2006)

Applying Text Mining to Assist People Who Inquire HIV/AIDS Information from Internet

Yungchang Ku, Chaochang Chiu, Bo-Hong Liou,
Jyun-Hong Liou, and Jheng-Ying Wu

Department of Information Management, Yuan Ze University,
135, Far-East Rd., Chung-Li, Taoyuan, Taiwan, R.O.C. 320
{ycku,imchiu}@saturn.yzu.edu.tw,
{s941624,s941618,s941718}@mail.yzu.edu.tw

Abstract. Inquire health information from Internet or virtual community is one of hot activities on the web. But no one can guarantee the treatments or remedy work or not for the health questioners. The present research proposes an Internet health information governance mechanism (IHIGM) for support the diseases control and health authority to do their efforts in Internet health information. In the experiment, the research takes "People Inquire HIV/AIDS Information from Internet" as example and explains the procedure of IHIGM. In the experiment result, the accuracy ratio of IHIGM can at least classify 85% positive HIV/AIDS Internet health information inquiry for intervention.

Keywords: text mining, HIV/AIDS, Internet health information.

1 Introduction

Since the mid 1980s, Information technology enables people with common interests and topics pleasurable communication to share experiences and opinions in Internet. There are various virtual communities developed, such as bulletin board systems (BBS), Internet forums, blogs, and wiki, etc. Virtual community is a kind of social synthesis that enables people to converge, discuss, share experiences, and develop social relationship through the Internet [18]. It is one of important key to advance the development of e-society with respect to the critical channel to real world society. Furthermore, virtual community gradually improves not only in knowledge sharing and communication between people, but also in daily life aspects such as economy, society, culture, education, and personal interaction [20][21].

Finding useful information to handle daily various problems is one of reasons for people in participating virtual community. Johnson (1997) considered that information searching is purposive to obtain useful information from summary data [14]. Considering health relative problems, people would rather inquire same illness condition and treatment via Internet than ask help directly from the doctors [10][12]. It is also true for people to discuss with the doctor about illness treatment information which he/ she collected from Internet [7]. The phenomenon tells that health

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 440–448, 2008.

information getting from Internet for reference will become more and more important than real world society. However, health information getting from Internet still lacks of treatment validation. Some people believed some remedy of certain illness but injuring themselves. Others distributed unofficial folk prescription to Internet to help people but causing potential dangerous in health. Thus, health information from virtual community or Internet should have had well governance mechanism to avoid causing serious injury to the health.

The present research will take "People Inquire HIV/AIDS Information from Internet" as example to demonstrate the Internet health information governance mechanism (IHIGM). This paper is organized as follows: First is the introduction to the concept of virtual community and health information getting from Internet. Section 2 provides related works in Internet health information researches and tells the reason why people inquire HIV/AIDS Information from Internet. The research method and proposed governance mechanism will be showed in section 3. Section 4 represents the experiment and section 5 provides the conclusions and possible future research directions.

2 Related Works

In clinic-patient relationship, there are many factors to let patients asking help from Internet, such as the afraid of hospital and illness situation, cognitive deficit of treatment, insufficient discussion with doctor, and mistrust to clinic description [17]. Patients would rather ask and collect same symptom and treatment from Internet to further understanding their illness. Therefore, inquiring health information from Internet or asking help from Internet become one of important communication topics in Internet, virtual communities or knowledge web sites [6]. Tardy and Hale (1998) mentioned that Internet health information is to help people decrease the uncertainty information of health issues. The health information will assist people to construct their health cognition and handle well when those in illness [19]. The related researches of Internet health information will include the quality evaluation of health information [11][16], trust in clinic-patient relationship [8], and the effectiveness of Internet health information to patients and people who coping with the patient. But, there is seldom research in identifying and classifying the need of the questioner who inquired health information from Internet.

For example, HIV/AIDS is a stigmatic illness, especially in Asia. HIV/AIDS will cause patient to death and always closely link with homosexual, prostitute and drugs. [5]. There is no well-development treatment to HIV/AIDS till now. When a person infected HIV/AIDS, the social stigma would tag along with him/ her directly. The social stigma would push those HIV/AIDS patients out from any normal social group. These patients would psychologically in worry, fear, shame, avoidance, guilt, and afraid to abandon. Then, they would close themselves, narrow down his social relationship and even to be the social marginal person. Thus, Internet becomes a critical channel for these HIV/AIDS patients and their family to get further treatment

information for themselves. In fact, Internet provides the patients with HIV/AIDS in higher obscuration, lower exposure, and multi-channel treatment and health information.

Furthermore, Borzckowski and Rickert (2001) processed the research to teenagers in America and discovered that teenagers would try to explore the Internet Information of sexually transmitted diseases [1]. Fox and Rainie (2002) also summarized that 33% health information questioners would inquire sensitive health information from Internet [13]. Moreover, people who live with HIV/AIDS patient would frequently employ Internet to inquire health information and special symptom of HIV/AIDS [15]. Similar research of Bull et al.(2001) discovered that 61% respondents would also browse sexually transmitted diseases and HIV/AIDS preventing resources through Internet [2]. More and more evidence reveal that HIV/AIDS patients, people who living with chronic ill, or possible infected persons accept Internet health information for their health inquiry needs [9].

3 The Internet Health Information Governance Mechanism

Correct diseases preventing strategies and health information for public is an important affair to the diseases control and health authority [3]. But in practice, there is no good way in dealing with considering whether Internet health information correct or not, even if in assisting HIV/AIDS need people from Internet. Thus, the authority is urgently needs to construct the assistance mechanism for HIV/AIDS questioners from Internet. The present research proposes the Internet health information governance mechanism (IHIGM) to respond to the practical needs. First, the concept of health information inquiry from Internet will be showed in Figure 1.

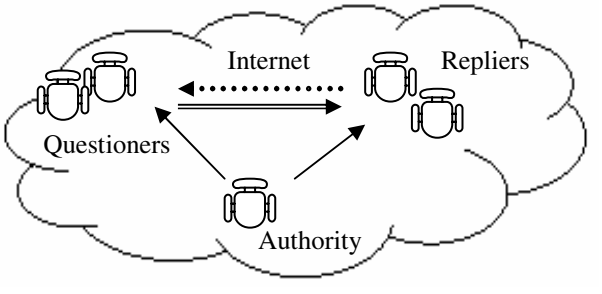


Fig. 1. Concept of health information inquiry from Internet

Figure 1 tells that:

1. Questioners: People inquiry health information form Internet to expect getting others treatments or suggestions.
2. Repliers: Some other people provide the treatments which based on their self-experience or remedy which hear form others to questioners. However, in Questioner-Replier relationship, no one can guarantee the treatments or remedy work or not for

the health questioners. More serious situation is the treatments or remedy may make the illness getting worse or cause people to death.
3. Authority: The diseases control and health authority should do their efforts in replying the health information needs for health questioners and check the accuracy of health treatments and remedy from Repliers.

After understanding the concept of health information inquiry from Internet and the relationship between Questioners, Repliers, and Authority, the detail description of the proposed IHIGM will be presented in the following.

The IHIGM will meet the need of the diseases control and health authority and designs as an automatic analysis mechanism. The IHIGM analyzes Internet health information based on the scenarios which are provided by domain expert from the authority. For example of HIV/AIDS, there are four scenarios that should be assisted in time, such as HIV/AIDS pathology, HIV/AIDS contagion with sex (e.g. one night stand, sex crossing), HIV/AIDS contagion without sex (e.g. living with HIV/AIDS people, blood or share syringe needle), and carrier or contagium. Thus, the framework of IHIGM will be showed in Figure 2.

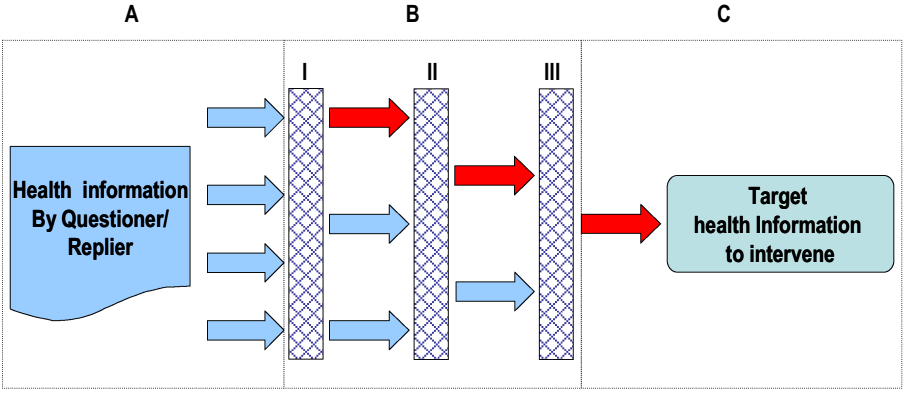


Fig. 2. The framework of IHIGM

Figure 2 presents the framework of IHIGM and divides into three parts. Part A is health information from Internet by Questioners or Repliers. There is a web crawler to gather related HIV/AIDS health information by Questioners or Repliers from Internet then into part B. In part B, the HIV/AIDS health information document will be recorded in stage I, such as Questioner/ Replier ID, title, content, link and time. The title and content of each document will be extracted by word segment technique [4] [22] [23] to find valuable terms to represent the document in stage II. Then in stage III, there is a baseline matching skill for four scenarios which provided by domain expert and mentioned above to classify the target health information to intervene. Finally, in part C, the authority can make responding strategies for the target health information to show efforts for health information in Internet.

4 The Experiment of IHIGM

4.1 Dataset and Analysis Procedure

The research develops a crawler program to collect HIV/AIDS health information inquiry from Yahoo!Knowledge (http://tw.knowledge.yahoo.com/) that employ "AIDS" or "愛滋" as keywords to gather matching inquiry document and get 1205 documents as dataset. Each document has its ID, title, content, link and time. Then, three domain experts provide the intervention judgments for there HIV/AIDS health information inquiry and reply based on four scenarios in (1) and (2) recursively.

$$Judgment = \begin{cases} positive\ to\ intervene, & 1 \\ otherwise, & 0 \end{cases} \tag{1}$$

$$Intervention = \begin{cases} 1, same\ posive\ \ judgment\ by\ three\ experts \\ 0, same\ negasive\ \ judgment\ by\ three\ experts \\ otherwise, discuss\ \ positive\ or\ negative\ by\ three\ experts \end{cases} \tag{2}$$

Furthermore, the title and content of each document by word segment technique to find valuable terms to represent the document. For example:

Original document:

怎麼辦！昨天晚上我和網友發生一夜情 她留言給我說：歡迎加入愛滋家族

(What can I do! Last night, I got one night stand with a friend met online. She gave me a message that welcomes to join AIDS family.)

Document terms extraction by word segment

"怎麼辦 (what can I do") "網友(a friend met online)" "一夜情(one night stand)" "愛滋家族(AIDS family)"

Next, the domain experts pick valuable terms from the positive intervention documents by *UCINet* of social network analysis to built baseline of scenarios for document matching. To avoid the same baseline terms includes in negative intervention documents, the mixed terms between positive and negative will be eliminated from the baseline and be showed in Figure 3. The result baseline of scenarios is showed in Figure 4.

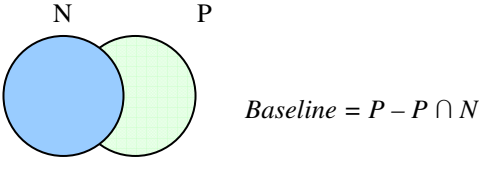
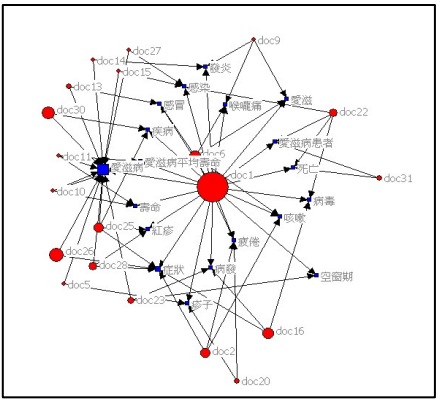
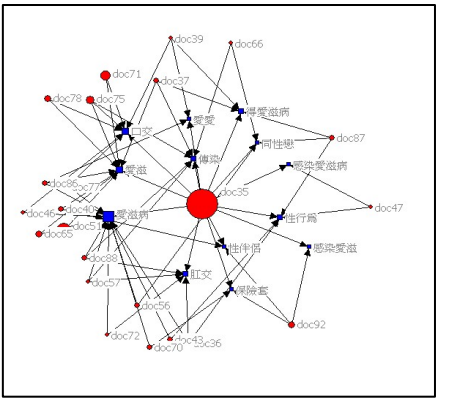


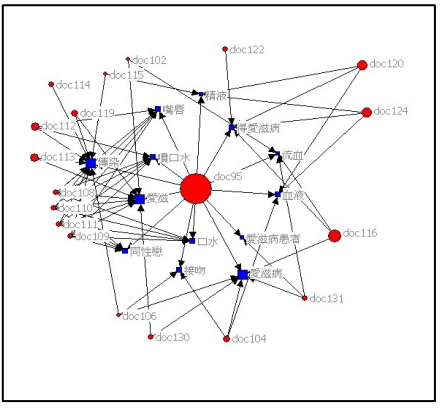
Fig. 3. Eliminating same positive and negative form baseline



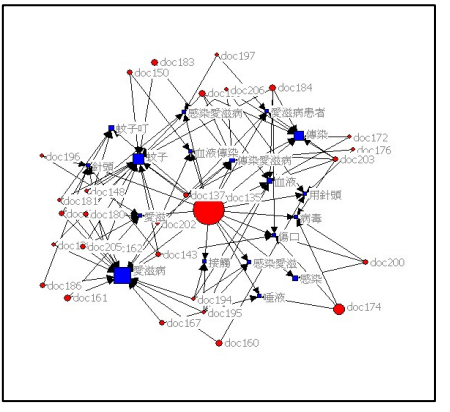
BSI: HIV/AIDS pathology



BSII: HIV/AIDS contagion with sex



BSIII: HIV/AIDS contagion without sex



BSIV: carrier or contagium

Fig. 4. Scenarios baseline terms extraction from positive document

4.2 Experiment Result

In analysis procedure, there is a baseline terms which picked by domain experts for positive health information document matching. Furthermore, each health information document has its valuable terms to represent itself. In this step, the research employs TFIDF and document similarity technique to classification model to classify the positive intervention health information document. TFIDF and Document similarity technique are showed in (3) and (4).

$$TFIDF(T_i, D_j) = TF(T_i, D_j) * \log \frac{|D|}{|DF(T_i)|} \tag{3}$$

where

$TFIDF(T_i, D_j)$ is the weight measure of word T_i within text D_j;

$TF(T_i, D_j)$ is the number of times word T_i occurs in text D_j;

$|D|$ denotes the total number of all text; and

$|DF(T_i)|$ is the number of texts in which word T_i occurs at least one.

$$Document\ \ similarity\ (\cos ine\ \ coefficient) = \frac{|X \cap Y|}{|X|^{\frac{1}{2}} \bullet |Y|^{\frac{1}{2}}} = \frac{\sum_{i=1}^{t} X_i \times Y_i}{\sqrt{\sum_{i=1}^{t} X_i^{\,2} \bullet \sum_{i=1}^{t} Y_i^{\,2}}} \tag{4}$$

where

X represents baseline terms of scenarios;

Y represents certain document terms which will compare with baseline.

Table 1. Accuracy ratio of training and test of 5-fold cross-validation

Training:

n	30% documents	Number of intervention by model	Number of intervention by domain experts	Same Judgment	precision	recall	Similarity Threshold
Fold 1	362	45	64	39	86.67%	60.94%	0.32
Fold 2	362	37	59	34	91.89%	57.63%	0.38
Fold 3	362	35	53	30	85.71%	56.60%	0.43
Fold 4	362	32	57	29	90.63%	50.88%	0.38
Fold 5	362	30	49	25	83.33%	51.02%	0.36

Testing:

n	30% documents	Number of intervention by model	Number of intervention by domain experts	Same Judgment	precision	recall	Similarity Threshold
Testing	362	41	63	36	87.80%	58.56%	0.38

Precison = Same judgment/ Number of intervention by model;
Recall = Same judgment/ Number of intervention by domain experts;
The Similarity Threshold will be produced by random selecting documents.

Therefore, the research can advance to process 5-fold cross-validation training and test to 1205 health information document. Each fold will random select 30% document to process and the accuracy ratio will be showed in Table 1.

After 5-fold cross-validation training and testing to 1205 health information document, the best accuracy ratio in precision is 91.89%, while the worst accuracy ratio in precision is 83.33%. The average accuracy ratio in precision is 87.65%. In testing, the precision is 87.56%. The result tells that the research model of IHIGM can classify at least 85% HIV/AIDS Internet health information inquiry and provide the authority to take advance assistance strategies.

5 Conclusions and Future Research Directions

Searching useful information to handle daily various problems is one of reasons for people in participating virtual community or Internet. Many researches pointed out that Internet or virtual communities become a critical channel for people to get further treatment information for themselves. This research takes take "People Inquire HIV/AIDS Information from Internet" as example to tell the phenomenon. There is a big problem for Internet health information that no one can guarantee the treatments or remedy work or not for the health questioners. More serious in Internet treatments or remedy may cause people getting worse illness or death. Thus, the diseases control and health authority should do their efforts in managing the health information in Internet and response the needs for those who have health information inquiry, especially in HIV/AIDS. The present research proposes an Internet health information governance mechanism (IHIGM) for support the authority to do their efforts in Internet. In the experiment result, the accuracy ratio of IHIGM can at least classify 85% positive HIV/AIDS Internet health information inquiry for intervention.

In future work, the following researchers can improve the model accuracy ratio of intervention Internet health information inquiry. Moreover, the researchers can also extent to develop useful response strategies to those who need assistance from Internet. More human relationship and social network analysis researches can also be held in the future to take advantage to "stigmatic illness".

References

1. Borzckowski, D.L.G., Rickert, V.I.: Adolescent Cybersurfing for Health Information: A New Resource that Crosses Barriers. Archives of Pediatrics & Adolescent Medicine 155(7), 813–817 (2001)
2. Bull, S.S., Mcfarlane, M., King, D.: Barriers to STD/HIV Prevention on the Internet. Health Education Research 16(1), 611–670 (2001)
3. Centers for Disease Control (CDC), Introduction to HIV/AIDS (Taiwan) (2005), http://www.cdc.gov.tw/
4. Chien, L.F.: PAT-Tree-Based Keyword Extraction for Chinese Information Retrieval. In: Proceedings of the 1997 ACM SIGIR, pp. 50–58 (1997)
5. Chuang, P., Liu, C.T.: AIDS Storm: The Stigma of HIV/AIDS Meands to the Client and the Public. Nursing Research 5(1), 52–64 (Taiwan) (1997)

6. Cline, R.J.W., Haynes, K.M.: Consumer Health Information Seeking on the Internet: The State of the Art. Health Education Research: Theory & Practice 16(6), 671–692 (2001)
7. Diaz, J.A., Griffith, R.A., Ng, J.J., Reinert, S.E., Fredmann, P.D., Moulton, A.W.: Patients' Use of the Internet for Medical Information. Journal of General Internal Medicine 17(3), 180–185 (2002)
8. Erdem, S.A., Harrison-Walker, L.J.: The Role of the Internet in Physician-Patient Relationships: The Issue of Trust. Business Horizons 49(5), 387–393 (2006)
9. Escoffery, C., Miner, K.R., Adame, D.D., Butler, S., McCormick, L., Mendell, E.: Internet Use for Health Information among College Students. Journal of American College Health 53(4), 183–188 (2005)
10. Eysenbach, G., Diepgen, T.L.: Patients Looking for Information on the Internet and Seeking Teleadvice. Archives of Dermatology 135(2), 151–156 (1999)
11. Eysenbach, G., Powell, J., Kuss, O., Sa, E.R.: Empirical Studies Assessing the Quality of Health Information for Consumers on the World Wide Web. The Journal of the American Medical Association 287(20), 2670–2691 (2002)
12. Achenbach, G.: Patient-to-Patient Communication: Support Groups and Virtual Communities. In: Lewis, et al. (eds.) Consumer Health Informatics – Informing Consumers and Improving Health Care, pp. 97–106 (2006)
13. Fox, S., Rainie, L.: Vital Decisions.: How Internet Users Decide What Information to Trust When They or Their Loved Ones Are Sick. The Pew Internet & American Life Project (2002), http://www.pewinternet.org
14. Johnson, J.D.: Cancer-Related Information Seeking, Baker & Taylor (1997)
15. Kalichman, S.C., Weinhardt, L., Benotsch, E., DiFonzo, K., Luke, W., Austin, J.: Internet Access and Internet Use for Health Information among People Living with HIV-AIDS. Patient Education and Counseling 46(2), 109–116 (2002)
16. Latthe, P.M., Khan, K.S.: Quality of Information on Female Sterilisation on the Internet. Journal of Obstetrics and Gynaecology 20(2), 167–170 (2000)
17. McMullan, M.: Patients Using the Internet to Obtain Health Information: How This Affects the Patient-Health Professional Relationship. Patient Education and Counseling 63(1-2), 24–28 (2006)
18. Rheingold, H.: The Virtual Community: Homesteading on the Electronic Frontier (1998), http://www.rheingold.com/vc/book/
19. Tardy, R.W., Hale, C.L.: Getting "plugged in": A Network Analysis of Health-Information Seeking among "Stay-at-home moms". Communication Monographs 65(4), 336–357 (1998)
20. Teo, H.H., Chan, H.C., Weib, K.K., Zhang, Z.: Evaluation Information Accessibility and Community Adaptivity Features for Sustaining Virtual Learning Communities. International Journal of Human-Computer Studies 59(5), 671–697 (2003)
21. Wang, F., Carley, K.M., Zeng, D., Mao, W.: Social Computing: From Social Informatics to Social Intelligence. IEEE Intelligence Systems 22(2), 79–83 (2007)
22. Sproat, R., Shih, C.: A statistical method for finding word boundaries in Chinese text. Computer Processing of Chinese & Oriental Languages 4, 336–351 (1990)
23. Sproat, R., Shih, C., Gale, W., Chang, N.: A Stochastic Finite-State Word-Segmentation Algorithm for Chinese. In: Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL), pp. 66–73 (1994)

Polarity Classification of Public Health Opinions in Chinese

Changli Zhang[1,4], Daniel Zeng[2,3], Qingyang Xu[1], Xueling Xin[1], Wenji Mao[2], and Fei-Yue Wang[2]

[1] College of Computer Science and Technology, Jilin University, China
[2] Institute of Automation, Chinese Academy of Sciences
[3] MIS Department, University of Arizona
[4] Artillery Command College of Shenyang, China
shenyangzcl@163.com

Abstract. Public health events with major consequences are occurring globally. Increasingly people are expressing their views on these events and government agencies' responses and policies online. Recent years have seen significant interest in investigating methods to recognize favorable and unfavorable sentiments towards specific subjects, including public health opinions, from online natural language text. However, most of these efforts are focused on English. In this paper, we study Chinese opinion mining in the context of public health opinions. We explore two complementary approaches—a Chinese opinionated word-based approach and a machine learning approach. We also conduct related comparative analysis and discuss the important role Chinese NLP techniques play in polarity classification.

1 Introduction

Many health-related topics are being debated globally, often with passion. Public health events are also occurring globally on a frequent basis and the public is expressing their views on these events (e.g., SARS, bird flu, etc.) and government agencies' responses and policies through various channels. Examples include: Should the euthanasia be made legal? What are the prevailing attitudes towards the quality of life among the AIDS patients? Increasingly, opinions and comments from individuals and groups are being voiced on the Web through various kinds of publicly-available user-generated contents (e.g. online forums and Blogs). Mining such public health opinions can be of great interest to and serve as an ideal testbed for social computing researchers.

Classifying opinions into positive or negative groups according to their sentimental orientation or polarity is the first step in opinion mining. Recently, an increasing number of researchers have devoted their efforts to investigating automated methods to recognize favorable and unfavorable sentiments towards specific subjects [1,2,3]. However, these studies have almost exclusively focused on English, with only a few exceptions [4,5]. In this paper, we conduct a systematic study of Chinese opinion mining, with emphasis on mining conflicting opinions on some public health issues.

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 449–454, 2008.

There are a number of technical challenges facing Chinese opinion mining. Very often, opinions in natural language are expressed in subtle and complex ways, especially in reviews written in Chinese. Negative reviews may contain many apparently positive phrases even if their authors maintain a strong negative tone, and vice versa. In this paper, we investigate two complementary approaches to Chinese opinion mining. One approach is based on Chinese opinionated words. The other is based on machine learning methods. We have also conducted a comparative study comparing these two methods using real-world data.

The rest of the paper is organized as follows: Section 2 presents our proposed Chinese opinion mining approaches. The experiments and results are presented in Section 3. We conclude the paper in Section 4 with a summary of our presented work and ongoing studies.

2 Chinese Opinion Mining Approaches

Chinese opinion mining poses some unique challenges, tendering many English-based opinion mining methods ineffective. 1)Processing Chinese text requires an additional step of segmentation, which can be complicated. 2)Chinese comparative and superlative sentences use degree adverbs, while English sentences mainly rely upon the suffix such as 'er' or 'est.'. 3)In many cases, Chinese text is more subtle and can lead to higher degree of ambiguity than English text.

2.1 Approach Based on Chinese Sentiment Words

2.1.1 An Overview

In this section, we propose an approach to classify Chinese views based on 'HowNet' sentiment words dictionary. HowNet is an online common-sense knowledge system unveiling inter-conceptual relations and inter-attribute relations of concepts as connoting in Chinese and English bilingual lexicons. Among other things, HowNet provides a comprehensive Chinese sentiment words dictionary, including word lists of: 1)3730 positive opinionated adjectives (e.g. 承认/admit) and 3116 negative opinionated words (e.g. 丑陋/ugly). 2)836 positive affective adjectives (e.g. 爱/love) and 1254 negative affective words (e.g. 悲伤/sad). 3)219 'degree adverbs' (e.g. 非常/very much). 4)some 'negation adverbs' (e.g. 否/not).

Note that HowNet quantifies its 'degree adverbs' according to their intensity degree. For example, '非常' has weight 2 and '很' has weight 1.5, though they have the same meaning as 'very much' in English. We make use of this information in our polarity classification algorithm as stated later.

Compared with English, Chinese opinion mining poses greater challenge for lack of adequate NLP resources and corpus. Furthermore, Chinese is an extremely subtle and complicated language. For example, the following three Chinese sentences have subtly different semantic orientations and strengths, though only a couple of different characters can be spotted.

sentence	degree	polarity
我很高兴	strong	positive
我不很高兴	weak	positive
我很不高兴	strong	negative

To deal with the subtlety of Chinese opinion classification, we adopt an intermediate step to identify sentence polarity before deciding upon the overall polarity of the entire Chinese review text. First, an entire review is decomposed into its constituting sentences. Then the polarity of each single sentence is synthesized to form the final classification result of the whole text. We propose this sentence-level (as opposed to word-level) approach as in [2] based on the following observations:

- Sentences are the semantic units in natural language text.
- Differentiation can be made at the sentence level. For example, we can give more weights to the 'thematic sentences' when calculating the whole text's overall polarity.

Next we present our two-step Chinese opinion mining procedure based on Chinese sentence segmentation.

2.1.2 Computing the Polarity Score of Single Sentences

First, we determine the polarity score of subjective sentence by using HowNet's sentiment words. As most of interrogative sentences have no semantic orientation, we disregard them. For each single sentence, a quadruple <S, P, D, N> is to be extracted, whose components are given as follows:

- S: opinion **S**ubject or object, noun or pronoun
- P: semantic **P**olarity, $+1$ or -1, denoting sentiment word's polarity in HowNet
- D: **D**egree modifier, weight value as given in HowNet
- N: **N**egation modifier, -1 or $+1$, denoting whether negation modifier is present or not

A complicated algorithm is used to transform each sentence into its corresponding quadruple (details omitted due to page limit). For example, below are the corresponding quadruples for the three Chinese sentences above:

- 我很高兴 <我, +1, 1.5, +1 >
- 我不很高兴 <我, +1, 1.5, +1 >
- 我很不高兴 <我, +1, 1.5, −1 >

The computation of the polarity score of a single sentence is straightforward. Assuming that the negation modifier does not appear before the degree modifier, the score is defined as $P * D * N$. When the negation modifer does appear, the score is set to $\alpha * P$, where α denotes a 'balancing coefficient' whose value we choose 0.5 empirically. Note that fine-grained differentiation of sentiment words apart from their polarity can lead to further improvement of our algorithm. We are exploring this direction in our current work. The basic approach as presented in this paper delivers satisfying results, as demonstrated in Section 3.

2.1.3 Synthesizing the Polarity of Entire Text

After the polarity score is calculated for single sentences, we synthesize the overall polarity score of the entire review. We note that different types of sentences have different contributions to the overall polarity. Especially, we take two special types of sentences into account:

1)first-person sentences: those sentences containing such words as '我(I)/作者(author)/笔者(writer)'. 2)topical sentences: those containing keywords with respect to the topic being discussed and those occur in the title. We give more weights to these sentences when synthesizing the polarity scores. Our experimental results shown later validate our approach.

2.2 Machine Learning Approach

Considering the great success of topic-based classification with machine learning-based approaches, we examine the effectiveness of applying machine learning techniques to the sentiment classification problem. It turns out that this problem is quite different from traditional topic-based classification. While topics are often identifiable by keywords alone, sentiment can be expressed in a more subtle manner. For example, the sentence '与其痛苦地活着还不如安乐的死去' contains no single word that is obviously negative. Thus, classifying sentiments requires deeper level of understanding than the usual topic-based classification. Early attempts to classifying movie reviews used standard bag-of-words techniques with only limited success[1]. We have conducted a series of polarity classification task based on a variety of machine learning algorithms including SVM, Naive Bayes, and Decision Tree. Specifically, we have experimented with the following three kinds of feature sets.

- bag-of-words: an unordered collection of words which represent a text.
- appraisal phrases: phrases indicating the feelings toward things or objects. e.g. 很好(very good)/很高兴(very happy) etc.
- parts-of-speech the eight traditional grammar classified words.

3 Experimental Results

Some public health events invoke dramatically conflicting sentiments on people. For example, people's views on 'euthanasia' could be vastly different. They either accept it passionately or reject it completely, but seldom stand on the neutral ground. As part of our experiments, we have extracted 851 Chinese reviews on 'euthanasia' from various Web pages, Blog postings, and online forums, and manually labeled them into 502 positive reviews and 348 negative ones. The standard precision, recall and F1 value are chosen as the evaluation metrics. Table 1 are the evaluation results of the approach based on Chinese sentiment words.

These results indicate that differentiating the three kinds of sentences can lead to great performance improvement.

Table 1. Performance of Chinese Sentiment Word Approach

Method	Polarity	Precision	Recall	F1	Accuracy
common sentences	positive	64.12%	88.65%	74.42%	60.75%
	negative	66.06%	20.63%	31.44%	
common sentence + first-person sentences	positive	66.87%	88.85%	76.31%	64.98%
	negative	70.86%	30.66%	42.80%	
common sentence + topical sentences	positive	77.35%	88.45%	82.53%	76.15%
	negative	80.32%	58.45%	67.66%	
common + first-person + topical sentences	positive	76.44%	89.84%	82.60%	76.26%
	negative	81.82%	56.73%	67.01%	

Table 2. Performance of Machine Learning Approach

Feature Set	Polarity	Precision	Recall	F1	Accuracy
bag-of-words	positive	81.19%	89.44%	85.12%	81.52%
	negative	82.15%	70.11%	75.66%	
bag-of-words + parts-of-speech	positive	80.65%	88.84%	84.55%	80.82%
	negative	81.14%	69.25%	74.73%	
appraisal phrases	positive	78.44%	84.06%	81.15%	76.91%
	negative	74.28%	66.57%	70.21%	
bag-of-words + appraisal phrases	positive	81.57%	89.04%	85.14%	81.65%
	negative	81.79%	70.98%	76.00%	

The evaluation results of machine learning-based methods are listed below, Our experiment is based on ten-fold cross validation, and information gain (IG) feature selection. Table 2 are the performance results using support vector machine (SVM).

The results show that, to achieve the best performance we should use the bag of words and appraisal phrases feature sets, though traditional bag-of-words feature sets alone are sufficiently good.

Next, we compare the performances of three different machine learning algorithms on the Chinese euthanasia data set. The results using Support Vector Machine, Naive Bayes, and Decision Tree are shown in Fig.1.

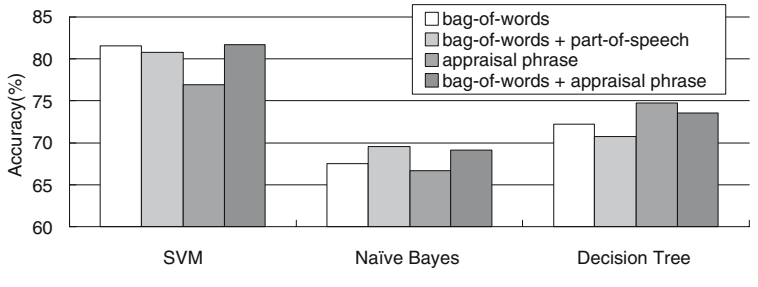


Fig. 1. Accuracies of Various Machine Learning Algorithms

4 Conclusion and Future Work

In this paper, we present our work on Chinese opinion mining, with emphasis on mining conflicting opinions on public health issues. We have developed two approaches: one based on HowNet and the other based on machine learning methods. Using a real-world dataset on opinions about euthanasia, we have conducted comparative experimental studies and conclude that both approaches seem to be effective. Though the machine learning based method outperforms its alternatives (especially the method based on SVM), such methods would require large labeled training instances, which are usually time consuming and labor intensive to acquire.

One promising line of future research is to improve the performance of non machine learning-based methods. Another direction is to use semi-supervised learning methods with a small set of manually labeled train data plus many additional unlabeled data. Considering the complexity of Chinese opinion mining, our current work draw on more heavily Chinese NLP techniques, such as Chinese parsing and semantic annotation.

Acknowledgements

This work is supported in part by NNSFC #60621001 and #60573078, MOST #2006AA010106, #2006CB705500 and #2004CB318103, CAS #2F05N01 and #2F07C01, and NSF #IIS-0527563 and #IIS-0428241.

References

1. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? sentiment classification using machine learning techniques. In: Proceedings of EMNLP-02, the Conference on Empirical Methods in Natural Language Processing, pp. 79–86. Association for Computational Linguistics, Philadelphia, US (2002)
2. Turney, P.: Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: Proceedings of ACL 2002, 40th Annual Meeting of the Association for Computational Linguistics, pp. 417–424. Association for Computational Linguistics, Philadelphia, US (2002)
3. Dave, K., Lawrence, S., Pennock, D.M.: Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In: Proceedings of WWW 2003, 12th International Conference on the World Wide Web, Budapest, HU, pp. 519–528. ACM Press, New York (2003)
4. Kobayashi, N., Iida, R., Inui, K., Matsumoto, Y.: Opinion mining on the web by extracting subject-aspect-evaluation relasions. In: Proceedings of AAAI-CAAW 2006, the Spring Symposia on Computational Approaches to Analyzing Weblogs, Stanford, US (2006)
5. Wang, B., Wang, H.: Bootstrapping both product properties and opinion words from chinese reviews with cross-training. In: Web Intelligence, pp. 259–262 (2007)

Parallel Crawling and Capturing for On-Line Auction

Cheng-Hsien Yu[1,2] and Shi-Jen Lin[1]

[1] National Central University，320 Tao-yuan, Taiwan
[2] China University of Technology，303 Hsin-chu, Taiwan
chyu@cute.edu.tw, sjlin@mgt.ncu.edu.tw

Abstract. The on-line auction is one of the most successful types of electronic marketplaces and has been the subject of many academic studies. In recent years, empirical research on on-line auction has been flourishing because of the availability of large amounts of high-quality bid data from on-line auction sites. Researcher can provide buyers and sellers with useful information, and help prevent fraud and reduce purchase cost. However, the increasingly large volumes of bid data have made data collection increasingly complex and time consuming, and there are no effective resources that can adequately support this work. So this study focuses on the parallel crawling and capturing of on-line auctions from the automatic agent perspective to help researchers collect auction data more effectively. The issues in this study include parallel crawling architecture, crawling strategies, content capturing strategies, and prototype system implementation. Finally we conduct an empirical experiment on eBay U.S. and Ruten Taiwan to evaluate the performance of our parallel crawling and capturing. The results of this study show that our parallel crawling and capturing methodology is able to work in the real world and generate good performance outcomes.

Keywords: parallel crawling, on-line auction, agent.

1 Introduction

The on-line auction is one of the most successful types of electronic marketplaces (Bajari, 2003). Over 10 million items can be found daily for sale on on-line auction sites, like eBay, Amazon, and Yahoo. The success of on-line auctions has given buyers access to greater product diversity with lower prices. It has provided sellers with large numbers of potential buyers (Lucking-Reiley, 2000). However, this success comes at a price, as buyers must pay higher search costs to locate desired products and sellers have to face greater competition from many other sellers (Bichler, 2002).

In order to avoid auction fraud, reduce the buyers' costs and increase sellers' competence, increasing numbers of researchers have begun to study issues surrounding on-line auctions. In recent years, empirical research on on-line auction has been flourishing because of the availability of large amounts of high-quality bid data from on-line auction sites. Researchers want to study price, bid behavior and auction market characteristics to locate suggestions for buyers and sellers participating in on-line auctions. However, the increasingly large volumes of bid data have made data collection increasingly complex and time consuming, and there are no

Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 455 – 466, 2008.

effective resources that can adequately support this work. So this study focuses on the parallel crawling and capturing of on-line auctions from the automatic agent perspective to help researchers collect auction data more effectively. For these reasons, our research will focus on parallel crawling architecture, crawling strategies, content capturing strategies, and prototype system implementation. Finally we conduct an empirical experiment on eBay U.S. (www.ebay.com) and Ruten Taiwan (www.ruten.com.tw) to evaluate the performance of our parallel crawling methodology. We hope the results will help researchers to study on-line auctions more effectively.

2 Literature Review

Due to the rapid development of on-line auctions, and the large number and special structure of auctions, data collection and analysis of auction research is increasingly difficult. With this in mind, the literature review surveys issues related to on-line auctions and review related literature regarding current on-line data services and on-line auction data structure.

2.1 Issues Related to On-Line Auction

There are few analytic data services provided by on-line auction sites, so researchers also have to collect auction data for their studies by themselves. Lucking-Reiley (2000) and Bajari (2003) studied the price determination by bid history of eBay.com. Bapna et al. (2004) used bid histories to study the bid strategies of uBid.com. Roth (2002) compared last bidding behaviors between eBay.com and amazon.com. Dellarocas (2001) analyzed the trust in the reputation mechanism of buyers and sellers of eBay.com. Bajar (2003) and Borle et al. (2005) compared the bid strategies between general value products and private value products on eBay.com. Jank et al. (2005) studied price dynamics and evolution of eBay.com. Kauffman et al. (2000) studied auction fraud detection of eBay.com. The research literature we reviewed was all created using empirical methodologies, showing the importance of the real auction data collection. On the other hand, only Kauffman (2000) and Bapna (2004) used agent technology to collect auction data automatically, the other researches collected data manually. So researchers also studied for short term without the support of automatic data collection tools. With regard to empirical methodologies, the researchers also used real auction data to verify their models, so an automatic data collection tool for on-line auction would provide strong benefits for researches.

2.2 Current On-Line Data Services

The current on-line data services for on-line auctions also provide suggestions for sellers and buyers based on bid histories. Hammertap.com (figure 1) provides a hot merchandise list, a most profitable category list, and a most successful sellers list for eBay.com. whizanalysis.com (figure 2) provides sales analysis reports for earnings of different merchandise and sellers. The online data services typically archive the

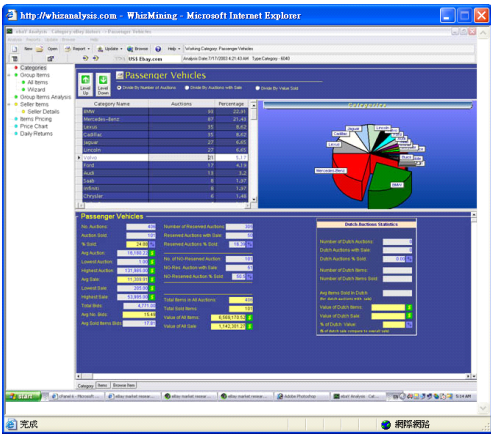


Fig. 1. Whizanalysis.cmo

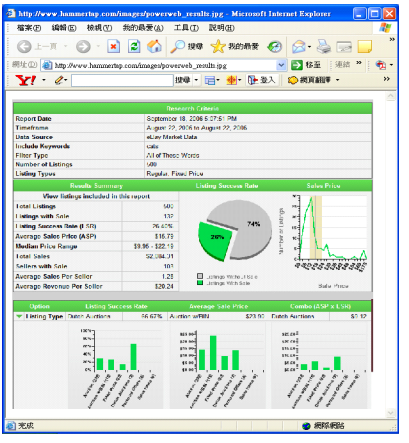


Fig. 2. Hammertap.com

selected merchandise or categories over a long period and report statistics, demographics, and pivot tables of bid histories, as well as providing suggestions by auction experts. The reports include only overall information, and do not provide details of specific auctions. They also archive selected merchandise and categories to reduce data size, so researches are not able to use this kind of data for empirical research.

2.3 Data Structure of On-Line Auction

On-line auctions have different data structure to on-line shopping. Users can be sellers and buyers. Sellers can post auctions and set the category, title, start time, end time,

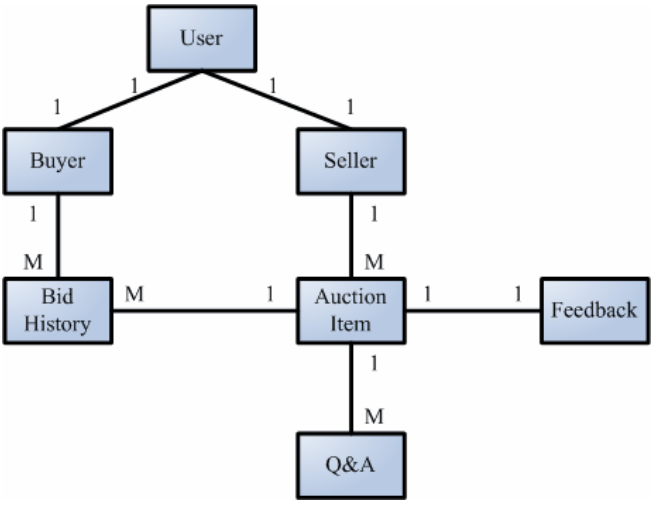


Fig. 3. Data structure of on-line auction

minimum bid, reserve price, area, condition, and description. Buyers can search auctions, and place bids, or make a buy-it-now transaction. The auction system records every bid, and buyers and sellers can give feedback to each other after the auction is completed. The auction system also provides a Q&A interface for communication between buyers and sellers. We can illustrate the data structure of online auctions as shown in figure 3.

Information for unclosed auctions is uncertain for the purpose of data analysis, as the state of auction is constantly changed by bid dynamics. We can only analyze an auction after it is closed. Unfortunately, the categories of auction sites are designed for sales, so information on closed auctions can not be found in any auction category. We can only locate closed auctions by item number. Researchers have to record all the item numbers of auctions they want to study before the auction closes. Since closed information can be used to reference reputation data, auction sites store the closed auction data for a period of time to help users to evaluate reputation scores. But due to limitations in storage capacity, auction sites delete the historical data after a given period of time. Currently, researchers face the problems of rapid increasing data and short data reserve time, making data collection increasingly difficult and time-consuming. As a result, there is a trend towards developing automatic data collection tools.

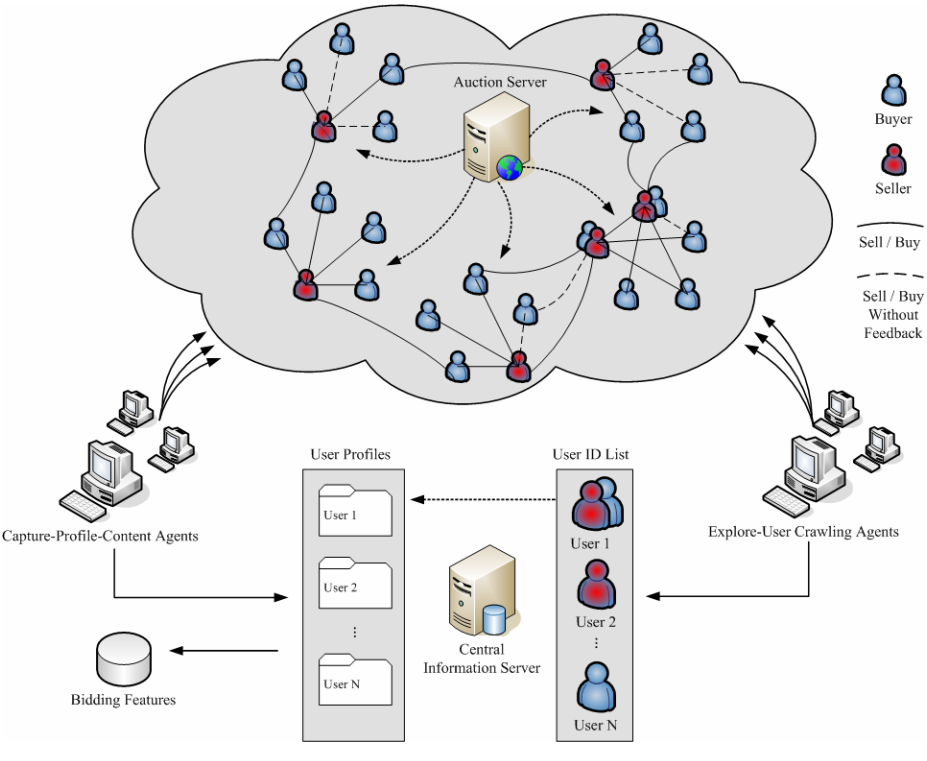


Fig. 4. Parallel Crawling Architecture

3 Parallel Crawling Architecture

Large amount of bid data are generated daily. Auction sites delete over-valid-time historical data to reduce data size, which means that researchers can not find this data on auction sites. Automatic tools, like the Net-Spider system, can crawl and capture web data faster than manual collection, and is therefore used to increase efficiency. Different of agent systems use different crawling and capturing methodologies. First, we will discuss web crawling first, content capturing will be discussed in the next section. In our research, the multi-agent system is used for parallel crawling and two types of agents are designed for crawling and capturing, as shown in figure 4. The explorer-user crawling agent crawls on the auction site to find users and maintains a user ID list to avoid duplicate crawls. The capture-profile-content agent checks the user ID list and selects the unvisited users to crawl, and captures the content of auction pages to create a profile of the user. Then the bidding features can extract from the user profiles. The crawling methodologies can be classified in four types as below:

3.1 Crawling by Monitored Items

As mentioned above, closed auctions can not be found in any sales category on auction sites. We can only locate the closed auction through the recorded item number. Researchers can monitor auction categories, check bid dynamics and record item numbers. When auctions are closed, researchers can locate them by the recorded item numbers, as shown in figure 5. On the other hand, data is lost when researchers do not record relevant item numbers.

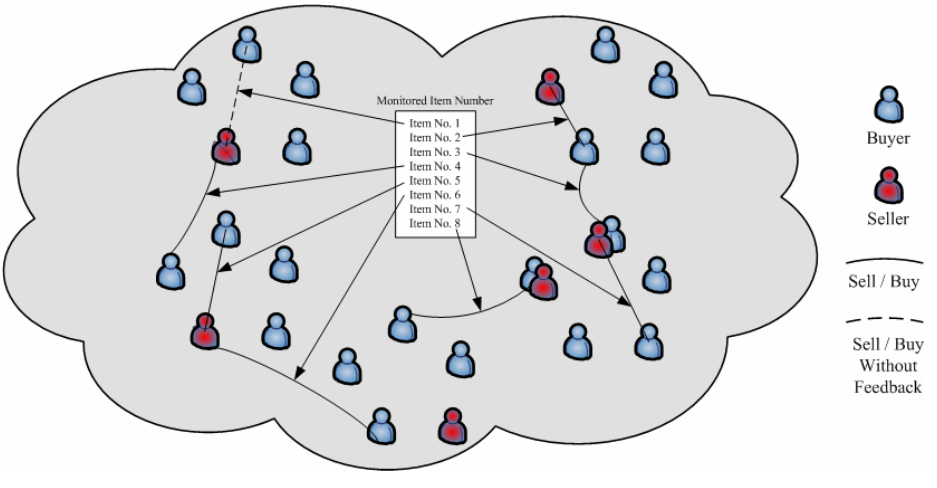


Fig. 5. Crawling by Monitored Items

3.2 Crawling by Random Item Number

Generally, auction sites use the item number to locate the auction, so researchers can change the parameter to search randomly for auction numbers, as shown in figure 6.

This kind of search covers a wide area of data, but can not be executed efficiently as unsuitable information will also be included in the search results, like invalid, deleted and unclosed auctions. In addition, this type of search is not compatible with all auction sites to enable search and capture for specific data. The different auction sites use different item number encoding, and researchers have to decode this information to increase the hit rate of searches.

3.3 Crawling by Reputation Relationship

Besides searching by random item number, we can also search by reputation relationship. Generally, sellers and buyers will release feedback to each other after an

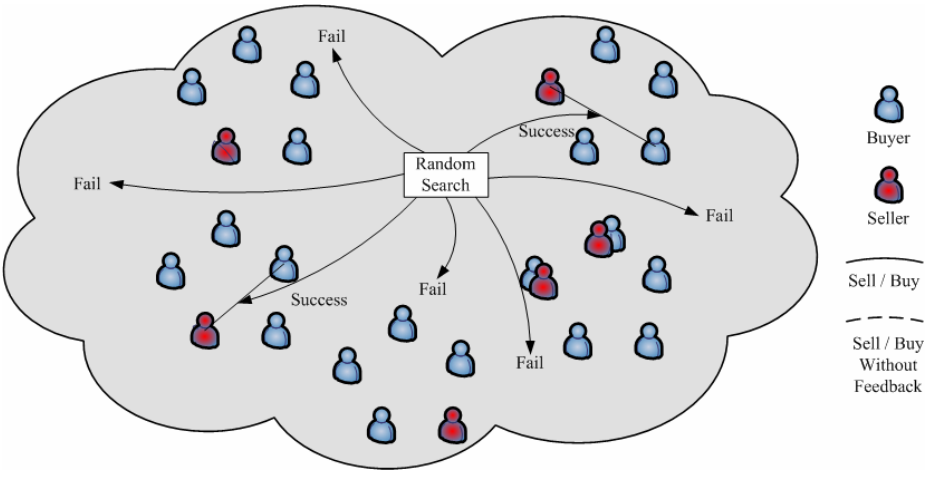


Fig. 6. Crawling by Random Item Number

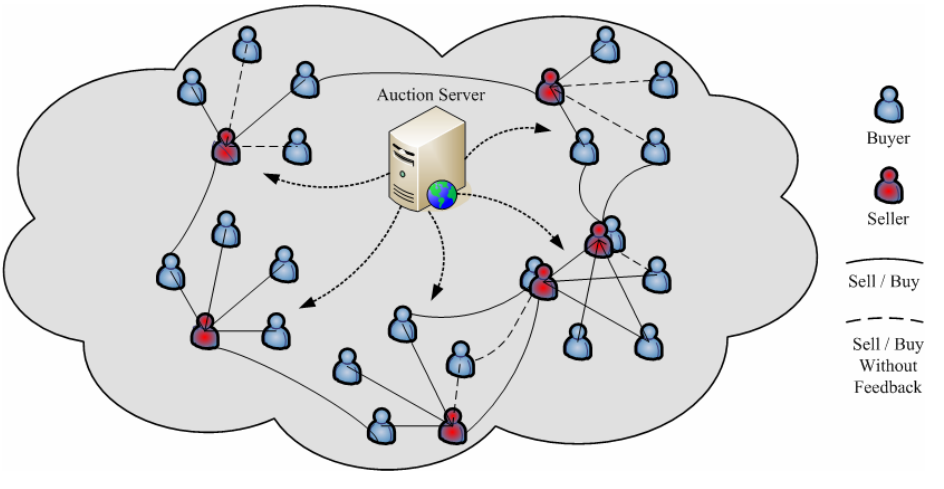


Fig. 7. Crawling by Reputation Relationship

auction is closed. Each reputation score of a buyer or seller is accumulated through this feedback. To show a clear bid history of reputation scores, auction sites will preserve the related history auction pages and provide users with links to them, so enabling them to evaluate the reputation of buyers and sellers. For this reason, we can crawl on auction sites by reputation data, as in figure 7. First, the crawling can start from the high reputation score of a buyer or seller, and crawl by his reputation relationship. Then, the crawling can be expanded level by level until we harvest enough data.

3.4 Crawling by Reputation Relationship with Mutation

On the other hand, users can decline to release feedback to the other side, making it impossible to find the auction page. This situation presents a problem in crawling by reputation relationship. The mutation strategy is designed to address this problem and improve the effectiveness of crawling. We expect the mutation strategy will change a small part of the item number and reach the auction pages with no feedback data, as in figure 8. Our study found that the item number encoding of auction sites is often over 10 bits. So if an all-bits mutation strategy is used, it will expand the search space more than 10^{10} complexity and slow down crawling. Therefore, we have to limit mutation bits to increase the crawling speed. In addition, the mutation strategy will not break the item number encoding structure and generate a new similar item number to cover the area over the original result. As such, this mutation strategy can overcome the limitations of crawling by reputation relationship.

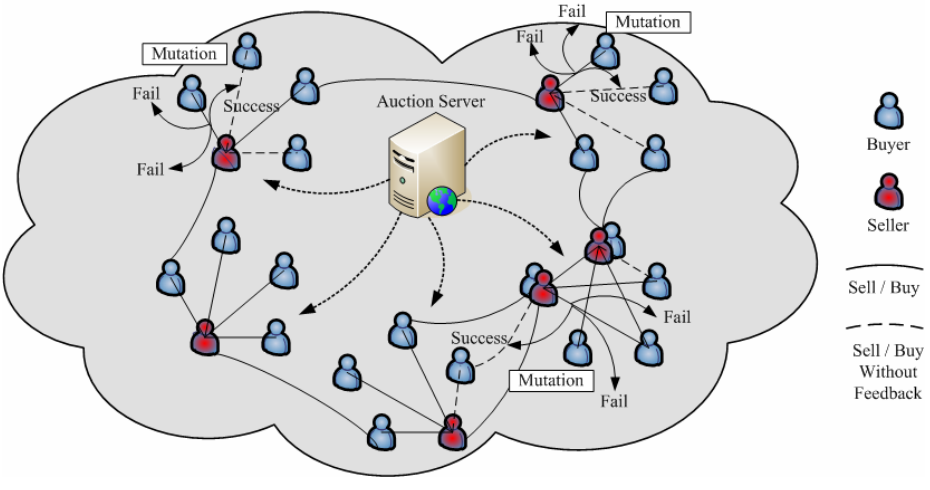


Fig. 8. Crawling by Reputation Relationship with Mutation

4 Content Capture Strategy

The current content capture strategy on the web can be classified in two types as below:

4.1 Capture Content by Semantic Parsing

Semantic parsing will analyze the structure of the web document to find the valid syntax description. This method can be used to check the syntax error and capture specific content in the document. If we want to use semantic parsing to capture the specific content of auction pages, we have to obtain the web page structure for parsing. In our study, the document standard of auction sites, HTML, is a static structure web page document. It can descript the typesetting of web pages but can not descript the data structure. Researchers can not differentiate between data fields and their values. For example in figure 9, "Item number: 12345678901234" is a part of the auction page, "Item number" is a data field and "12345678901234" is its value. It is hard to differentiate in HTML standard. XML is an extended version of HTML, and can descript the data structure. Unfortunately, XML is not the current standard on auction sites, so we are still unable to use semantic parsing to capture specific content on auction sites.

4.2 Capture Content Based on Ordering of Typesetting

Currently, we can not capture the content of auction pages based on semantic parsing, so some researchers have found other methods, such as capturing content based on the ordering of typesetting. The idea is to locate the specify content by tag ordering. With regard to this idea, we can define a descript syntax to locate the specific content in auction pages as definition (1).

(1)The text between the N^{th} tag and the $(N+1)^{th}$ tag, N is the tag ordering of the document.

For example in figure 9, the value of item number is defined as "The text between the 8^{th} tag and the 9^{th} tag of the document". It seems to be rational to capture content based on the ordering of typesetting. But our study found that this method will frequently capture incorrect data because of the dynamics of client side scripts and interactive advertisements especially in the "<Script></Script>" section. The active script code will disturb the tag ordering and make the ordering unstable. To solve this problem, we propose a revised idea which combines the feature-anchor to improve the defects in tag ordering. The revised syntax is defined as definition (2) and (3).

```
<html><head><title>Online Auction</title></head>
<body>
<b>Item number:</b>12345678901234<br>
<b>Title:</b>Nintendo Wii <br>
……
</body>
</html>
```

Fig. 9. HTML document example

(2)The text between the N^{th} tag and the $(N+1)^{th}$ tag after the feature-anchor, N is the tag ordering after the feature-anchor of the document.

(3)The text between the N^{th} tag and the $(N-1)^{th}$ tag before the feature-anchor, N is the tag ordering before the feature-anchor of the document.

For example in figure 9, the value of item number is defined as "The text between 1st tag and 2nd tag after feature-anchor ('Item number')" or "The text between 2nd tag and 3rd tag before feature-anchor ('Title')".

5 Result and Discussion

This research constructed a prototype system with crawling by reputation relationship and content capturing based on ordering of typesetting. We also conducted an empirical experiment on eBay U.S. and Ruten Taiwan for verify our parallel crawling architecture and evaluate system performance. The system ran with two auction sites concurrently for 8 hours on the same machine. The result is described below in two sections. In addition, we will evaluate the performance of the mutation strategy and discuss in a later section.

5.1 Result on eBay U.S.

The system ran on eBay U.S. for 8 hours to capture 7,682 auction pages and 9,773 user IDs, as in figure 10 and 11. On eBay U.S., bid history data is not provided after a 90 days limit. EBay U.S. has been operating for many years, and many auctions have taken place. However the problem of deleted data still exists due to this time limitation. This situation caused some faults when searching from reputation relationship. As expected from out discussion above, the number of user IDs is more than the number of auction pages.

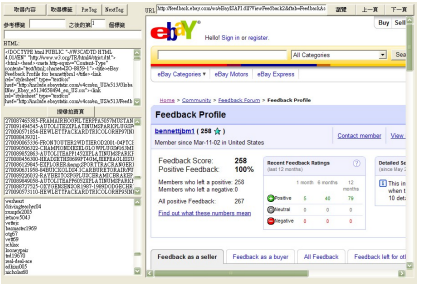


Fig. 10. System result on eBay U.S.

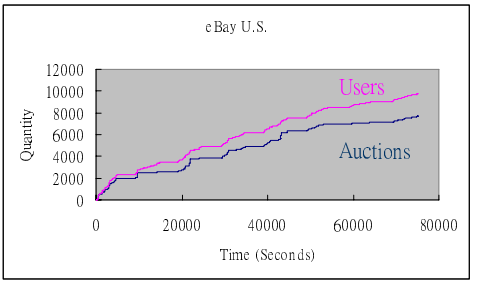


Fig. 11. Performance evaluation of eBay U.S.

5.2 Result on Ruten Taiwan

The system ran on Ruten Taiwan for 8 hours to capture 117,163 auction pages and 51,266 user IDs, as in figure 12 and 13. The result on Ruten Taiwan are significantly

different to the result on eBay U.S., however, since the number of auction pages is larger than the number of user IDs. We believe the reason is as follows. PC home and eBay Taiwan were combined to one auction site "Ruten" for just a short time, for the most part of auctions were valid and did not run over the preserve time. The crawling was more successful than the crawling on eBay U.S. because of the high recall rate of crawling. The advantage of domestic area networking makes the amount of data found on Ruten Taiwan's fifteen times larger than the amount of eBay U.S.'s. Regarding the results of our study, we found our crawling strategy will be efficient and stable over time. So our proposed strategies are very suitable to use for in automatic auction data collection.

Fig. 12. System result on Ruten Taiwan

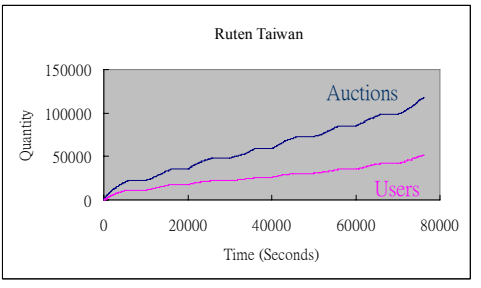


Fig. 13. Performance evaluation of Ruten Taiwan

5.3 Evaluation of Mutation Strategy

To evaluate the performance of mutation strategy, we use several mutation situations in the evaluation. We mutate the bits from the last-1-bit to the last-7-bits, and all-bits of item number. Every type of mutation is performed 1000 times and the results compared in table 2 and table 3. The field "closed auctions" refers to the number of found items which were closed. The field "on bidding auctions" refers to the number of found items which were not closed. The field "duplicate crawling" refers to the number of found items which were found in early crawling.

An analysis of the numbers in table 1 shows that the mutation performs well on eBay U.S. When the mutation-bit length increases, the crawling performance increases too. The mutation-bit length over 3 has significant performance improvement, while the mutation-bit length 5 has the best performance. When mutation-bit length is over 6, performance worsened. The all-bits-mutation (random search) has a very bad performance on eBay U.S. and we do not recommend researchers use this strategy.

On the other hand, the mutation strategy did not performed well on Ruten Taiwan, as shown in table 2. The best performance is only one-thirtieth of the performance of eBay U.S. The result shows the mutation strategy is ineffective for Ruten Taiwan. To find out the reason for this, we compare the environment of the two auction sites and find the difference as listed below:

1. The number of auction items on eBay U.S. is much larger than the number of auction items on Ruten Taiwan

Table 1. Evaluation of mutation strategy on eBay U.S.

Result / Mutation	Closed auctions	On-bidding auctions	Duplicate crawling	Test times
Last-1-bit	140	0	101	1000
Last-2-bits	239	0	28	1000
Last-3-bits	713	0	3	1000
Last-4-bits	711	0	0	1000
Last-5-bits	735	0	0	1000
Last-6-bits	634	0	0	1000
Last-7-bits	593	0	0	1000
All-bits	0	0	0	1000

Table 2. Evaluation of mutation strategy on Ruten Taiwan

Result / Mutation	Closed auctions	On-bidding auctions	Duplicate crawling	Test times
Last-1-bit	5	13	85	1000
Last-2-bits	16	52	10	1000
Last-3-bits	5	69	7	1000
Last-4-bits	14	56	4	1000
Last-5-bits	24	47	1	1000
Last-6-bits	8	24	0	1000
Last-7-bits	0	0	0	1000
All-bits	0	0	0	1000

2. EBay U.S. uses the 12 bits item number coding while Ruten Taiwan uses the 14 bits item number coding.

According to the comparisons above, we can find many unused item numbers on Ruten Taiwan. These unused item numbers make the mutation-crawling fail and decrease the performance of mutation strategy on Ruten Taiwan.

6 Conclusion and Future Directions

This research started from automatic agent perspective and studied issues including parallel crawling architecture, crawling strategies, content capturing strategies, and prototype system implementation. We have proposed ideas for improvement and constructed a prototype system to verify our revised strategy in the real world environment. Finally we evaluated the performance of proposed crawling architecture, strategies and prototype systems, and provided conclusions and recommendations for researchers doing related research. The conclusions are listed as below:

1. The prototype system was evaluated in the real world environment and it is useable and performed well on both eBay U.S. and Ruten Taiwan.

2. The proposed crawling strategy and content capture strategy can be used to crawl and capture the large amounts of on-line auction content.
3. The mutation strategy can break through the data closure of crawling by reputation relationship. It performs differently when used on different auction sites. How to implement the mutation strategy effectively will be the focus of our future work.
4. This parallel crawling architecture has a high performance to help researchers collect large amounts and rapidly increasing on-line auction data. Using this kind of tool can expand the scale of research in the long term.

Future work includes application model construction and verification. Performance comparison between crawling strategies is also an important direction.

References

1. Bajari, P., Hortacsu, A.: The Winner's Curse, Reserve Prices and Endogenous Entry: Empirical Insight from eBay Auctions. Rand Journal of Economics 3(2), 329–355 (2003)
2. Bapna, R., Goes, P., Gupta, A., Jin, Y.: User Heterogeneity and its Impact on Electronic Auction Market Design: An Empirical Exploration. MIS Quarterly 28(1), 21–43 (2004)
3. Bichler, M., Kalagnanam, J., Katircioglu, K., King, A.J., Lawrence, R.D., Lee, H.S., Lin, G.Y., Lu, Y.: Applications of Flexible Pricing in Business-to-Business Electronic Commerce. IBM Systems Journal 41(2), 287–302 (2002)
4. Borle, S., Boatwright, P., Kadane, J.B.: The Timing of Bid Placement and Extent of Multiple Bidding: An Empirical Investigation Using eBay Online Auctions, Working paper, Rice University (2005)
5. Dellarocas, C.N.: Analyzing the Economic Efficiency of eBay-like Online Reputation Reporting Mechanisms. In: Proceedings of the 3rd ACM Conference on Electronic Commerce, Tampa, FL, October 14-16 (2001)
6. Jank, W., Shmueli, G.: Profiling Price Dynamics in On-line Auctions Using Curve Clustering, Working Paper, Smith School of Business, University of Maryland (2005)
7. Kauffman, R.J., Wood, C.A.: Running up the Bid, Modeling Seller Opportunism in Internet Auctions. In: Proceedings of the AMCIS 2000 Conference, Long Beach CA (2000)
8. Lucking-Reiley, D.: Auctions on the Internet: What's Being Auctioned and How. Journal of Industrial Economics 48(3), 227–252 (2000)
9. Roth, A.E., Ockenfels, A.: Last-Minute Bidding and the Rules for Ending Second-Price Auctions: Evidence from eBay and Amazon Auctions on the Internet. The American Economic Review 92(4), 1093–1103 (2002)
10. Shmueli, G., Jank, W., Aris, A., Plaisant, C., Shneiderman, B.: Exploring Auction Databases through Interactive Visualization. Decision Support Systems (2006)

A New Credit Card Payment Scheme Using Mobile Phones Based on Visual Cryptography

Chao-Wen Chan and Chih-Hao Lin

Graduate School of Computer Science and Information Technology,
National Taichung Institute of Technology,
404 Taichung, Taiwan
ccwen@ntit.edu.tw, s18953202@ntit.edu.tw

Abstract. With the increasing usage of credit cards, secure requirements for handling credit card payment have become more critical. Today, it is being deployed for a wide application services in mobile commerce over public networks. Based on the security of Naor-Shamir's visual cryptography, Diffie-Hellman key agreement scheme and symmetric cryptosystems, we propose a new credit card payment scheme using mobile phones based on visual cryptography. Compared with traditional credit card payment methods', the proposed scheme can provide more secure communication for many electronic commerce transactions.

Keywords: payment, secret sharing, visual cryptography, key agreement.

1 Introduction

In the last decade, many advances have been made in the technology of mobile commerce, such as hardware of mobile devices, qualities of application services, etc. With the increasing usage of mobile commerce applications over public networks, it is more and more important to have reliable security. In recent years, identity theft has surged and led to serious damage for both cardholders and payment companies. Traditional credit card payment schemes in Taiwan require cardholders show credit card information to merchants, allowing strangers access to private business operations. The issuer only provides verification for a payment request from a credible requester, for instance a phone call or an SMS (Short Message Service) is needed to check the cardholder's credit limit and currently charged amount. This method does not ensure a secure credit card payment. The verification process is only performed according to the parameters of a cardholder's available credit. In the past, the above security problem has attracted little attention from researchers [2]. Observe that leakage of private information stored in a credit card is the attraction. To prevent the leakage of cardholder's information, a new secure credit card payment scheme should satisfy the following requirements.

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 467–476, 2008.

1. The merchant learns nothing about the private information stored in a credit card during a credit card payment.
2. No one else can use the secret information about a credit card payment to perform any other credit card payments.
3. Both cardholder and issuer bank need to authenticate each other at beginning of a credit card payment.
4. The credit card payment scheme may support the electronic cash or coin-like application.

Owing to it is being deployed for a wide application services in mobile commerce over public networks. Based on Diffie-Hellman key agreement method [5], Naor-Shamir's Visual Cryptography [9], and symmetric cryptosystems, we propose a novel secure credit card payment scheme which could provide the above requirements. The proposed scheme is different from traditional credit card payment methods. We briefly describe the new idea of our method. In general, credit cards include private information about users, such as card numbers, validity dates, expiration dates and identifications etc. To prevent a card user from sharing the credit card with others, the user must get an authorization code from the bank before making purchases. The bank checks the quota of the cardholder and confirms the identification of the user from the database. If the information of the user is correct, the bank generates an authorization code and transfers it to the user over mobile communication networks. The user must input the authorization code to accomplish the payment without giving his credit card to the merchant. Compared with traditional credit card payment methods', the proposed scheme can provide more secure communication for many electronic commerce transactions.

The rest of the paper is organized as follows. In the next section, we present the necessary related works of our scheme. In Section 3, a new credit card payment scheme is proposed. Discussions are presented in Section 4. Finally, we give some conclusions in Section 5.

2 Preliminaries

Before a new credit card payment scheme is proposed, we first introduce the properties of Diffie-Hellman key agreement method [5], Naor-Shamir's Visual Cryptography [9], and symmetric cryptosystems that will allow us to discuss our scheme's security in Section 4.

2.1 Diffie-Hellman Key Agreement Scheme

The famous key agreement scheme was proposed by Diffie and Hellman in 1976. If Alice and Bob want to transfer a secret by Diffie-Hellman key agreement scheme, then Alice selects a random secret α and sends $g^{\alpha} \pmod{p}$ to Bob, where p is a large prime number. Then, Bob selects another random secret β and sends $g^{\beta} \pmod{p}$ to

Alice. Finally, Alice and Bob compute a common session key $K = g^{\alpha\beta} = g^{\beta\alpha} \pmod{p}$. Thus, Alice and Bob can use the common key K for encryption and decryption in the session. No one can derive the session key K from the public information $g^{\alpha} \pmod{p}$ and $g^{\beta} \pmod{p}$. The security is based on the computational Diffie-Hellman problem. In addition, neither of these two parties computes anything about the random secret value selected by another provided Discrete Logarithm Problem [1,8].

2.2 Visual Cryptography (VC)

In 1995, Naor and Shamir [9] proposed a secret sharing scheme in which secrets and shadows are represented in binary image format. This scheme opened the research on Visual Cryptography (VC); in fact it was the first paper about VC. Visual cryptography is a secret sharing method for digital images without cryptographic computations for decryption. In their scheme, a dealer wants to break a secret image into n shared images and sends n shared images to other participants, such that any t or more participants can recover the secret image by stacking t or more shared images. However, the secret image cannot be recovered by $l(< t)$ shadows. The above method is defined as (t, n)-threshold visual cryptography.

For example, a $(2, 2)$-threshold VC is depicted below. We set up three pairs of 2×2 pixel blocks as shown in Fig. 1. The pixel block consists of four sub-pixels. If a pixel of the secret image is black, we can randomly select a pair of pixel blocks, then, randomly assign one of the pair to the corresponding pixel block of share 1, and the other of the pair is assigned the corresponding pixel block of share 2. If the pixel of the secret image is white, we randomly select one of the three pairs, and randomly assign one of the pair to the corresponding pixel block of share 1. And, randomly select one pixel block of the other two pairs to the corresponding pixel block of share 2. As shown in Fig. 2, the pixel is white when stacked with one horizontal share and the other vertical share. In addition, the pixel is black when stacked with two horizontal shares. Now, taking a 100×100 digital image as another example, as shown in Fig. 3, we break a secret image into two shares, share 1 and share 2, respectively. Based on $(2, 2)$-threshold VC, we can recover the secret image by stacking share 1 and share 2. Still today, much research of visual cryptography for gray-level images is proposed in [3,6,7].

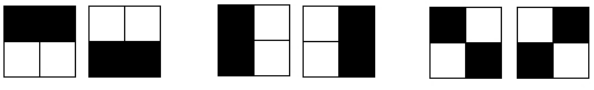


Fig. 1. Three Pairs of 2×2 Pixel Blocks

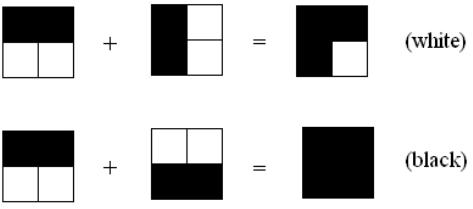
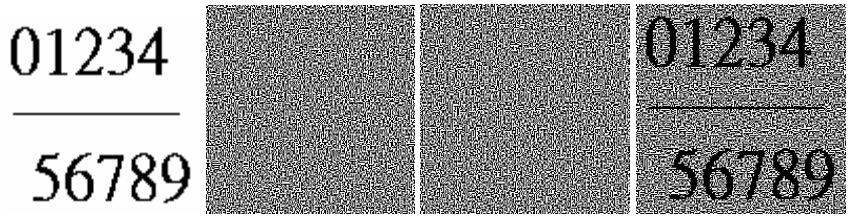


Fig. 2. Example 1 of $(2, 2)$-Threshold VC



(a)secret image (b)share 1 (c)share 2 (d)recovered image

Fig. 3. Example 2 of $(2, 2)$-Threshold VC

2.3 Symmetric Cryptosystems

Symmetric cryptosystems are used to encrypt or decrypt messages by the same secret key to provide authorized securities. It can provide more efficient than asymmetric cryptosystem for message encrypting and decrypting. One of the most popular symmetric cryptosystems is the Advanced Encryption Standard (AES) [10]. AES is an encryption standard provided by National Institute of Standards and Technology in 2001. However, the encryption method of images may be different from text. In 2001, Chang et al. [4] proposed an encryption algorithm for image cryptosystems based on vector quantization. It reduces the computational complexity of the encryption and decryption schemes.

3 The Proposed Scheme

In this section, we elaborate on our proposed credit card payment scheme. Our method is based on DH key agreement scheme and $(2, 2)$-threshold VC scheme to generate the authorization code $AuthC$. In order to confirm data integrity of the shared image, we use technologies of symmetric cryptosystems to transform the shared image. The proposed scheme consists of three parties: a user U, a bank B, and merchants M.

1. U: The uer who makes a purchase with his credit card.
2. B: The bank issues credit cards for the user. To easily present the method, we assume that the bank B is the issuer and acquirer, and the bank is a trusted third party.
3. M: Merchants which sell commodities to users.

Table 1. Notations

UID	A user U's identification number is defined as UID.
PW_U	A user U's password is defined as PW_U.
PO	A purchase order is a commercial document issued by a buyer to a seller, including information for products or services.
p	p is a large prime integer.
g	g is a public primitive integer in modular p, where g is not equal to 1.
$AuthC$	$AuthC$ are meaningful words or signs made by stacking two shared images for making a purchase, called authorization code.
$U \rightarrow B : \{m\}$	A party U sends a message m to a party B.
$E_k(m)$	A message m is encrypted by the symmetric key k.
$D_k(m)$	A message m is decrypted by the symmetric key k.
$PRNG$	A cryptographic pseudo random number generator is defined as $PRNG$.
$S_j = E_{S_i}^{VC(2,2)}(AuthC)$	The formula denotes using $(2,2)$ -threshold VC to generate the other shared image S_j by a shared image S_i and $AuthC$, where $i \neq j$.
$AuthC = D_{S_i}^{VC(2,2)}(S_j)$	The formula denotes using $(2,2)$-threshold VC to get $AuthC$ by stacking two shared images S_i and S_j according to human visual system, where $i \neq j$.

The basic notations are used in the description which is defined as Tab.1.

When U makes a purchase with his credit card, U must get the authorization code $AuthC$ for accomplishing a transaction. Hence, the proposed scheme consists of registration phase and transaction phase. The details of two phases will be described as follows:

3.1 Registration Phase

1. U sends his identification number UID and password PW_U to the bank B through a secure channel.
2. Upon receiving the messages UID and PW_U, B generates the shared image S_1 and stores the information $\{UID, S_1\}$ in the database. Next, B computes $S = E_{PW_U}(S_1)$ and stores S into the user's mobile phone through a secure channel.

3.2 Transaction Phase

1. User(mobile phone)→Bank: $\{PO, UID, a\}$
   U selects $\alpha \in_R \mathbb{Z}_p^*$ and computes $a = g^{\alpha} \pmod p$. Then, U sends the payment request PO, UID, and a to B.

2. Bank→ User(mobile phone): $\{b, R\}$

   Upon receiving the request messages PO, UID, and a, B can find out the user $U's$ information $\{UID, S_1\}$ according to UID in the database. Then, B selects $\beta \in_R \mathbb{Z}_p^*$ and computes $b = g^{\beta} \pmod p$ and $r = a^{\beta} = g^{\alpha\beta} \pmod p$. Next, B can use the integer r to get the initial coordinate (X, Y) for displacing vectors of the shared image S_1. Using r as a seed in a $PRNG$ function, we can get a set of random integers. Then, two parties negotiate for two integers to be the initial coordinate. We define this formula as $PRNG(r) = (X, Y)$, where (X, Y) is the initial coordinate vector (point). Thus, B displaces vectors of S_1 from $(0, 0)$ to (X, Y) to get the small shared image S_2, after getting the initial coordinate. The schematic diagram is shown as Fig. 4. Next, B generates a secret image for the user making a purchase, called authorization code $AuthC$. According to $(2, 2)$-threshold VC, another shared image S_3 can be computed by S_2 and $AuthC$ as shown below.

$$S_3 = E_{S_2}^{VC(2,2)}(AuthC). \tag{1}$$

   Finally, B computes $R = E_r(S_3)$, where r is used as a session key, and sends b and R to the user U.

3. User(mobile phone)→Merchant: $\{PO, UID, AuthC\}$

   When U receives the messages, U computes $r = b^{\alpha} = g^{\beta\alpha} \pmod p$ and $S_3 = D_r(R)$. Then, U inputs the random integer r into a $PRNG$ function to get the initial coordinate vector (X, Y). Next, U is required to input his password PW_U for decryption of the shared image S_1 by computing $D_{pw_U}(S) = S_1$. Therefore, U can use the initial coordinate (X, Y) to displace vectors of the shared image S_1 and get the smaller shared image S_2. Hence, U can recover the secret image by stacking the shared images S_2 and S_3 to get the authorization code $AuthC$. The recovered formula is defined as follows:

$$AuthC = D_{S_2}^{VC(2,2)}(S_3). \tag{2}$$

   In the above steps, U sends UID and $AuthC$ for payment without giving his credit card to the merchant M . Then, M forwards the messages $\{PO, UID, AuthC\}$ to B. B verifies the messages $\{PO, UID, AuthC\}$ in the database. If the records are valid, then B responses with a message of acceptance to M and updates the financial information in their system.

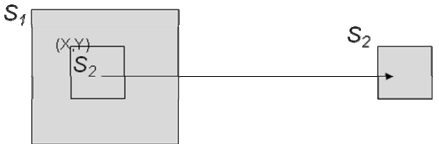


Fig. 4. Displace coordinates of S_1 to get the shared image S_2

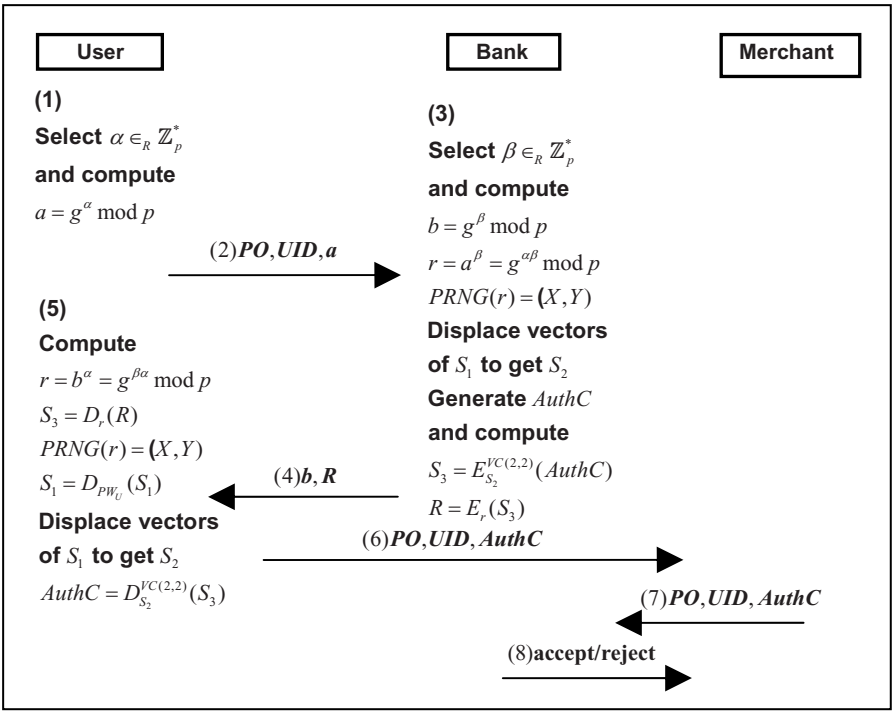


Fig. 5. The proposed scheme of the transaction procedure

Moreover, if U performs a new transaction in the next session, U and B can use the shared image S_1 to generate the other small shared image S_2 by implementing the DH key agreement scheme repeatedly. Then, U and B accomplish each payment according to our transaction phase. The above transaction procedure is briefly illustrated in Fig. 5.

4 Discussions

In this section, we are going to analyze the securities and performances of the proposed scheme.

4.1 Security Analysis

In the proposed scheme, the security is based on DH scheme, VC, and symmetric cryptosystems. Therefore, we consider some possible attacks. Suppose that an adversary attempts to impersonate the user U. Without knowing the random integer r generated by DH key agreement scheme, the adversary cannot get $AuthC$ to perform a transaction. The adversary only knows a and b on insecure public networks, and he learns nothing about the random integer r to get the initial coordinate (X, Y). The adversary must achieve the computational DH problem. However, an adversary may

try to compute α from $a = g^{\alpha} \pmod{p}$ or β from $b = g^{\beta} \pmod{p}$ to get the integer $r = a^{\beta} = g^{\alpha\beta} = g^{\beta\alpha} = b^{\alpha} \pmod{p}$. This means that the adversary still has to solve the discrete logarithm problem. The discrete logarithm problem is hard to solve in polynomial time. Hence, the proposed scheme can resist an illegal transaction by an impersonation attack. Even if an adversary knows the random integer r by man-in-the-middle attacks, he still cannot get $AuthC$ without the shared image S_1 which is shared by U and B. The shared image S_1 is protected by symmetric cryptosystems with the user's password PW_U. Therefore, our scheme can prevent an illegal transaction by man-in-the-middle attacks.

Next, we consider a replaying attack. A replaying attack is a method where an active adversary stores "old" intercepted messages and retransmits them at a later time. In our scheme, the authorization code $AuthC$ is generated by a trusted third party B. Suppose that an adversary will attempt to impersonate a legal user U by replaying old messages UID and a, he still cannot get $AuthC$ without knowing the shared image S_1. Hence, the proposed scheme can withstand an illegal transaction from replaying attacks. For each transaction, the bank records the information $\{PO, UID, AuthC\}$ which is used in the database. When a new transaction request is performed, the user will get a new $AuthC$ different from the above session. Therefore, the proposed scheme can prevent a double spending problem. In addition, both U and B need to authenticate each other by using $(2, 2)$-threshold visual cryptography. A valid $AuthC$ can only be recovered through the shared image S_1 which is shared by U and B. Hence, our scheme can provide the mutual authentication property.

Moreover, the proposed method is based on $(2, 2)$-threshold visual cryptography. The user (mobile phone) must stack two shared images S_2 and S_3 and use the human visual system to obtain the authorization code $AuthC$. According to decryption by human visual system, it is difficult to design a malicious program to get $AuthC$ for an adversary, such as brute force attacks. In the worst situation, suppose that a user's mobile phone is lost, the important information S_1 is still protected by a secure symmetric cryptosystem with the user's password PW_U. This reduces the threat of a lost or stolen mobile phone for the user before the emergency event happens.

4.2 Performance Analysis

To the best of our knowledge, the traditional (t, n)-threshold VC scheme can only be used once. It is similar to a one-time pad (OTP). When the secret image is recovered by stacking any $t'(\geq t)$ shares, the shared images cannot reuse the information in the next session. A dealer has to break a secret image into n shared images and retransfers n shared images to other participants. Therefore, with regard to efficiency, we generate a random seed r into a $PRNG$ by using DH scheme to reuse the shared image S_1 which is stored in the user's mobile phone. The user does not need to update the shared image S_1 in each session. Next, we consider the proposed scheme's computations to accomplish a transaction.

For convenience, we define related notations to analyze the computational complexity. The notation T_e means the time for one multiplication over a prime

integer p, T_{sym} denotes the time for one symmetric encryption or decryption, and T_{VC} denotes the time for stacking two shares by $(2, 2)$-threshold VC scheme. We summarize the computational complexity and communication in Tab. 2. As shown in Tab. 2, the proposed scheme only requires one symmetric encryption computation (T_{sym}) for the bank in the registration phase. In the transaction phase, our scheme performs two multiplication operations over a prime integer $p\,(2T_e)$, two symmetric decryption computation $(2T_{sym})$, and one stacking shares operation (T_{VC}) for a user in the transaction phase. For a bank, our scheme performs two multiplication operations over a prime integer p $(2T_e)$, one symmetric decryption computation (T_{sym}), and one stacking shares operation (T_{VC}). Therefore, in the fast progress of wireless communication, the proposed new credit card payment scheme using mobile phones based on visual cryptography would easily be implemented for electronic commerce transactions.

Table 2. Computational complexity and communication of the proposed scheme

	User	Bank
Number of communications	6	3
Number of rounds	2	2
Computation of the registration phase	No	T_{sym}
Computation of the transaction phase	$2T_e$+$2T_{sym}$+T_{VC}	$2T_e$+T_{sym}+T_{VC}

5 Conclusions

In this article, we propose a new credit card payment scheme by using DH scheme, VC, and symmetric cryptosystems for transferring secrets in insecure mobile communication networks. The security analyses show that the proposed scheme is more secure than traditional credit card payment methods. Our scheme can prevent impersonation attacks, man-in-the-middle attacks, replaying attacks, double spending problems, malicious program attacks and provides a mutual authentication property. In the future, our scheme has potential for application services in mobile commerce. We can perform similar application services, such as an electronic ticket system or an electronic auction market, etc.

References

1. Blake-Wilson, S., Menezes, A.: Authenticated Diffie-Hellman Key Agreement Schemes. In: Tavares, S., Meijer, H. (eds.) SAC 1998. LNCS, vol. 1556, pp. 339–361. Springer, Heidelberg (1999)
2. Bottoni, A., Dini, G.: Improving Authentication of Remote Card Transactions with Mobile Personal Trusted Devices. Computer Communications 30, 1697–1712 (2007)
3. Blundo, C., De Santis, A., Naor, M.: Visual Cryptography for Grey Level Images. Information Processing Letters 75(6), 255–259 (2000)
4. Chang, C.C., Hwang, M.S., Cheng, T.S.: A New Encryption Algorithm for Image Cryptosystems. The Journal of Systems and Software 58, 83–91 (2001)

5. Diffie, W., Hellman, M.E.: New Directions in Cryptography. IEEE Transactions on Information Theory 22(6), 644–654 (1976)
6. Lin, C.C., Tsai, W.H.: Visual Cryptography for Gray-level Images by Dithering Techniques. Pattern Recognition Letters 24(1-3), 349–358 (2003)
7. Lukac, R., Plataniotis, K.N.: Bit-level Based Secret Sharing for Image Encryption. Pattern Recognition 38(5), 767–772 (2005)
8. Maurer, U.: Towards the Equivalence of Breaking the Diffie-Hellman Scheme and Computing Discrete Logarithms. In: Desmedt, Y.G. (ed.) CRYPTO 1994. LNCS, vol. 839, pp. 271–281. Springer, Heidelberg (1994)
9. Naor, M., Shamir, A.: Visual Cryptography. In: De Santis, A. (ed.) EUROCRYPT 1994. LNCS, vol. 950, pp. 1–12. Springer, Heidelberg (1995)
10. National Institute of Standards and Technology: Advanced Encryption Standard, FIPS 197 (2001)

Detecting Hidden Hierarchy in Terrorist Networks: Some Case Studies

Nasrullah Memon[1,2] , Henrik Legind Larsen[1], David L. Hicks[1], and Nicholas Harkiolakis[2]

[1] The European Center for Counterterrorism Research and Studies
Department of Computer Science and Engineering
Aalborg University, DK-6700 Esbjerg, Denmark
[2] Hellenic American University, Athens Campus, Greece

Abstract. This paper provides a novel algorithm to automatically detect the hidden hierarchy in terrorist networks. The algorithm is based on centrality measures used in social network analysis literature. The advantage of such automatic methods is to detect key players in terrorist networks. We illustrate the algorithm over some case studies of terrorist events that have occurred in the past. The results show great promise in detecting high value individuals.

Keywords: Centrality, Destabilizing terrorist networks, Hidden hierarchy, Social network analysis.

1 Introduction

After the tragic terrorist attacks in New York and Washington in September, 2001, media interest in the Al Qaeda organization rose immediately. Experts and analysts all over the world started to offer various explanations of Al Qaeda's origins, membership recruitment, modes of operation, as well as possible ways for its disruption. One could thus read or hear that Al Qaeda is "a net that contains independent intelligence", that "functions as a swarm", that "gathers from nowhere and disappears after action", that is "an ad hoc network", "an atypical organization", and that is extremely hard to destroy, especially by traditional anti-terrorist or counterterrorist methods [1].

One common criticism of efforts for analyzing terrorism by focusing on tensions in defined hierarchies is to argue that the current terrorist threat is not organized with clear lines of authority. Instead they are organized as loose networks and so belong to an analytically distinct category. According to many counterterrorism analysts today, Al Qaeda has evolved from a centrally directed organization into a worldwide franchiser of terrorist attacks [2]. Since war in Afghanistan, which significantly degraded Osama bin Laden's command and control, Al Qaeda does appear to have become increasingly decentralized. It is now seen by many as more of a social movement than coherent organization [3].

Al Qaeda did not decide to decentralize until 2002, following the ouster of the Taliban from Afghanistan and the arrest of a number of key Al Qaeda leaders including Abu Zubaydhah, Al Qaeda's Dean of students, Ramzi bin Al Shibh, the organizer of the Hamburg cell of 9/11 hijackers, Khalid Sheikh Mohammed, the

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 477–489, 2008.

mastermind of 9/11 and the financier of the first World Trade Center attack, and Tawfiq Attash Kallad, the master mind of the USS Cole attack. In response to these and other key losses, Al Qaeda allegedly convened a strategic summit in northern Iran in November 2002, at which the group's consultative council decided that it could no longer operate as a hierarchy, but instead would have to decentralize [4].

Intuitively it is possible to argue that as all social structures experience inertia to changes, similar phenomena will be also observed in terrorist organizations. Based on the assumption that remnants of a hierarchical structure would also exist in the non-hierarchical networks it is imperative to develop and apply tools that can convert network structures to their underpinning/underlined hierarchies. Apart from studying terrorism these tools could be further used by law enforcement agencies to study the structure of existing terrorist networks and follow the best method of intervention in their effort to eliminate terrorism. Hierarchy, as one common feature of many real world networks, has attracted special attention in recent years [5] [6] [7] [8]. Hierarchy is one of the key aspects of a theoretical model to capture statistical characteristics of terrorist networks.

In the literature, several concepts are proposed to measure the hierarchy in a network, such as the hierarchical path [7], the scaling law for the clustering coefficients of nodes in a network [5], etc. These measures can tell us of the existence and extent of hierarchy in a network. We address herein the problem of how to construct hidden hierarchy in terrorist networks (which are known as horizontal networks). Discovering hierarchy in a terrorist network is *a process of comparing different centrality values of different nodes to identify which node is more powerful, influential or worthy to neutralize than others*.

From the above definition, it is clear that we may use different centrality measures for finding the corresponding hierarchical view of a graph/network. Every centrality measure is associated with a particular meaning, for example, degree centrality tells us about powerful nodes, and betweenness centrality gives us an idea about gatekeepers. Similarly these measures can be used to build hierarchies to detect an organizational view of a corresponding terrorist network/organization. Currently, experts have agreed that real destabilization is about isolating leaders from enough followers, thus disabling them from executing any terrorism plans. This idea has certainly been a central motive for investigative tools.

This paper is structured as follows: Section 2 discusses centrality measures we have applied in framing the algorithm for detecting the hidden hierarchy of terrorist networks. Section 3 introduces a new algorithm for detecting hidden hierarchy. We illustrate the algorithm using three case studies. Section 4 provides illustrations for the algorithm. Section 5 provides information on how the data on terrorist connections was collected. Section 6 concludes the paper.

2 Centrality Measures

Centrality is one of the network properties that frequently has been used to study actors or events in terrorist social networks. The general notion of centrality encompasses a number of different aspects of the "importance" or "visibility" of actors within a network. A review of key centrality concepts can be found in the papers by Freeman, *et al*. [9]. Their work has significantly contributed to the

conceptual clarification and theoretical application of centrality. This paper provides two measures of centrality termed *degree*, and *Eigen vector centrality*. We also discuss a recently introduced *dependence centrality* measure. This work was partially motivated by the structural properties of the center of a star graph. The most basic idea of degree centrality in a graph is the adjacency count of its constituent nodes.

Formally a centrality measure is a function C which assigns every vertex $v \in V$ of a given graph G a value $C(v) \in \mathbb{R}$. As we are interested in the ranking of the vertices of the given graph G we choose the convention that a vertex u is more important than another vertex v, if $C(u) > C(v)$. We will now explain two centrality measures.

"The *degree* of a node v is simply the count of the number of encounters with other nodes that are adjacent to it, and with which it is, therefore, in direct contact" [23]; it is known as a measure of activity. The degree centrality $C_d(v)$ of a vertex v is simply defined as the degree $d(v)$ of v if the considered graph is undirected. The degree centrality is, e.g., applicable whenever the graph represents something like a voting result. These networks represent a static situation and we are interested in the vertex that has the most direct votes or that can reach most other vertices directly. The degree centrality is a local measure, because the centrality value of a vertex is only determined by the number of its neighbors. The most commonly employed definition of degree centrality is:

$$C_d(u) = \sum r(u, v) \tag{1}$$

Where $r(u, v)$ is a binary variable indicating whether a link exists between nodes u and v.

A more sophisticated version of the same idea is the so-called Eigenvector centrality. *Eigenvector centrality* C_{ev} of a node in a network is defined to be proportional to the sum of the centralities of the node's neighbours, so that a node can acquire high centrality either by being connected to a lot of others (as with simple degree centrality) or by being connected to others that themselves are highly central.

The eigenvector centrality can be understood as a refined version of degree centrality in the sense that it recursively takes into account how neighbour nodes are connected. The idea is that even if a node has a few ties, if those few nodes influence many others (who themselves influence still more others), then the first node in that chain is highly influential. For example, in terrorist networks, if a person has the potential to get bomb making training from a few neighbours', and those neighbours' have already trained in terrorist camps and are on high security risk, the potential risk of getting bomb making training for the first person is very high.

It is defined as:

$$x_i = \frac{1}{\lambda} \sum_{j=1}^{n} A_{i,j} x_j \tag{2}$$

where n is the total number of nodes in the network and λ is the largest Eigen-value to assure the centrality is non-negative. Thus, x_i is the i^{th} component of eigenvector associated with the largest Eigen-value λ of the network. While the eigenvector centrality of a network can be calculated via standard methods [10] using the adjacency matrix representation of the network, it can also be computed by an

iterative degree calculation method, known as the accelerated power method [11]. This model is not only more efficient, but is also consistent with the spirit of the refined version of degree centrality.

2.1 Hurdles

The main hurdle in detecting high value individuals using centrality measures is that centrality values of different nodes can be equal, thus minimizing the chances of identifying the powerful nodes among them. The other problem is that if two or more nodes qualify as powerful nodes over another particular node, then it becomes very difficult to identify, which node will be the real parent of that particular node when attempting to construct a hierarchy. These points are explained as below:

1. If too many nodes within a graph equalize in terms of values of centralities, it is cumbersome to identify which one will be a parent and which one will be a child, if we consider a hierarchy as a tree.
2. If two nodes '*A*' and '*B*' both qualify as influential over another node '*C*' in a graph, where should '*C*' be placed in the hierarchy of that graph.

There are two options: to place node *C* as a child of *A,* or as a child of *B*. But we have to choose an option and it should be correct. To make that choice we can compare the values of *A* and *B*, but this approach then suffers from point 1 described above. As well is the other problem that it doesn't consider '*C*' in this comparison, which should be included somewhere in the comparison of '*A*' and '*B*' because it is '*C*' whose parent is going to be decided in that comparison. We call this dilemma the '*ABC*' problem in the rest of discussion.

The node over which 'C' depends mostly is its parent. If we take the *ABC* example, suppose *C* depends more on **B** than on **A**. In this case *B* will be considered the parent of *C*. Of course the most important question, which arises here, how to determine which node depends on which node? The answer is the dependence centrality of a node, which gives us an idea of how much it depends on other nodes as discussed in the next Section.

2.2 Dependence Centrality

Dependence centrality represents how much a node is dependent on other nodes in a network [12]. Consider a network representing a symmetrical relation, "communicates with" for a set of nodes. When a pair of nodes (say, *u* and *v*) is linked by an edge so that they can communicate directly without intermediaries, they are said to be adjacent. Consider a set of edges linking two or more modes (*u*, *v*, *w*) such that node *u* would like to communicate with *w*, using node *v*. Dependence centrality can discover how many times node *u* uses node *v* to reach node *w* and how many shortest paths node *u* uses node *v* to reach node *w*. There can, of course, be more than one geodesic, linking any pair of nodes.

Let $\zeta_{(u,v)}(\mathrm{w})$ = dependence factor of the node *u* on node *v* to reach any other node (*i.e.*, node *w*) in the graph of communication as shown in (3):

$$\zeta_{(u,v)}(w) = \frac{occurence(u,v)}{d(u,v) \times path(u,w)} \tag{3}$$

where $occurence(u,v)$ = the number of times (shortest paths), the node *u* uses node *v* in the communication with one another, $path(u,w)$ = the number shortest paths between node *u* and node *w*, and $d(u,v)$ = the geodesic distance between node *u* and node *v*.

As mentioned earlier, the dependence centrality of a node represents how much a node is dependent on other nodes. Usually the nodes which are adjacent to a node are always important for that node, as all activities of that node depend on the nodes which are adjacent to it (or directly connected to that node).

Now we define dependence centrality as the degree to which a node, *u*, must depend upon another, *v*, to relay its messages along geodesics to and from all other reachable nodes in the network. Thus, for a network containing *n* nodes, the dependence centrality of *u* on *v* can be found by using:

$$C_{dep(u,v)} = 1 + \sum_{w=1}^{n} \zeta_{(u,v)}(w), \qquad u \# v \# w \tag{4}$$

We have used 1 in the above formula, because we expect every graph/ network we use is connected. We can calculate the dependence centrality of each vertex on every other vertex in the network and arrange the results in a matrix $D = [C_{dep(u,v)}]$. The value of the dependence matrix can be normalized by dividing each value with (*n*—1) where *n* represents the total number of nodes in the network.

Each entry in D is an index of the degree to which the node designated by the row of the matrix must depend on the vertex designated by the column to relay messages to and from others. Thus *D* captures the importance of each node as a gatekeeper with respect to each other node—facilitating or perhaps inhibiting its communication. The dependence matrix benefits an analyst by providing an even clearer picture than betweenness or closeness centrality alone, not only identifying how much a particular node is dependent on others, but also how much others depend on that particular node.

By the combination of centrality measures we have developed a model to detect the hidden hierarchy in terrorist networks. The main steps of our model are discussed in the next section.

3 Main Steps for Detecting Hidden Hierarchy

1. Using an undirected graph (we consider terrorist networks as undirected graphs); we first convert it into a *directed graph* using degree centrality and Eigenvector Centrality. For Example, if the degree centrality of one node is higher than another, then the directed link is originated from that node and points towards the other. If they are equivalent in terms of degree, the link will originate from the node with the higher Eigenvector centrality. If the Eigenvector centrality values for both nodes are equal, then we ignore the link.
2. Now we identify the parents and children pairs. For example, if we have two nodes which are competing for being the parent of a node, then we have to identify its correct parent. The correct parent will be the one which is connected with maximum neighbours. This represents the fact that the true leader, with respect to a node, is more influential on its neighbourhood.

3. Then we identify hierarchical relationships among the parents of a node.
4. Finally, we detect the parent of the node (among the possible parents) by using dependence centrality.

We identify parents in such a way that we traverse all the nodes. Then a tree structure is obtained, which we call a hierarchical chart or command structure.

4 Case Studies

In this section, we present three case studies of terrorist attacks that occurred or were planned to test the model for detecting hierarchy in terrorist networks as discussed in the previous Section.

4.1 Case 1: Oplan Bojinka Terrorist Plot[1]

Oplan Bojinka was a planned large scale attack on airliners in 1995. The term refers to "airline bombing plot" alone, or that combined with the "Pope assassination plot" and the "CIA plane crash plot". The first refers to a plot to destroy 11 airliners on January 21 and 22, 1995, the second refers to a plan to kill Pope John Paul II on January 15, 1995, and the third refers to a plan to crash a plane into the CIA headquarters in Fairfax County, Virginia and other buildings. Oplan Bonjika was prevented on January 6, and 7, 1995, but some lessons learnt were apparently used by

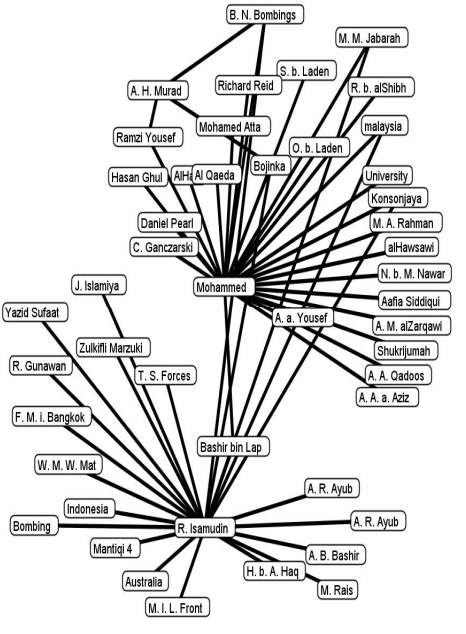


Fig. 1. Bojinka Terrorist Network

[1] http://en.wikipedia.org/wiki/The_Bojinka_Plot

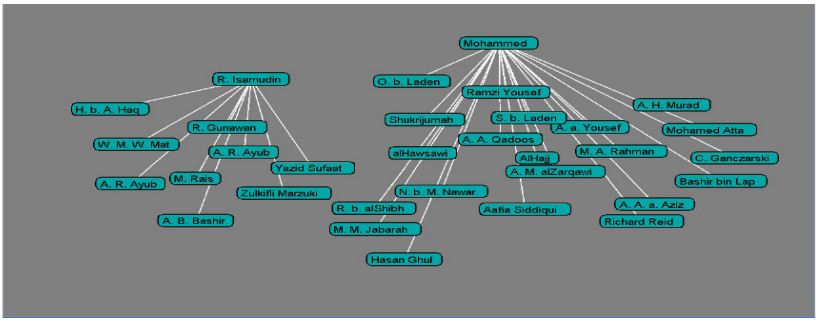


Fig. 2. The hierarchy suggests that Khalid Shaikh Mehmood and Hambali were leaders of the plot

the planners of the September 11 attacks. The money that funded operation Bojinka came from Osama Bin Laden and (R. Isamuddin) Hambali, and also from organizations operated by Jamal Khalifa, Bin Laden's brother in law. We have collected the dataset of this terrible plot which is depicted in Figure 1.

Using the model for detecting hidden hierarchy, we identified the hierarchical structure of the network which is shown in Figure 2.

The hierarchical chart shows two clusters; one is headed by Khalid Shaikh Mehmood and another by R. Isamuddin (known as Hambali). In reality *Khalid Shaikh Mehmood* was known as a key conspirator of the plot and *Hambali* was a key financier.

4.2 Case 2: March 11 Madrid Bombing Plot[2]

The 2004 Madrid train bombings (also known in Spanish as 11-M) consisted of a series of coordinated bombings against the Cercanias (commuter train) system of

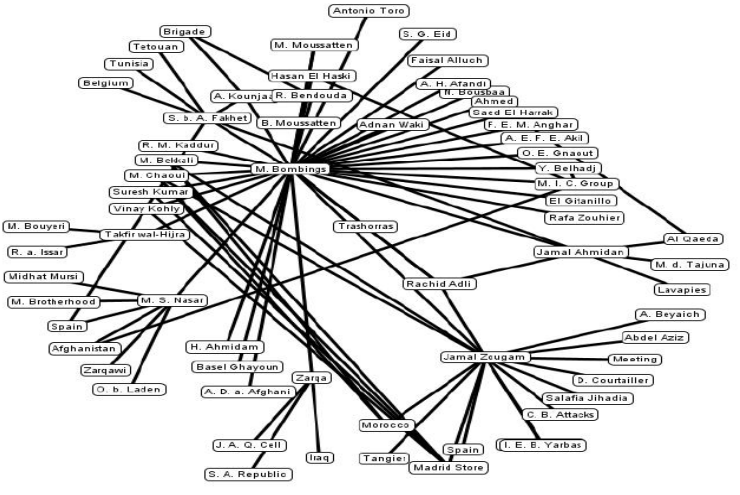


Fig. 3. 11-M Terrorist Network

[2] http://en.wikipedia.org/wiki/2004_Madrid_train_bombings

Madrid, Spain on the Morning of 11th March 2004 (three days before Spain's general elections), killing 191 people and wounding 2,050. The official investigation by the Spanish Judiciary determined the attacks were directed by an Al-Qaeda inspired terrorist cell although no direct Al-Qaeda participation has been established. We have collected data about this network and it is depicted in Figure 3.

Using the model for detecting hidden hierarchy we found that in the 7-M terrorist network, the main role was played by Jamal Zougam as shown in Figure 4. In reality after 21 months of investigation of the 7-M plot, judge Juan del Olmo ruled Moroccan national Jamal Zougam guilty of physically carrying out the attack [13].

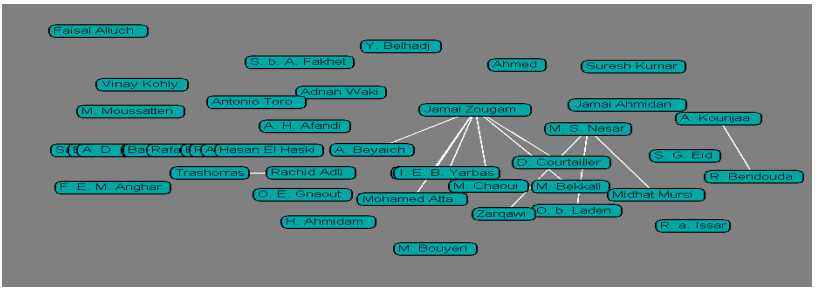


Fig. 4. The hierarchical chart of terrorists involved in 7-M

4.3 Case 3: July 07 London Bombing Plot[3]

The 7 July 2005 London bombings (also called the 7/7 bombings) were a series of coordinated terrorist bomb blasts that hit London's public transport system during the

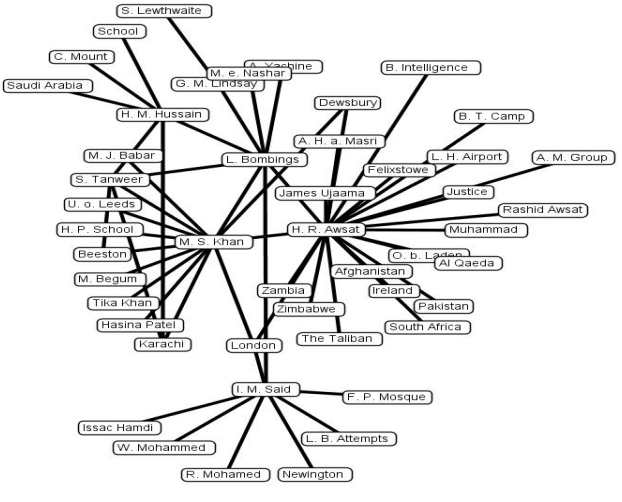


Fig. 5. The 7/7 London Bombing terrorists Network

[3] http://en.wikipedia.org/wiki/7_July_2005_London_bombings

morning rush hour. At 8:50 a.m., in which three bombs exploded within fifty seconds of each other on three London Underground trains. A fourth bomb exploded on a bus nearly an hour later at 9:47 a.m. in Tavistock Square. The bombings killed 52 commuters and the four suicide bombers, injured 700, and caused a severe day-long disruption of the city's transport and mobile telecommunications infrastructure countrywide.

We have collected data about the network involved in this tragic attack. The network is shown in Figure 5.

Using the model for detecting hidden hierarchy in the 7/7 terrorists network, we found that Haroon Rashid Aswat was the key person behind this plot. In reality Haroon Rashid Aswat was the mastermind behind the 7/7 bombing in London [14].

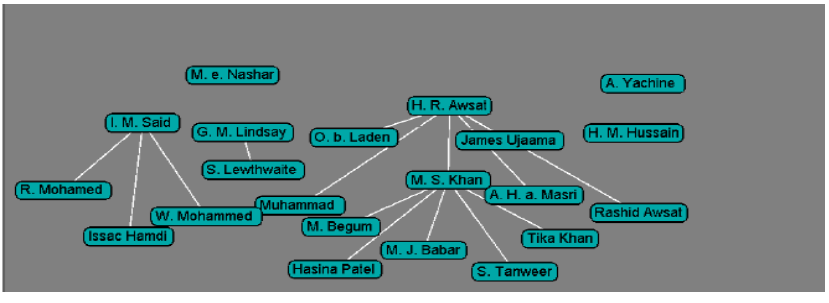


Fig. 6. The hierarchical chart of terrorists involved in 7/7

5 Data Collection

Data collection is difficult for any network analysis because it is difficult to create a complete network. It is not easy to gain information on terrorist networks. It is a fact that terrorist organizations do not provide information on their members and the government rarely allows researchers to use their intelligence data [15]. A number of academic researchers [16] [17] [18] focus primarily on data collection on terrorist organizations, analyzing the information through description and straightforward modeling.

One promising activity is the development of a major terrorism web portal at the University of Arizona's Artificial Intelligence Center. This website makes social network tools and data related to terrorism publicly available. One example is the Terrorism Knowledge Portal [19], a database consisting of over 360,000 terrorism news articles and related Web pages coming from various high-quality terrorism Web sites, major search engines, and news portals. By providing publicly available network tools and data, the research opens itself to a number of new scholars. Academics can double-check the work of others to ensure quality. New scholars can enter the field without the lengthy time commitment and financial cost of developing basic tools and getting data. Such activities, combined with the federal government's support, will help push the field of terrorism-related social network analysis to new heights in the future.

Despite their strength, researchers' works have a few key drawbacks. By dealing with open sources, these authors use limited data sets. With open sources, if the author does not have information on terrorists, he or she assumes they do not exist. This can be quite problematic as the data analysis may be misleading. For example, if one cannot find an Al Qaeda operative in Denmark in publicly available sources, the researcher could assume there is no Al Qaeda network. However, it is highly probable that this is not the case, since terrorists generally try to keep a low profile before committing an attack. The data collectors can also be criticized because their work is more descriptive and lacks complex modeling tools. Fostering relationships with modelers could augment the work being conducted by data collectors, as statistical analysis might be able to take into account some of the limitations of the data and provide an additional analytical framework.

To counter the information scarcity, we have developed a knowledge base (we call it iMiner Knowledge base) about the terrorist attacks that have occurred in the past and the information about terrorist organizations that were involved in those events. This information is mostly collected from open source media (but authenticated websites), such as http://www.trackingthethreat.com/.

The focus of the knowledge base we have developed is the agglomeration of publicly available data and the integration of the knowledge base to an investigative data mining software prototype. The main objective is to investigate and analyze terrorist networks to find hidden relations and groups, prune datasets to locate regions of interest, find key players, characterize the structure, trace a point of vulnerability, detect efficiency of the network and discover the hidden hierarchy of the non-hierarchical networks. The iMiner knowledge base consists of various types of entities. Here is an incomplete list of the different entity types:

- Terrorist organizations such as Al Qaeda
- Terrorists such as Osama Bin Ladin, Ramzi Yousef, etc.
- Terrorist facilities such as Darunta Training Camp, Khalden Training Camp, etc.
- Terrorist events/ attacks such as 9/11, London 7/7, *etc*.

The iMiner system applies a spider management system as well as spiders to import data that is available in the Web. We have developed a prototype system to get information from online repositories and save it in its knowledge base for analysis [20].

6 Discussion and Conclusion

Terrorist networks consist of many cells that attend to particular parts of their environment. Hierarchy is used to manage the interdependencies between the different subunits that are not easily resolved through direct interaction [21]. It also resolves issues by flat, balancing incentives at the sub-organizational level with the interest of (terrorist) network as a whole [22]. A separate benefit comes from the ability of a hierarchy to allow those *lower down* in the network to deal with the routine issues, reserving more unusual issues for specialist problem (strategic, tactical, executive) solvers through exception management [23]. It also has the power to change the direction of the organization through substantive decisions on *what should be done* or through asset allocations [24].

Hierarchy's ability to take the role of network (organizational) architect and devise new ways of dividing labour allows a new set of frames to emerge. A new decomposition allows a new set of informational inputs to percolate through the network, and for new views to be formed [25] [26]. This is the *cybernetic control* function of hierarchy.

Hierarchy can also help to provide some real-time control of terrorist network's (organization) routine mode of operation, at the level of both actions and cognition. In terms of action, it can step in to block routines that are not functional and override or restructure a proposed course of action to the extent possible. In the terms of cognition, it can help assimilate crucial information that may not be evident to any terrorist network cells, and provide the opportunity to reframe the challenges faced.

In this paper, we presented the results of our findings based on limited exercises in exploring the utility of the algorithm for detecting a hidden hierarchy in terrorist networks. The proposed model could be used for law enforcement agencies for the destabilization of terrorist networks by capturing key nodes. We are also confident that real-time or near real-time information from a multiplicity of databases could have the potential to generate early warning signals of utility in detecting and deterring terrorist attacks. It is necessary, of course, to have *experts* in the loop. This research has provided substantive and in-depth analysis of terrorist networks. Furthermore this analysis has provided a richer and deeper understanding and insight into terrorist networks and has provided approaches to destabilize the networks. The results achieved from the datasets serve as an excellent representation of reality.

Although real world data is hard to come by, the datasets we collected by harvesting the Web provides an excellent starting point for analyzing terrorist networks. Also, we have applied the algorithm for detecting hidden hierarchy on the datasets we collected in our knowledge base, and the results achieved are promising.

This investigation is, of course, not without shortcomings. The datasets are collected from the Internet, without any validation of data records from counterterrorism agencies. So despite thorough investigation, elements of this paper cannot help but be speculative, especially to the extent to which it relies on open source datasets. Also, there remains the possibility that other factors, unknown to the authors might have been at play. Thus, the paper offers a lens, and although this hopefully provides some new perspective, it does not offer a comprehensive or exclusive account.

Future research directions that we plan to work on and study include but are not limited to:

- Connection diversity: the links between nodes could have different weights, directions and signs - some of the nodes could indicate counterterrorism agents affecting signal transmission.
- Network evolution: the wiring diagram could change over time. Terrorists get eliminated and new ones are recruited.
    - Develop a tool that simulates network evolution including hidden node creation – terrorists that exist but are not yet identified.
    - Allow nodes to connect dynamically over time (set probabilities of creating connections). This way we could simulate terrorist recruiting process.

  - Critical times: what is the network topology before an attack?
  - Study the emergence of leadership – especially after a leader's elimination
- Dynamical complexity: the nodes could be nonlinear dynamical systems.
  - We could study signal transmission across networks.
  - Sub network synchronization – especially right before and during attacks.

References

1. Penzar, D., Srbljinović, A.: About Modeling of Complex Networks with Applications to Terrorist Group Modeling. J. Interdisciplinary Description of Complex Systems 3(1), 27–43 (2005)
2. Grier, P.: The New Al Qa'ida: Local Franchiser, Christian Science Monitor (2005) (July 11, 2005), http://www.csmonitor.com/2005/0711/p01s01-woeu.html (Accessed on May 26, 2006)
3. Wiktorowicz, Q.: The New Global Threat: Transnational Salafis and Jihad. Middle East Policy 8(4), 18–38 (2001)
4. Felter, J., et al.: Harmony and Disharmony: Exploiting al-Qa'ida's Organizational Vulnerabilities, pp. 7–9. United States Military Academy, West Point (2006)
5. Ravasz, E., Barabasi, A.L.: Hierarchical Organization in Complex Networks. J. Physical Review E 67, 026112
6. Costa, L.D.F.: Hierarchical Backbone of Complex Networks. Physical Review Lett. 93, 098702 (2004)
7. Trusina, A., Maslov, S., Minnhagen, P., Sneppen, K.: Hierarchy Measures in Complex Networks. Physical Review Lett. 92, 178702
8. Variano, E.A., et al.: Networks Dynamics and Modularity. Physical Review Lett. 92, 188701 (2004)
9. Freeman, L.C., Freeman, S.C., Michaelson, A.G.: On Human Social Intelligence. Journal of Social and Biological Structures 11, 415–425 (1988)
10. William, P., et al.: Numerical Recipes: The Art of Scientific Computing. Cambridge University Press, Cambridge (2002)
11. Hotelling, H.: Simplified calculation of principal components. Psychometrika 1, 27–35 (1936)
12. Memon, N., Hicks, D.L., Larsen, H.L.: How Investigative Data Mining Can Help Intelligence Agencies to Discover Dependence of Nodes in Terrorist Networks. In: Alhajj, R., Gao, H., Li, X., Li, J., Zaïane, O.R. (eds.) ADMA 2007. LNCS (LNAI), vol. 4632, pp. 430–441. Springer, Heidelberg (2007)
13. 11 March Madrid Train Bombing, http://en.wikipedia.org/wiki/2004_Madrid_train_bombings (Accessed on December 22, 2007)
14. John Loftus On Fox TV Claims London Bombings 'Mastermind' Is MI6 Double Agent, http://www.btinternet.com/~nlpWESSEX/Documents/WATLoftusMI6.htm
15. Ressler, S.: Social Network Analysis as an Approach to Combat Terrorism: Past, Present, and Future Research (2006)
16. Krebs, V.E.: Mapping Networks of Terrorist Cells. Connections 24(3), 43–52 (2002)

17. Sageman, M.: Understanding Terrorist Networks. University of Pennsylvania Press, Philadelphia (2004)
18. Rodriquez, J.A.: The March 11th Terrorist Network: In its weakness lies its strength. In: Proc. XXV International Sunbelt Conference, Los Angeles (2005)
19. Reid, E., et al.: Terrorism Knowledge Discovery Project: A Knowledge Discovery Approach to Addressing the Threats of Terrorism. In: Chen, H., Moore, R., Zeng, D.D., Leavitt, J. (eds.) ISI 2004. LNCS, vol. 3073, pp. 125–145. Springer, Heidelberg (2004)
20. Memon, N., et al.: Harvesting Terrorists Information from Web. In: Proc. 11th International Conference Information Visualization (IV 2007). IEEE Press, Los Alamitos (2007)
21. Thompson, J.D.: Organizations in Action. McGraw-Hill, New York (1967)
22. Williamson, O.E.: The Economic Institutions of Capitalism. The Free Press, New York (1985)
23. Garicano, L.: Hierarchies and Organization of Knowledge in Production. J. Political Econom. 108(5), 874–905 (2000)
24. Bower, J.L.: Managing the Resource Allocation Process. Harvard Business School Press, Boston (1974)
25. Simon, H.A.: The Architecture of Complexity. Proc. Amer. Philos. Soc. 106, 467–482 (1962)
26. Jacobides, M.: The Architecture and Design of Organizational Capabilities. Indust. Corporate Change 15(1), 115–169 (2006)

Keyphrase Extraction from Chinese News Web Pages Based on Semantic Relations

Fei Xie[1,4], Xindong Wu[1,2], Xue-Gang Hu[1], and Fei-Yue Wang[3]

[1] School of Computer Science and Information Engineering,
Hefei University of Technology, Heifei 230009, China
[2] Department of Computer Science, University of Vermont, Burlington, VT 50405, U.S.A.
[3] Institute of Automation, Chinese Academy of Sciences, Beijing, China
[4] Department of Computer Science and Technology, Hefei Teachers College,
Hefei 230061, China
xiefei9815057@sina.com, xwu@cs.uvm.edu, jsjxhuxg@hfut.edu.cn,
feiyue.wang@ia.ac.cn

Abstract. Keyphrases are very useful for saving time on browsing through the news web pages. A new keyphrase extraction method from Chinese news web pages based on semantic relations is presented in this paper. Semantic relations between phrases are analyzed, and a lexical chain is used to construct a semantic relation graph. Keyphrases are extracted and a semantic link graph is built on the lexical chains. News web pages with core hints are selected from www.163.com to test our method. The experimental results show that the proposed method substantially outperforms the method based on term frequency, especially when the number of keyphrases extracted is 3 - the precision is improved by 26.97 percent, and the recall is improved by 20.93 percent.

Keywords: keyphrase extraction, semantic relation, word similarity, word co-occurrence, lexical chain.

1 Introduction

With the rapid development of the Internet, more and more electronic news resources have been made available and many people are spending much time on browsing the Internet news. Keyphrases are defined as phrases that capture the main topics discussed in a document. Only a minority of web pages have keyphrases assigned to them because manual assignments of keyphrases are expensive, time-consuming and quite subjective. This motivates the research in finding automated approaches to keyphrase extraction from news web pages.

Research in keyphrase extraction began in early 1950's. Existing work can be categorized into two major approaches: supervised approaches and unsupervised approaches. Supervised approaches view keyphrase extraction as a classification task. Turney [1] designed a keyphrase extraction system GenEX based on C4.5. Witten [2] used Naive Bayes to extract keyphrases, and designed a Kea system. Supervised methods require a large amount of training data. Many documents with known

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 490–495, 2008.

keyphrases are needed. Unsupervised keyphrase extraction does not need training dada that exploits the structure of the text itself to determine keyphrases that capture the topic of the text. Mihalcea [3] presented a graph-based ranking approach to keyphrase extraction.

The study of Chinese keyphrase extraction began in recent years. Li Su-Jian [4] probed into the keyphrase extraction using the Maximum Entropy (ME) model. Liu Yuan-Chao [5] mined the manually labeled keyphrase corpus which come from People's Daily and attained the constructed rules for Chinese keyphrase extraction. Suo Hong-Guang [6] presented a lexical-chain-based keyphrase extraction method for Chinese documents. The lexical chain is constructed based on HowNet-based word semantic similarity by Li Su-Jian [7]. Word similarity is computed by HowNet, but the similarity of the phrases not in HowNet is difficult to compute.

A new keyphrase extraction method for Chinese news web pages based on semantic relations is presented in this paper. Two kinds of phrase relations based on HowNet and the word co-occurrence model are studied. Semantic relations of the phrases not in HowNet are computed by a word co-occurrence model. Lexical chains are constructed to represent semantic relations and build semantic links between phrases.

The rest of the paper is organized as follows. Section 2 reviews related work on word similarity and lexical chains. Section 3 presents our algorithm for keyphrase extraction based on semantic relations. Section 4 provides experimental results. Section 5 concludes with suggestions for future work.

2 Related Work

2.1 Word Similarity Based on HowNet

HowNet is a common-sense knowledge base unveiling inter-conceptual and inter-attribute relations of concepts as connoting in lexicons of the Chinese and their English equivalents [8]. There are two important terms in HowNet: concept and sememe. Concept is the semantic description of phrases. Each phrase has several concepts. Concept is defined by a kind of knowledge representation language named sememe that is the smallest basic semantic unit.

Given two phrases W_1 and W_2, W_1 has n concepts, S_{11}, S_{12}, …, S_{1n}, and W_2 has m concepts, S_{21}, S_{22}, …, S_{2m}, the similarity between W_1 and W_2 is defined as follows:

$$Sim(W_1, W_2) = \max_{i=1...n,\, j=1...m} Sim(S_{1i}, S_{2j}) \tag{1}$$

A concept is described by sememes. Sememe similarity is the basis of concept similarity. Sememes in HowNet compose a hierarchical tree by the Hypernym-Hyponym relation. Suppose the length of two sememes p1 and p2 in the hierarchical tree is d. The semantic distance of two sememes is defined as follows:

$$Sim(p_1, p_2) = \frac{\alpha}{d + \alpha} \tag{2}$$

where p1 and p2 represent two sememes, d is the length of p1 and p2 in the sememe hierarchical tree, and α is a parameter that can be adjusted.

2.2 Lexical Chains

The notion of cohesion, introduced by Halliday and Hasan [9], is a device for "sticking together" different parts (i.e., phrases, sentences and paragraphs) of the text to function as a whole. Five cohesion relations are defined: reference, substitution, ellipsis, conjunction, and lexical cohesion. Lexical cohesion occurs not only between two terms, but also among sequences of related words, called lexical chains [10]. Morris and Hirst first introduced the concept of lexical chains to segment text. Later, lexical chains were used in many areas such as text retrieval and information extraction [11].

The construction of lexical chains needs thesaurus for determining relations between phrases. It is difficult to compute the relations of phrases that do not appear in the thesaurus. For Chinese, there are many new phrases that do not appear in the thesaurus. Word co-occurrence [12] is an important model based on statistics widely used in natural language processing that reflects the relatedness of the words in the document. The frequency of two words co-occurring in the same window unit (i.e., sentence and paragraph) can be computed without thesaurus. A new lexical chain constructing algorithm based on word similarity computed by HowNet and the co-occurrence model is presented in this paper to extract news keyphrases. Semantic links of phrases in the same lexical chain are built that can string the related phrases together.

3 Keyphrase Extraction Based on Semantic Relations

Keyphrase are mainly the nouns in academic journals [13]. However, verbs also play key roles in representing the news topics. Therefore, all phrases except stop words in the news web pages are considered. Our Keyphrase Extraction based on Semantic Relations (KESR) algorithm is designed as follows:

(1) Non-news content in the news web page is filtered. Words are segmented, and stop word are removed.

(2) Compute the TFIDF [14] of each word w_i by Equation (3).

$$TFIDF_i = \frac{tf_i \times \log(N / n_i)}{\sqrt{\sum_j (tf_j \times \log(N / n_j))^2}} \tag{3}$$

where tf_i is the frequency of word w_i in the given web page, N is the number of the documents in the corpus, and n_i is the number of documents in the corpus that contain the word w_i.

(3) Select the top n words $\{w_1, w_2, \ldots, w_n\}$ from all the words segmented except stop words by TFIDF as candidate words.

(4) Compute the word similarities based on HowNet and word co-occurrence frequencies of candidate words.

(5) Select the first candidate word w_1 to construct the first lexical chain L_1.

(6) Select the candidate word w_i. If the word similarity between w_i and some word in the lexical chain L_j exceeds the threshold t_1 or the word co-occurrence frequency exceeds the threshold t_2, then w_i is inserted into L_j, else w_i constructs a new lexical chain.

(7) Repeat (5) until all the candidate words are computed.

(8) Compute the weight of the phrase w_i by Equation (4).

$$Weight(w_i) = a \times TFIDF_i + b \times |chain_i| \tag{4}$$

where $TFIDF_i$ is the TFIDF value of w_i, $|chain_i|$ is the length of the chain that contains w_i , and a and b are the parameters that can be adjusted.

(9) Select the top m words as the keyphrases extracted from the candidate words by their weights.

4 Experimental Results and Analysis

There are no standard news web pages for keyphrase extraction, and 120 news web pages with core hints are selected from the 163 website (http://www.news.163.com) as our experimental data to test KESR. Keyphrases extracted are compared with the phrases in the news title and the phrases in the core hints provided by the editor. We use recall and precision as measures of the system performance. The title recall R and core hint precision P are defined as follows:

$$R = \frac{the\ number\ of\ keyphrases\ \ extracted\ \ matching\ \ with\ \ the\ phrases\ \ in\ the\ title}{the\ number\ of\ \ phrases\ \ in\ the\ title} \tag{5}$$

$$P = \frac{the\ number\ of\ \ keyphrases\ extracted\ matching\ with\ \ the\ phrases\ in\ the\ core\ hints}{the\ number of\ \ keyphrases\ extracted} \tag{6}$$

The experimental results of KESR are compared with TFIDF (a keyphrase extraction implementation based on TFIDF). Two sets of experiments are conducted. In the first set, the news title and the core hints for each news web page are removed, and the title recall and the core hint precision are compared of the two methods. In the other set of experiments, the news title is kept while the core hints are removed, and the core hint precision is compared of the two methods. The parameters of n, a, and b selected are respectively 20, 1, and 1 by experiments. The thresholds of t1 and t2 are selected respectively 0.3 and 4. The number of keyphrases extracted is respectively selected 3, 5, and 7.

Figure 1 and Figure 2 show the title recall and core hint precision comparisons of KESR and TFIDF with the title and core hints removed.

From Figure 1 and Figure 2, we can find that KESR performs substantially better than TFIDF, and the superiority increases with the number of keyphrases extracted decreased. Especially when the number of keyphrases extracted is 3, the recall and the precision are improved respectively 20.93% and 26.97%. The semantic relations of phrases are considered in KESR on the basis of term frequency. The aim of KESR is to extract the words with a low frequency but a great contribution to the text subject

and filter the words with a high frequency but little contribution to the text subject, and the experiments have testified this purpose.

The precision comparisons of KESR and TFIDF with the title kept and the core hints removed are presented in Figure 3. Figure 3 demonstrates that KESR outperforms TFIDF with the title kept. Figure 4 shows the precision comparison of KESR with the title removed and the core hints kept. From Figure 4 we can see that the news title plays a key role in keyphrase extracting.

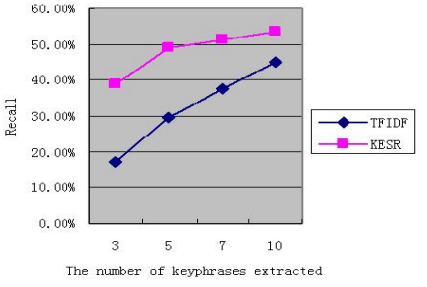


Fig. 1. Recall comparison of KESR and TFIDF with the title and core hints removed

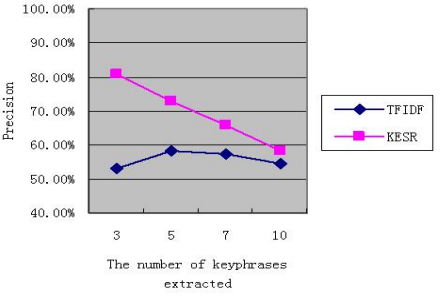


Fig. 2. Precision comparison of KESR and TFIDF with the title and core hints removed

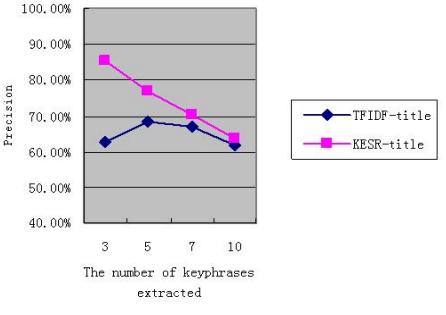


Fig. 3. Precision comparison of KESR and TFIDF with the title kept and the core hint removed

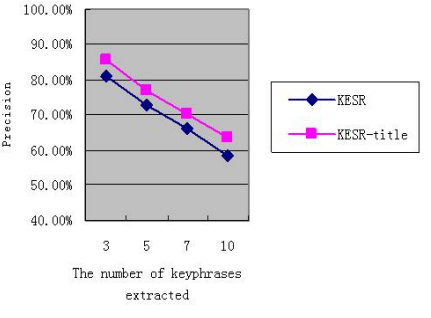


Fig. 4. Precision comparison of KESR with the title removed and the core hints kept

5 Conclusion

A new keyphrase extraction method based on semantic relations is presented in this paper. Semantic relations between phrases based on HowNet and co-occurrence are studied, and lexical chains are used to link the relations. Keyphrases with high quality are extracted based on the information in the lexical chains. The experiments have indicated that this method achieves better performances than the method based on term frequency; and the recall and the precision have both been improved

substantially. There is rich information in lexical chains, and only the lengths of chains are utilized in this paper. Future work can seek to construct better lexical chains and make a full use of the chains.

References

1. Turney, P.D.: Learning to extract keyphrases from text. National Research Council, Canada, NRC Technical Report ERB-1057 (1999)
2. Witten, I.H., Paynter, G.W., Frank, E., Gutwin, C., Nevill-Manning, C.G.: KEA: Practical automatic keyphrase extraction. In: Proceedings of the 4th ACM conference On Digital Libraries, Berkeley, California, US, pp. 254–256 (1999)
3. Mihalcea, R., Tarau, P.: Graph-based ranking algorithms for sentence extraction, applied to text summarization. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004) (companion volume), Barcelona, Spain (2004)
4. Su-Jian, L., Hou-Feng, W., Shi-Wen, Y., Cheng-Sheng, X.: Research on maximum entropy model for keyword indexing. Chinese Journal of Computers 27(9), 1192–1197 (2004)
5. Yuan-Chao, L., Xiao-Long, W., Zhi-Ming, X., Bing-Quan, L.: Ming constructing rules of Chinese keyphrase based on rough set theory. Acta Electronica Sinica 35(2), 371–374 (2007)
6. Hong-Guang, S., Yu-Shu, L., Shu-Ying, C.: A keyword selection method based on lexical chains. Journal of Chinese Information Processing 20(6), 25–30 (2006)
7. Qun, L., Su-Jian, L.: Word Similarity Computing Based on How-net. Computational Linguistics and Chinese Language Processing 7(2), 59–76 (2002)
8. Zhen-Dong, D., Qiang, D.: HowNet, http://keenage.com.cn
9. Halliday, M.A.K., Hasan, R.: Cohesion in English. Longman, London (1976)
10. Morris, J., Hirst, G.: Lexical cohesion computed by thesaural relations as an indicator of the structure of text. Computational Linguistics 17(1), 21–48 (1991)
11. Barzilay, R., Elhadad, M.: Using lexical chains for text summarization. In: Proceedings of the ACL/EACL 1997 Workshop on Intelligent Scalable Text Summarization, Madrid, Spain, July 1997, pp. 10–17 (1997)
12. Peat, H.J., Willet, P.: The limitations of term co-occurrence data for query expansion in document retrieval systems. Journal of American Society for Information Science 42(5), 378–383 (1991)
13. Chun, D.: On indexing of key words. Acta Editologica 16(2), 105–106 (2004)
14. Salton, G., Wong, A., Yang, C.S.: On the specification of term values in automatic indexing. Journal of Documentation 29(4), 351–372 (1973)

Automatic Recognition of News Web Pages

Zhu Zhu[1], Gong-Qing Wu[1], Xindong Wu[1,2], Xue-Gang Hu[1],
and Fei-Yue Wang[3]

[1] School of Computer Science and Information Engineering,
Hefei University of Technology, Heifei 230009, China
[2] Department of Computer Science, University of Vermont, Burlington, VT 50405, U.S.A.
[3] Institute of Automation, Chinese Academy of Sciences, Beijing, China
zz2001_99@sohu.com, wugongqing@gmail.com, xwu@cs.uvm.edu,
jsjxhuxg@hfut.edu.cn, feiyue.wang@ia.ac.cn

Abstract. The information on the World Wide Web is congested with large amounts of news contents. The filtering, summarization and classification of news Web pages have become hot topics of research, aiming for useful news contents. Accurately identifying news Web pages is a crucial problem in these research topics. To solve this problem, this paper proposes an automatic recognition method for news Web pages based on a combination of URL attributes, structure attributes and content attributes. Our experimental results demonstrate that this method provides a high accuracy of above 96% with the recognition of news Web page.

Keywords: news Web pages, attribute extraction, classification, automatic recognition.

1 Introduction

Along with the explosive development of the World Wide Web, the information on Web pages is rapidly inflated and congested with large amounts of news contents. To identify useful information that satisfies the users' requirements, the filtering and summarization of news Web pages have become hot topics in Web intelligence research. The primary task of our research is to quickly and accurately identify news Web pages.

At present, the number of Web pages is increasing significantly. The studies of classification methods of Web pages mainly utilize automatic classification. The main classification approaches can be summarized as follows.

(1) Content-based analysis[1,2,3]. This approach utilizes traditional text classification methods, based on feature vectors, which turn a text document into a characteristic vector (w_1, w_2, …, w_n) and then utilize different kinds of classification algorithms to carry out the classification. The accuracy of classification will be reduced when the Web pages contain few feature words and descriptive contents.

(2) The link-based analysis approach[4,5,6]. This approach utilizes the hyperlinks contained in the Web pages for classification. Many studies indicate that efficient utilization of the links contained within Web pages can improve the performance of classification[6,11]. However, though a large number of hyperlinks provide much

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 496–501, 2008.

valuable information for Web page classification, there are still many potential problems. First, information of the hyperlinks cannot directly reflect the contents of Web pages. Secondly, there are many noisy hyperlinks in Web pages. The links that are useful to classification are usually scarce. Whether noisy hyperlinks can be effectively filtered directly affects the performance of classification of Web pages.

(3) Combined analysis[7, 8, 9, 10]. This approach combines the content-based analysis and the link-based analysis to enhance the accuracy of classification. However, the utilization of hyperlinks based analysis to strengthen the content-based analysis can lead to two problems. First, when combining the two approaches, the content-based analysis would probably filter away some hyperlinks that are important to the performance of classification [7]. Secondly, it might increase the error rate of classification if we simply add neighbor documents to the contents based analysis, and the reason is that there exist a large number of noisy feature words [12, 13]. Therefore, these studies indicate that we need to be careful when selecting neighboring Web pages.

We can find that Web page classification is difficult and is not reasonable to represent Web pages using pure text, because the Web pages contain far more abundant information than just text. First, the forms of Web pages are flexible and many different kinds of formats coexist. The same format could have many different standards. Besides, Web pages contain abundant structural information, and whether this information can be utilized effectively will certainly affect the performance of the classifiers. The rest of the paper is organized as follows. In Section 2, we introduce the selection of related attributes for automatic recognition of news Web pages. In Section 3, we present and analyze our experimental results. At last, Section 4 provides conclusions.

2 Interrelated Attributes in Automatic News Web Page Recognition

Our method of automatic recognition of news Web pages is based on the selection of important attributes and then the utilization of classification algorithms to identify news Web pages. Existing classification algorithms are well developed, and the key to accurate news Web page recognition is which attributes are used to represent the features of the given news Web page. In this paper, a decision tree learning technique, C4.5, is applied for the automatic recognition of news Web pages. Based on the Web pages that are classified as the news category, we can carry out further studies, for example, filtering and summarization of news Web pages.

2.1 URL Attributes of News Web Pages

We notice that the URL of a news Web page has the following characteristics. First, most of the time, the URL contains time related attributes. Second, the structures of the URLs from different websites are similar. URL attributes include positive attributes and passive attributes.

Positive attributes. Time attributes, second level domain attributes, and first level catalog attributes.

(1) Time attributes. We randomly selected 478 URL addresses of news Web pages through Baidu news search, and the test results demonstrate that 97.4% of these news Web pages contain time attributes.

(2) Second level domain attributes. We discovered that the same departments of different Web pages share similar structure attributes. In this paper, we selected news Web pages' URL addresses from the 10 most visited news websites in China, including 163, sohu, sina, tom, QQ, CCTV, China news net, the Chinese news net, the New China net, and the Chinese youths, to create the training corpus, and our training obtained 59 attributes of the second level domains. Second level domain attributes are also crucial for the correct recognition of news Web pages.

(3) First level catalog attributes. We discovered through statistics that there are only 2.6% URLs of news Web pages that do not contain time attributes in the 478 Web pages that were randomly selected. There are several first level catalog attributes of news Web pages such as "news", "newshtml" and "newscenter".

Passive attributes. Passive attributes can be utilized to eliminate the Web pages that are not news Web pages. In this paper, passive attributes selected include "index", "blog", "bbs", "video", and the end of URL with "/".

2.2 Structure Attributes of News Web Pages

The structure and content of a news Web page affect the performance of the classifiers, in addition to the URL attributes. The Web pages contain rich structure information, and if used correctly, it can enhance the accuracy of recognition. With this starting point, we have found that some structure attributes contribute to the recognition, in addition to pure text in the news Web pages, including the title and subtitle of a Web page, labeled by <title> and <H_n>tags, and that carves up the hierarchy of the Web page. Picking up and analyzing these structure characters from news Web pages contribute to the news Web page recognition.

In this paper, structure attributes selected include tags such as <H_1>, …, <H_n>, <title>, and . Through analyzing several websites, we have found that the title of a Web page is usually labeled by <title>. The content extraction includes the title of a Web page and the information of the website which is spaced out by the conjunction. The information of the website can be an important attribute of the news Web page. Moreover, though statistics we have found the hierarchy labeled by between the text and the title of news Web pages, including the time attributes. Judging if the tag exists is an important attribute of classifying news web pages, such as "year, month, date, hour and minute".

2.3 Content Attributes of News Web Pages

The classification of Web pages sets an HTML file as an object, gets relevant HTML content from the URL address, and picks up some key words as characters. These key words should affect the accuracy of recognition of the news Web page.

Though statistics on 881 HTML files of non-news Web pages and 1087 HTML files of news Web pages on 163, sohu, sina, tom, QQ, CCTV, China news net, the Chinese news net, the New China net, and the Chinese youths, we find that the frequency of the key word "news" in a Web page is an important attribute for the recognition of a news Web page. It can be observed that, whether the keyword "news" appears twice can discriminate a news Web page from a non-news Web page more efficiently than three times. So, in this paper, we select "the appearing frequency of keyword "news" > 2 as another attribute of a news Web page.

In this paper, we also select the words "news centre", "report", "reporter", "editor", and "relative news" as content attributes of a news Web page.

2.4 Combined Attributes of News Web Pages

We have extracted news Web pages' attributes from the popular news websites of 163, sohu, sina, tom and QQ. Table 3 shows the extraction of attributes from news Web pages from these websites.

Table 1. Combined attributes of news web pages

URL attributes:	**Positive attributes** 1: time attributes; 2: second level domain: news/tech/stock1/ent/sports/auto/finance/book/edu/comic/games/baby/ astro/lady/chanye/www/mil/bj/ladies/2008/business/money/it/digi/ teamchina/yule/house/cul/learning/health/travel/women/nba/golf/weiqi/ music/mobile/war/discover/history/ ; 3: first level catalog: news/newshtml/newscenter **Passive attributes** 4: index; 5: the end of URL with "/"; 6: bbs; 7: blog; 8: video
Content attributes:	9: "news centre"; 10: "main body"; 11: "report"; 12: "correspondent" "author"; 13: "home news"; 14: "editor"; 15: "source" "article source"; 16: "related report" "related subject" "related link" "related news"; 17: "hot news" " list of hot comments" "hot spot comments"; 18: "news forum" "news search" "news subscription" "ranking of news" "news on move"; 19: "comments" 20: sum up of the number of times of "news" appear in the HTML page.
Structure attributes:	21: has <H_1> tag or not; 22: has <H_2> tag or not; 23: has "news" in <title> tag or not; 24: includes time feature in tag or not
	25: Strategy attributes

3 Experimental Results and Analysis

We selected 1087 news Web pages' URLs from popular news websites, such as 163, sohu, sina, tom, QQ, CCTV, China news net, the Chinese news net, the New China net, and the Chinese youths, and 881 non-news Web pages' URLs also from these websites as our experimental objects. The performance of the induced classifier is

evaluated in terms of precision. Depending on the different attributes from the different websites, we conduct 3 different sets of experiments accordingly. For the first set of experiments, we extract the attributes of news Web pages from five websites respectively which are the same as the training and testing websites. Table 2 shows the results of these experiments. For the second set of experiments, we extract the attributes of news Web pages from five websites respectively which are different from the training and testing websites. Table 3 shows the results of these experiments. For the third set of experiments, we extract the combined attributes of news Web pages from five websites, and evaluate the results using five different websites from the five websites for training and testing. Table 4 shows the results of these experiments. (Note that the websites of experiments are denoted as follows. s1: QQ, s2: sina, s3: sohu, s4: 163, s5: tom, s6: New China net, s7: Chinese news net, s8: CCTV, s9: Chinese net news, s10: Chinese youths.)

Table 2. Evaluation results of using the same websites for training, testing and feature selection

feature selection	S1	S2	S3	S4	S5	news pages	non-news pages	url	url+con	average
QQ	√					101	95	97.96%	99.49%	98.73%
sina		√				152	104	96.88%	99.47%	98.18%
sohu			√			108	70	100%	100%	100%
163				√		40	30	100%	100%	100%
tom					√	126	84	99.05%	100%	99.53%

Table 3. Evaluation results of using different websites for training, testing and feature selection

feature selection	S1	S2	S3	S4	S5	S6- S10	news pages	non-news pages	url	url+con	average
QQ		√	√	√	√	√	985	786	99.43%	99.66%	99.55%
sina	√		√	√	√	√	934	777	99.36%	99.59%	99.48%
sohu	√	√		√	√	√	978	811	98.99%	99.39%	99.19%
163	√	√	√		√	√	1046	851	99.05%	99.47%	99.26%
tom	√	√	√	√		√	960	797	98.98%	99.32%	99.15%

Table 4. Combined feature selection: evaluation results of using five different websites from the five websites for training and testing

feature selection	s1-s5	s6- s10	news pages	non-news pages	url	url+con	average
QQ,163,sina,tom,sohu	√		559	498	98.98%	99.89%	99.39%
QQ,163,sina,tom,sohu		√	527	383	99.62%	99.34%	99.48%

From Tables 2, 3 and 4, we can see that a comprehensive usage of URL features, structure features and content features of news Web pages in the selection of attributes can get an accurate C4.5 classifier when identifying news Web pages. The results of experiments indicate that the accuracy of our proposed automatic recognition method is above 96%.

4 Conclusions

In this paper, we have designed a news Web page classifier based on the analysis of news Web page's URL structure, and the Web page's structure and content information. We have used this classifier to identify news Web pages and have demonstrated a satisfactory performance of above 96% accuracy. Because of the selection of attributes at the present stage is based on statistics, the issue of attribute noise may possibly exist, and how to choose more reasonable attributes more effectively will be our next research effort. We will also consider other machine learning methods than C4.5, for news Web pages' recognition.

References

1. Guan, T., Wong, K.F.: KPS-A Web Information Mining Algorithm. In: The 8th International World Wide Web Conference, pp. 1495–1507 (1997)
2. Kleinberg, J.M.: Authoritative Sources in a Hyperlinked Environment. Journal of the ACM 46(5), 604–632 (1999)
3. Kwon, O.W., Lee, J.H.: Web Page Classification based on k-Nearest Neighbor Approach. In: The 5th International Workshop on Information Retrieval with Asian Languages, pp. 9–15. ACM, New York (2000)
4. Yang, Y., Slattery, S., Ghani, R.A.: A study of app roaches to hypertext categorization. Intelligent Information Systems 18(2/3), 219–241 (2002)
5. Furnkranz, J.: Exploiting structural information for text classification on the WWW. In: DA 1999, pp. 487–497. Springer, Amsterdam (1999)
6. Shen, D., Sun, J.-T., Yang, Q., Chen, Z.: A Comparison of Implicit and Explicit Links for Web Page Classification. In: The World Wide Web Conference Committee (IW3C2). ACM 1-59593-323-9/06/0005
7. Chakrabarti, S., Joshi, M., Tawde, V.: Enhanced Topic Distillation using Text, Markup Tags, and Hyperlinks. In: The ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 208–216. ACM, New York (2001)
8. Kuo, Y.H., Wong, M.H.: Web Document Classification based on Hyperlinks and Document Semantics. In: Mizoguchi, R., Slaney, J.K. (eds.) PRICAI 2000. LNCS, vol. 1886, pp. 44–51. Springer, Heidelberg (2000)
9. Kan, M.-Y.: Web page categorization without the web page. In: WWW 2004, May 17–22, ACM, New York (2004) 1-58113-912-8/04/0005
10. Yan, F., et al.: Using Naive Bayes to Coordinate the Classification of Web Pages. Journal of Software (in Chinese) 12(9), 1386–1392 (2001)
11. Xie, W., Mammadov, M., Yearwood, J.: Using Links to Aid Web Classification. In: ICIS 2007 (2007) 0-7695-2841-4/07
12. Ng, A.Y., Zheng, A.X., Jordan, M.I.: Link Analysis, Eigenvectors and Stability. In: The 7th International Joint Conference on Artificial Intelligence, pp. 903–910. Morgan Kaufmann, San Francisco (2001)
13. Ng, A.Y., Zheng, A.X., Jordan, M.I.: Stable Algorithms for Link Analysis. In: The ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 258–266. ACM, New York (2001)

Understanding Users' Attitudes Towards Using a VoIP Survey

Hsiu-Mei Huang, Chun-Hung Hsieh, Pei-I Yang, Chung-Min Lai, and Wei-Hong Lin

Graduate School of Computer Science and Information Technology, National Taichung Institute of Technology,
129, Sec. 3, Saming Rd., Taichung 404, Taiwan, ROC,
{s18943112,s18953114}@ntit.edu.tw, xyz3028@msn.com

Abstract. The Internet phone is a popular new technology. Several researches have been done on Internet phone technology. Many researches were done on how to make use of the Internet phone. The purpose of this paper is to explore users' perceptions on the use of the Internet phone as a survey tool.

There are five factors that include usefulness, ease-of-use, perceived liking, perceived self-efficacy and computer anxiety in this study. This questionnaire was designed based on prior factors. Next, the questionnaire is used to collect data that explores users' intentions to using a VoIP survey. The analysis tool of this study is to apply cluster analysis, discrimination analysis and an independent-sample T test.

The results show that the k-means clustering is able to detect factor relationships and patterns for users' attitudes towards a VoIP survey. In this study, k-means clustering classifies users' attitudes into high pleasure and low pleasure towards a VoIP survey. After that, classification results indicate 97.3% of original grouped cases correctly classified and 97.3% of cross-validated grouped cases correctly classified.

Finally, this study contributes to understanding that both self-efficacy and perceived anxiety are main factors to the impact users' attitudes toward a VoIP survey.

Keywords: Technology Acceptance Model, Internet Phone Survey, Two-stage, Clustering Approach, C5.0 Decision Tree.

1 Introduction

In the early 1960s, the Internet brought some progressive thinking by people who saw a mighty change in allowing computers to share information and send images or voices. The public domain of the questionnaire has been provided to explore users' attitudes and intentions. Advanced technology has been speedily applied in research individuals' attitudes for example email survey, Web survey. VoIP (Voice over Internet Protocol) is technology at present, so we have attempted to explore users' attitudes towards the VoIP as a survey tool. It involves the transmission of common telephone calls over the Internet. In other words, VoIP can send voices over the Internet that are not through the normal telephone network. Traditional telephones were at high costs in market. The purpose of the study will lead factors in individuals' adoption of the VoIP survey.

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 502–507, 2008.

2 Theoretical Background

2.1 Technology Acceptance Model (TAM)

Theory of reasoned action (TRA) is usually discussed human behavior action [1] [2]. Davis [3] first developed the technology acceptance model (TAM) as a theoretical extension of theory of reasoned action (TRA). Davis [3] [4] suggested TAM, now robust and powerful, has been a widely cited model for predicting and explaining user behavior and IT usage [3] [4]. TAM, suitable for TRA, suggests that two particular beliefs, perceived usefulness and perceived ease of use, are the primary factors for technology acceptance [7]. Perceived usefulness is defined as "the degree to which a person believes that using a particular system would enhance his/her job performance". Perceived ease of use is defined as "the degree to which a person believes that using a particular system would be free of physical and mental effort" [3]. Therefore, perceived usefulness and perceived ease of use both effect upon individual intention.

2.2 Perceived Anxiety

Computer anxiety has been defined as an aversion, fear or apprehension toward interacting with computers [5]. Computer anxiety is described as an effective response. That is, an emotive fear of potential negative outcomes.

3 Survey Instrument and Data Collection

This questionnaire was applied in a seven-point Likert scale (1=very strongly disagree, 4=neutral, 7=very strongly agree). The questionnaire consisted of six major parts. The first part of the questionnaire is basic personal information such as gender, educational background, computer experience, internet experience, VoIP experience, and using VoIP weekly. The second part of the questionnaire is perceived anxiety such as voice quality, question number, answer time, and different VoIP callers. The third part of the questionnaire is perceived liking such as liking VoIp, convenience, anonymity, and security. The fourth part the questionnaire is perceived self-efficiency such as computer use and VoIP use. The fifth part the questionnaire is ease of use such as computer use, VoIP use, and answering questions. The sixth part the questionnaire is usefulness such as the quick data collective questionnaire, saving effect, and saving money.

The total samples of 350 participants were collected from the survey. Since respondants checked, 55 respondants were excluded from the analysis because important questions were blank or written twice. Eventually, 295 respondants were used for data analysis. The return rate of the study was 84.3%. The sample was studying at a college in the central Taiwan area. From the total sample of 68.6% were female and 31.4% were male. The age constituent is broken down into the following categories: 16–20, 13.1%; 21–25, 75.1%; over 26, 11.8%, as shown in Table 1. The participants had VoIP experience and had been using it for more than 6 months (63%).

4 Results

4.1 Factor Analysis

The Factor analysis is a technique that needs a large sample size to hold stabilization because it is based on the correlation matrix of the variables and correlations involved. Tabachnick and Fidell [8] suggest regarding a sample size: 50 cases is very poor, 100 is poor, 200 is fair, 300 is good, 500 is very good and 1000 or more is excellent.

The main applications of the factor analytic techniques includes both to reduce the number of variables and to discern classification variables structure in the relationships between variables. Hence, the factor analysis is applied as a data reduction and variables structure detection method. Furthermore, principal component analysis uses a varimax rotation. The most usual and reliable criterion is the use of extracting eigenvalues factors. All factors with eigenvalues greater than 1 are held as being significant items; if factors with less than 1 are discarded.

The KMO value is equal to 0.864, therefore it shows the value is adaptive to create a factor analysis. As noted above, 22 factor items for the VoIP survey is factor analyzed as a known in Table 2. The factor analysis resulted in five components of factor: components have 4, 4, 3, 3, 8 items that are labeled as “perceived anxiety” , “perceived liking” ,“perceived self-efficacy” ,“ease of use” and “usefulness” ,respectively. All five components have eigenvalues greater than 1 and account for 70.169% of total variance.

4.2 Reliability

A reliability α (Cronbach α) was computed to check the internal consistency of items with each dimension. All factors with reliability α above 0.7 were deemed acceptable values. For reliability, the Cronbach’s α of this instrument were 0.822, 0.813, 0.849, 0.859 and 0.920 for perceived anxiety, perceived liking, perceived self-efficacy, ease of use and usefulness, respectively, as shown in Table 2.

4.3 Cluster Analysis

The call cluster analysis [9] includes a number of different algorithms and methods for combination objects of similar kind into each category. In other words, the cluster analysis is an inquiry data analysis tool which purpose is to sort different objects into groups in a way that the degree of proximity between two objects is maximal if they belong to the same group (http://www.statsoft.com/textbook/stcluan.html, 2007/6/14).

A non-hierarchical method to forming good clusters is to indicate a coveted number of clusters that is said k, then appoint each case to one of k clusters to the effect that minimize a measure of dispersion inside the clusters (http:// www.resample.com/ xlminer/help/kMClst/KMClust_intro.htm, 2007/6/14). Brief and to the point, k-means clustering is a statistic to classify or to group your cases based on sum of squared Euclidean distances from the meaning of each cluster into K number of groups. When

using the k-means method specifies positive integer number k, therefore authors decide number k to use Ward's method.

A two-stage clustering approach cluster analysis is to use classified subjects into mutually incompatible groups on the strength of the Ward method using the K-means clustering procedure. This cluster method includes both the Ward method and the K-means clustering, with the first step of all phases finding the maximum in total coefficient variation increments to look on as cluster numbers. The second step k-means clustering using the first step results indicate significant differences.

The results of the cluster analysis indicate that 218 phase to 217 phase find the maximum in total coefficient variation increments (increment=610.887) showing appropriate use of two cluster groups, as shown in Table 3. As shown in Table 4, k-means clustering uses two groups. Cluster I final class centroid of this instrument is 5.41, 4.57, 6.22, 5.81 and 5.23 for perceived anxiety, perceived liking, perceived self-efficacy, ease of use and usefulness, respectively. Cluster II final class centroid of this instrument is 4.85, 3.53, 4.20, 3.96 and 4.03 for perceived anxiety, perceived liking, perceived self-efficacy, ease of use and usefulness, respectively, as shown in Table 4. Both the two cluster groups, cluster I appeared to have the highest means score on " perceived self-efficacy" and cluster II highest means score on " perceived anxiety." Based on the means score characteristics with respect to the two factors, cluster I was labeled as "high pleasure" and cluster II was labeled as "low pleasure" to use the VoIP as a survey tool.

The Ward method achieves an easy look for the k value in the k-means clustering. Consequently, Cluster I's best factor is perceived as self-efficiency, Cluster II's best factor is perceived as anxiety. Table 4 reveals high pleasure with regards to perceived self-efficiency and low pleasure care about perceived anxiety.

5 Identifying Cluster Results

5.1 Discriminant Analysis

Discriminant analysis is a technique used to build a predictive model and classifying set of observations based on observed characteristics of each case into predefined classes. Fisher's discriminant analysis refers to "the use of multiple measures in taxonomic problems" (Fisher, 1936). In order to validate the results of pleasure degree for a VoIP survey, a discriminant analysis was performed with two cluster groups and five factors.

The selected Discriminant for the value of Wilks' lamda that is a kind of inverse measure, values of lambda which are near zero signify high discrimination between groups. Wilks' Lambda value is equal to 0.342 to indicate efficacious discrimination between groups. The results of the study reveal that for the "high pleasure" Fisher's linear discriminant function is Y high pleasure = -59.501 + 4.024 * Perceived anxiety + 2.568 * Perceived liking + 4.954 * Perceived self-efficacy + 3.447 * Ease of use + 6.051 * usefulness. "Low pleasure" Fisher's linear discriminant function is Y low pleasure = -34.414 + 3.566 * Perceived anxiety + 1.938 * Perceived liking + 3.466 * Perceived self-efficacy + 2.471 * Ease of use + 4.481 * usefulness.

Specifically, the testing of equality of groups means it reveals all five factors significant differences ($p<0.001$) the two pleasure degrees for the VoIP survey factor clusters. Classification results are indicated at 97.3% of the original grouped cases correctly classified and 97.3% of cross-validated grouped cases correctly classified.

5.2 Classification Results

The results of the independent-samples T test reveal that cluster I is strongly motivated by perceived self-efficiency, cluster II is strongly motivated by perceived anxiety. As show in table 6, the primary factor is perceived self-efficacy (means value=5.3582, $p<0.000$). Finally, the two clusters show different attitudes toward the VoIP survey.

Table 1. Results of independent-samples T test

Variable	Mean			T value
	Cluster I (high pleasure)	Cluster II (low pleasure)	total	
perceived anxiety	5.4314	**4.8844**	5.1781	3.280*
perceived liking	4.6991	3.5047	4.1135	8.023**
perceived self-efficacy	**6.2271**	4.4057	**5.3582**	11.731**
easy to use	5.8289	4.1258	5.0088	12.309**
usefulness	5.4115	3.9658	4.7118	11.567**
*p<0.001, **p<0.000				

6 Discussions and Conclusions

One interesting finding is that, both self-efficiency and perceived anxiety are the main factor to the impact users' attitudes toward the VoIP survey as shown in table 6. Our result implies that users have high self-efficiency will have a high need to adopt the VoIP as a survey tool. Thus, if users have a training course to improve their ability to use a VoIP, users will reduce their anxiety to rejecting the VoIP as a survey tool.

The results show that the k-means clustering is able to detect relationships and patterns in the VoIP survey attitudes. Results of k-means clustering were both high pleasure and low pleasure. The discriminating reasons were visualized by the k-means clustering with an understandable factor shown in Table 4. The foregoing, comparisons of the two clusters via a research worker enable the VoIP survey to obtain an overview of the understanding of users' attitudes toward it.

In summary, the contribution of this paper has confirmed the VoIP as innovative method as a survey method. In order to confirm the generalizability of this result to other complex technologies, further research into the utility of the VoIP on different types of dimensions will be beneficial.

References

1. Ajzen, I., Fishbein, M.: Belief, Attitude, Intention, and Behavior: An Introduction to Theory and Research. Addison-Wesley, Reading (1975)
2. Ajzen, I., Fishbein, M.: Understanding Attitudes and Predicting Social Behavior. Prentice-Hall, Inc., Englewood Cliffs (1980)
3. Davis, F.D.: Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly 13(3), 319–340 (1989)
4. Davis, F.D.: User acceptance of information technology: System characteristics, user perception and behavioral impacts. International Journal of Man–Machine studies 38, 475–487 (1994)
5. Huang, H.M., Liaw, S.S.: Exploring users attitudes and intentions toward the web as a survey tool. Computers in Human Behavior 21, 729–743 (2005)
6. Huang, H.M.: Do print and Web surveys provide the same results? Computers in Human Behavior (2004)
7. Jieun, Y., Imsook, H., Munkee, C., Jaejeung, R.: Extending the TAM for a t-commerce. Information & Management 42, 965–976 (2005)
8. Tabachnick, B.G., Fidell, L.S.: Using Multivariate Statistics, 4th edn. Allyn & Bacon, Needham Heights (2001)
9. Tyron, R.C.: Cluster Analysis. Edwards Brothers, Ann Arbor (1939)
10. Wen, S.L.: Research on the relationship of cellular phone customer satisfaction and loyalty- senior high (vocational) school student within Taipei area as an example, Graduate Institute of Business Management, Tatung University (2003)

Privacy-Preserving Collaborative E-Voting

Gary Blosser and Justin Zhan

SION Lab, Carnegie Mellon University
{gblosser,justinzh}@andrew.cmu.edu

Abstract. We propose a cooperative multi-party system where unrelated entities can both both vote and change their vote with continual accurate tabulation results while retaining vote privacy. The voting protocol presented uses homomorphic encryption, digital envelopes, Hamming weights, multiple parties to fulfill the needs of an e-voting system. The protocol is presented and limitations are explored.

Keywords: Privacy, Security, E-Voting.

1 Introduction

E-voting is an important topic with headlines being made as problems are discovered and published [6]. Many schemes have been proposed to solve the privacy and cryptographical problems inherit to e-voting. Traditionally these schemes are based upon centralized servers [9], dual-layered encryptions [5], or probability [11]. In this paper, we propose a cooperative multi-party e-voting scheme where unrelated entities can both vote and change their vote with continual accurate tabulation results while retaining vote privacy.

As with all e-voting the original physical voting system and its requirements define the needs and requirements of the electronic counterpart. Modern voting is concerned with three main points and a successful e-voting system needs to mirror or even enhance these points [6]. First, an individuals vote should only be counted once, either by preventing additional votes or, as is in our proposed case, causing the second vote to replace the first. Second, the tabulated result must be accurate. While most systems are focused upon this aspect some systems, such as the method proposed in [11], rely upon probability and statistics to give a statistically accurate count. After the recount issues in several recent American elections a 'statistically' accurate count would be found problematic when 100% verifiable accuracy is required. Third, each voters vote must be kept private. In homomorphic schemes like [5] and this proposed scheme, privacy is kept through the use of homomorphic encryption. However, the homomorphic cryptosystem and methods are quite different between [5] and this paper. Homomorphic encryption allows basic mathematical functions to be applied to encrypted data and is explained in more detail in the next section. In the probabilistic scheme proposed in [11] privacy is preserved through the use of masking the true vote via probability, preventing an attacker from knowing if the vote was truly the given value or had been masked.

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 508–513, 2008.

There are three main *contributions* with this proposed scheme. First, most homomorphic schemes depend upon the Pallier [7] homomorphic cryptosystem while we rely upon the cryptosystem proposed by Goldwasser and Micali [3]. Secondly, the reviewed schemes focus upon preventing a second vote while we allow an individual to vote and change their vote as often and many times as they wish without any negative effect to the final tabulation process. This allows an individual to change their mind, or even alter their vote if the first vote was induced by coercion. Lastly, our scheme is distributed amongst participants and does not rely upon a central trusted server as a point of failure, making this a truly social e-voting protocol.

2 Approach

We will first provide the building blocks which contain our definition of privacy, homomorphic encryption description, digital envelope technique, explanation of Hamming weights and the attack models that we consider.

2.1 Building Blocks

We will define privacy as given in [10], the basic idea is as follows: A privacy-oriented scheme S preserves data privacy if for any private Voting data V, the following is held:

$$|Pr(V|S) - Pr(V)| \leq \epsilon$$

where

- S: The privacy-preserving multi-Party reversible e-voting Scheme.
- V: Private voting data that needs to be preserved.
- ϵ: A probability parameter.
- $Pr(V|S)$: The probability of revealing the voting data V once the e-voting scheme is applied.
- $Pr(V)$: The probability of revealing voting data V without any privacy preserving method being applied.
- $Pr(V|S) - Pr(V)$: The probability of revealing voting data V both with and without the privacy-preserving e-voting scheme being applied.

Therefore, $1 - \epsilon$ represents the total privacy level that the privacy-oriented scheme S can achieve.

To use this defined measurement for e-voting privacy, we need to reduce the entirety of the e-voting algorithm to a set of privacy-oriented protocols. We can infer that the privacy-preserving e-voting scheme preserves privacy if each component part, as well as their combination, preserves privacy. A similar idea and the associated proof can be found in [2].

Theorem 1. *Suppose that g is privately reducible to f and that there exists a protocol for privately computing f. Then there exists a protocol for privately computing g.*

The formal definition for component protocol privacy is as follows:

Definition 1. *A privacy-oriented component protocol C preserves data privacy if for any private data T, the following is held:*

$$|Pr(T|C) - Pr(T)| \leq \epsilon$$

where

- *C: Component protocol.*
- *$Pr(T|C)$: The probability of revealing private data T once a privacy-preserving component protocol is applied.*
- *$Pr(T|C) - Pr(T)$: The probability of revealing private data T both with and without applying the privacy-preserving component protocol.*

In this case, $1 - \epsilon$ is the privacy level that the privacy-oriented component protocol C can achieve.

Now we will examine the fundamental building blocks of homomorphic encryption and Hamming weight. Information on the digital envelope and attack model building blocks used within this paper can be found in [1] and [10], respectively.

Homomorphic Encryption. While many homomorphic encryption schemes have been proposed the original concept was proposed in [8]. Our protocols are based upon the scheme proposed by Goldwasser and Micali in [3] which is semantically secure.

A crypto-system is homomorphic with respect to some operation $*$ on the message space if there is a corresponding operation $*'$ on the cipher-text space such that $e(m) *' e(m') = e(m * m')$. In our privacy-oriented protocols, we use the exclusive-disjunction ($\oplus$) homomorphism offered by [3] in which Goldwasser and Micali proposed a trapdoor mechanism based on the idea that it is hard to factor number $n = pq$ where p and q are two large prime numbers.

In our scheme, we utilize the following property of the homomorphic encryption functions: $e(m_1) \times e(m_2) = e(m_1 \oplus m_2)$ where m_1 and m_2 are the data to be encrypted. Because of the property of associativity, $e(m_1 \oplus m_2 \oplus \cdots \oplus m_n)$ can be computed as $e(m_1) \times e(m_2) \times \cdots \times e(m_n)$ where $e(m_i) \neq 0$. That is

$$d(e(m_1 \oplus m_2 \oplus \cdots \oplus m_n)) = d(e(m_1) \times e(m_2) \times \cdots \times e(m_n)) \tag{1}$$

Where the truth table for exclusive disjunction is

Table 1. Exclusive-Disjunction Truth Table

$\oplus$	1	0
1	0	1
0	1	0

Which allows reversibility due to

$$d(e(m_1 \oplus m_2 \oplus m_2)) = d(e(m_1)) \tag{2}$$

Hamming Weight. A Hamming weight [4] is the distance that a string or array of bits is from being all zeros. In this paper, we use Hamming weight to determine the total number of yes votes after tabulation. As examples, the Hamming weight of 0010000 is 1, 1011101 is 5, and 0000000 is 0.

2.2 Privacy-Preserving Collaborative Reversible E-Voting Framework

E-voting is the great-grandchild of the original yes or no question. Through time verbal systems have been replaced with paper, and paper with electronic systems. However, at their most fundamental level all voting systems are still just a yes and no vote, be it yes for one candidate and no for all other or just yes or no on a law.

Problem 1. An interesting research problem is as follows: Given three or more parties, we need to conduct a vote. Due to privacy issues, revealing the vote of any party is not allowed, but every party may alter their vote as often as wanted. The issue is how to tabulate each parties vote in a way that does not compromise data privacy for any party and still allows for the votes to be changed. The objective is to conduct the vote using multiple parties without violating privacy, but to still obtain an accurate result as in physical systems.

Privacy-Preserving Multi-Party Voting Protocol

Problem 2. Let us assume P_1 has a private bit $c.count_1$, P_2 has a private bit $c.count_2$, $\cdots$, and P_n has a private bit $c.count_n$ where $n \geq 3$. The goal is to compute the $\sum_{i=1}^{n} c.count_i$ without compromising data privacy. Furthermore, any party may change their bit at any time during the protocol execution. One party obtains $\sum_{i=1}^{n} c.count_i$, then shares the result with other parties.

Highlight of Protocol 1: In our protocol, we randomly select a key generator, e.g., P_k and a separate tabulator, e.g., P_t. P_k generates a cryptographic key pair (e, d) of a homomorphic encryption scheme and a set of at least i unique bit arrays B where i is the total number n of participants P_n. P_n then sends a random public-key encrypted bit array to each participant. P_k then sends the public key to P_t. P_t uses the public key to encrypt a random bit array of length i to use as a digital envelope. P_1 brings their bit array to a power to signify their vote where an odd power is a no and an even power is a yes; P_1 then sends their vote to P_t for tabulation and may send votes as often as they wish where the total power determine the yes and no; P_t multiplies the received vote into the digital envelope; Repeat until P_t obtains votes from all participants P_n. P_t then sends the digital envelope to P_k who decrypts it and returns the decrypted envelope to P_t. Finally, P_t obtains $c.count$ by using an exclusive-disjunction to remove the random bit array from the envelope then calculating the Hamming weight [4].

Limitations of Protocol 1: Due to the binary nature of the protocol the length of communications and time complexity scale rapidly with greater numbers of

participants. However, this limitation can be minimized by breaking the participants into random groupings and summing the resulting tabulation. To further preserve privacy the summing can be done with another homomorphic scheme, like [7] which provides additive homomorphism.

We present the formal protocol as follows:

Protocol 1

1. P_k generates a cryptographic key pair (e, d) of a semantically secure homomorphic encryption scheme. P_k also generates a set of n unique bit arrays b of the same length in which there is only a single yes bit, e.g. 00010000. P_k distributes a unique bit array to each participant P_n and transmits the public key to P_t.
2. P_t generates and encrypts a random bit array $e(b_{de})$ of length n, e.g. 11001010.
3. P_1 brings their bit array b_1 to a random even (for no) or odd (for yes) power p. With even computing $e(b_1)^2 = e(b_1 \oplus b_1) = e(\neg b_1)$ and odd computing $e(b_1)^3 = e(b_1 \oplus b_1 \oplus b_1) = e(b_1)$. The resulting value $e(c.count_1)$ is sent to P_t. P_t computes $e(b_{de}) \times e(c.count_1) = e(b_{de} \oplus c.count_1)$.
4. P_2 brings their bit array b_2 to a random even (for no) or odd (for yes) power p. With even computing $e(b_2)^2 = e(b_2 \oplus b_2) = e(\not{0}_2)$ and odd computing $e(b_2)^3 = e(b_2 \oplus b_2 \oplus b_2) = e(b_2)$. The resulting value $e(c.count_2)$ is sent to P_t. P_t computes $e(b_{de}) \times e(c.count_1) \times e(c.count_2) = e(b_{de} \oplus c.count_1 \oplus c.count_2)$.
5. Repeat until P_n sends $e(c.count_n)$ to P_t who computes $e(b_{de}) \times e(c.count_1) \times e(c.count_2) \times \cdots \times e(c.count_n) = e(b_{de} \oplus c.count_1 \oplus c.count_2 \oplus \cdots \oplus c.count_n)$.
6. P_t sends $e(b_{de} \oplus c.count_1 \oplus c.count_2 \oplus \cdots \oplus c.count_n)$ to P_k who decrypts it returning $(b_{de} \oplus c.count_1 \oplus c.count_2 \oplus \cdots \oplus c.count_n)$ to P_t.
7. P_t computes $b_{de} \oplus (b_{de} \oplus c.count_1 \oplus c.count_2 \oplus \cdots \oplus c.count_n) = c.count_1 \oplus c.count_2 \oplus \cdots \oplus c.count_n \Rightarrow \sum_{i=1}^{n} (c.count_i > 0)$

3 Conclusion

We have presented an e-voting scheme that allows for participants to cooperatively vote and change their vote while maintaining privacy. However, as noted before the nature of the scheme limits the number of participants due to increasing computational and time complexity preventing use in large-scale elections unless the participant group is broken into workable size groups.

In the future, an experiment to determine the actual time complexity of using both Goldwasser and Micali's homomorphic algorithm and the proposed e-voting scheme needs to be undertaken. The privacy and security effects of breaking a larger population into pieces and using Pallier homomorphic encryption to bind the results together may also be a topic of study.

References

1. Chaum, D.: Security without identification. Communication of the ACM 28, 1030–1044 (1985)
2. Goldreich, O.: The Foundations of Cryptography. Cambridge University Press, Cambridge (2004)

3. Goldwasser, S., Micali, S.: Probabilistic encryption and how to play mental poker keeping secret all partial information. In: Proceedings of the fourteenth annual ACM symposium on Theory of computing, pp. 365–377 (1982)
4. Hamming, R.W.: Error detecting and error correcting codes. Bell Syst. Tech. J. 29(2), 147–160 (1950)
5. Her, Y.S., Imamoto, K., Sakurai, K.: E-voting System with Ballot-Cancellation Based on Double-Encryption
6. Hisamitsu, H., Takeda, K.: The Security Analysis of e-Voting in Japan. LNCS, vol. 4896, p. 99. Springer, Heidelberg (2008)
7. Paillier, P.: Public-key cryptosystems based on composite degree residuosity classes. In: Stern, J. (ed.) EUROCRYPT 1999. LNCS, vol. 1592, pp. 223–238. Springer, Heidelberg (1999)
8. Rivest, R., Adleman, L., Dertouzos, M.: On data banks and privacy homomorphisms. In: DeMillo, R.A., et al. (eds.) Foundations of Secure Computation, pp. 169–179. Academic Press, London (1978)
9. Rjaskova, Z.: Electronic voting schemes. Unpublished masters thesis, Comenius University, Bratislava (2003)
10. Zhan, Z.: Privacy Preserving Collaborative Data Mining. PhD thesis, University of Ottawa (2006)
11. Zhan, Z., Matwin, S., Chang, L.W.: Privacy-Preserving Electronic Voting. Information & Security 15(2), 165–180 (2004)

Privacy-Aware Access Control through Negotiation in Daily Life Service

Hyun-A Park[1], Justin Zhan[2], and Dong Hoon Lee[1]

[1] CIST (Center for Information Security Technologies), Korea University Anam Dong, Sungbuk Gu, Seoul, Korea
{kokokzi,donghlee}@korea.ac.kr
[2] SION Lab, Carnegie Mellon University,
justinzh@andrew.cmu.edu

Abstract. As users are participating in various social contexts, some projects such as MobiLife and MyLifeBits are developing a facility to commence daily life service. This enables users to store all their daily events which can be collected using their mobile device. These data can be shared with other people or some service providers only if the user agrees. However, in the cases of inter-domain web service usage or sharing their data with others, there are some potential problems about privacy. To solve these problems, we propose a new method, privacy-aware access control through negotiation process (N-PAC). This method enables a user to accomplish self-determination and self-control of personal information in more realistic application environments.

Keywords: Privacy, personal information, self-determination, self-control, negotiation, encryption, daily life service, MobiLife, MyLifeBits.

1 Introduction

The Ubiquitous era, which is expressed as "5 any - anyone, anytime, anywhere, anynetwork, anydevice", is forthcoming. However, the mass-market-scale ubiquitous services and applications are still problems which we have to figure out. To find the solutions for these problems, some projects such as MyLifeBits and MobiLife are researching daily life service.

As users are participating in various social contexts in daily life, MyLifeBits and MobiLife try to develop a facility to maintain such relations to communicate, to share items and to manage today's complex lifestyles. This enables users to store all their daily events which can be collected using their mobile phone, for example, SMS, photos, call, movie, e-commerce information, web service log and usage information, location information, documents, media, battery charge, personal schedule, and so on. These stored data are transferred through the internet to each user's personal database, and stored and managed as a personal history with the passage of time. It is possible to share their data with other people or some service providers only if the user wants. However, without users' consent, the data stored in mobile phone database or personal database can be abused or misused by a server manager or unauthorized accesses.

Related works and contributions. The work for privacy preserving techniques in data management system has been researched in a variety of directions. P3P(Platform for Privacy Preferences), was developed by W3C(World Wide Web Consortium), provides a

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 514–519, 2008.

way for a web site to encode its data-collection practices in a XML format known as a P3P policy [17]. A Hippocratic database uses privacy metadata which consist of privacy policies and privacy authorizations stored in two tables. The policies and authorizations associate with each attribute and each user the usage purpose(s) [1].

In purpose-based access control by Byun et al. [4, 5], they proposed an access control model for privacy protection based on the notion of purpose. However, purpose management introduces a great deal of complexity at the access control level. In another aspect, [16] introduced an alternative privacy access control mechanism that is not based on purpose. It defined the intended purpose of personal information as a chain of acts on this type of information. Mun et al. [14] provided policy-based access control mechanism for the personal information directory system.

As for privacy policy negotiation, Hatakeyama and Gomi [11] introduced a means for providers to be able to confirm privacy policies before an attributes exchange takes place and to determine what kinds of attributes to exchange and how to manage these attributes. In other papers related to privacy policy negotiation, [8] extends the previous session level data privacy methods by adding transaction level data privacy.

In almost all of the previous papers, access control rules are setup based on purpose, intent, or policy at first time. However, the data for some goal can be used for other purpose. They cannot cover all the cases and had many difficulties for users to make a self-determination or self-control of their personal data. Even if it has negotiation or dynamic process, it is too abstract and conceptual.

In this paper, we focus on more realistic application, daily life service. Based on this service, we accomplish privacy-aware access control by adding negotiation protocol and encrypting data under the classified level. We introduce PAAC (privacy aware access controller) as a kind of TTP so that all the accesses to user's information should pass through PAAC. Negotiation is also processed by PAAC.

Moreover, these properties enable our proposed method N-PAC (Privacy-Aware Access Control through Negotiation) to make self- determination and self-control of personal data.

2 The Construction of N-PAC (Privacy-Aware Access Control through Negotiation) Introduction

2.1 Application Scenarios

In this section, we address our application scenarios, daily life services such as MobiLife or MyLifebits projects [18, 19]. We consider the privacy problems caused by these services. Especially, in this paper, we model our solution using MobiLife service as one of our application scenarios.

2.2 The Components of N-PAC

N-PAC has four main parties as follows;.

- **MA (Mobile user agent).** Instead of a user, a mobile user agent does everything related to daily life using their mobile phone.
- **PAAC (Privacy-Aware Access Controller).** PACC is a kind of TTP (Third Trust Party) and protects users' privacy. This manages users' secret keys, implements encryption and decryption of users' data, registers users' privacy policies, and evaluates and negotiates users' privacy policies with service providers'/ PIR ' policies.

▪**SP (Service provider)/ PIR (personal information requestor).** This party requires disclosing of users' data for enabling web services or sharing users' data. It negotiates privacy polices with PACC.
▪**PDM (Personal database manager).** As a kind of semi-trusted party, PDM just manages personal database and implements PAAC's requests.

2.3 The Processes of N-PAC

2.3.1 Privacy Policy Registration

We classify all data into 4-levels;

- Level 1: Not sensitive data. These data don't need to be encrypted.
- Level 2: Sensitive data. These data need to be encrypted but not to be negotiated by PAAC.
- Level 3: High sensitive data. These data need to be encrypted and negotiated by PAAC.
- Level 4: Top secret. These data should be encrypted only by user's secret key so that only the user can request and decrypt them. These data should not be negotiated and disclosed.

All mobile users register their privacy policies on each service in advance to PAAC. A user can determine which attributes can be disclosed. For this, one of the above 4-levels is assigned to all attributes of each table.
But, this process is provided as option. It means that there are two choices. One is that a user can choose what he wants by himself. The other is that a user can follow in advance setup classification under EU guideline or some other privacy rules. It can be updated and added as needed.

PAAC should keep each user's policy table like Table 1, where T_i means the table for service i and A_{ij} means j-th attribute of service i. If each table T_i has its attributes A_{ij}, it is expressed as: $T_i=(A_{i1}, A_{i2}, A_{i3,...}, A_{ij})$.

Table 1. Policy Table of User_1

	Level 1	Level 2	Level 3	Level 4
T_1 (service 1)	A_{11} A_{12}, A_{16}	A_{13}, A_{15}	A_{14}	A_{17}
T_2 (service 2)	A_{23}, A_{26}	A_{21}	A_{22}, A_{25}	A_{24}
……..	……..	……..	……..	……..

2.3.2 Data Processing

According to a user's policy, MA encrypts the data belonging to level 2, 3, and 4. In N-PAC, there are two kinds of keys, K_2 and K_3. K_2 is the shared key between a user and PACC in advance. K_3 is the key generated by a user. We explain the encryption method as one example with T_1 (service 1) in Table 1. E means efficient encryption function and D means efficient decryption function.

- Level 1: Let these data be in plaintexts.
- Level 2: $E_{k2}(A_{13})$, $E_{k2}(A_{15})$
- Level 3: $E_{k2}(E_{k3}(A_{14}))$
- Level 4: $E_{k3}(A_{17})$

The data stored in MDB are transferred to PDB through wireless network once a day or by certain period. All data of a user are stored in PDB (personal database) time by time, and day by day, i.e., according to the passage of time.

2.3.3 Privacy Policy Evaluation

When a user wants to be provided with something from different domain service or a PIR requests to share with the user's data, at first PAAC evaluates a user's privacy policies with SP or PIR's privacy policies. For more simple explanation, we consider just the case of SP. The following 3 conditions can be expected;

1. $\bigcup P_{SP\_T_i} \subseteq P_{MA\_T_i\_L_1} \cup P_{MA\_T_i\_L_2}$

If the condition satisfies with this equation, then PAAC implements Action Process, where P_{SP_Ti} is service provider's privacy policy on service table T_i. $\bigcup P_{SP\_T_i}$ means all the attributes which a SP wants to be disclosed on the user's data T_i. P_{MA} is mobile user agent's privacy policy and $P_{MA\_T_i\_L_4}$ means the attributes which a mobile user agent wants to be in plaintexts on service table Ti. Therefore, the above equation means that all the data which a service provider wants to share with on service table T_i are included to the attributes which can be disclosed without negotiation process. This time, in Action Process, PACC needs to decrypt the attributes of L_2 and replies with the attributes of L_1 and the decrypted attributes of L_2.

From next, the conditions are $\bigcup P_{SP\_T_i} \not\subset P_{MA\_T_i\_L_1} \cup P_{MA\_T_i\_L_2}$.

2. $P_{SP\_T_i} \subseteq P_{MA\_T_i\_L_3}$

This condition is that some attributes which a SP wants to know belong to the user's attributes of *Level 3*. Then, PAAC starts to implement Negotiation Process.

3. $P_{SP\_T_i} \subseteq P_{MA\_T_i\_L_4}$

This condition is that some attributes which a SP wants to know belong to the user's attributes of *Level 4*. Then, PAAC sends the message that the attributes cannot be disclosed, to the SP. If the SP accepts this message under the user's policy, PAAC starts Action Process. If not, the access trial is over.

2.3.4 Privacy Policy Negotiation

If the privacy policy evaluation satisfies with condition 2, PAAC gets to start Privacy Policy Negotiation Process.

1. PACC sends the message that the attributes cannot be disclosed to the SP.
2. If the SP accepts this acknowledgement under the user's policy, PAAC starts Action Process.
3. If the SP rejects this, then PAAC sends this message that the SP wants to know the attributes of *Level 3* for a service T_i to the MA.
4. If the MA rejects this acknowledgement under the SP's policy, the access trial is over.
5. If the MA accepts, MA replies to PAAC as "Yes".
6. Then, PAAC requests the attributes to PDM (personal database management) and decrypts the received data of *Level 3* from PDM with the shared key K_2; $D_{k2}(E_{k2}(E_{k3}(A_4)))= (E_{k3}(A_4))$
7. PAAC resends the half decrypted data $E_{k3}(A_4)$ to MA.

8. MA decrypts the data again with K_3 and sends it to PAAC; $D_{k3}(E_{k3}(A_4))= A_4$
9. PAAC starts Action Process.

2.2.5 Action

In this process, PACC implements the final results.

In the case of condition 1, PAAC requests the attributes of $\bigcup P_{SP\_T_i}$ to PDM and decrypts the attributes of L_2; $D_{k2}(E_{k2}(A_3))= A_3$, $D_{k2}(E_{k2}(A_5)) = A_5$. Then, PAAC provides the SP with the user's data which belong to *Level 1* and *2* and the user can be offered something from the SP.

In the case that SP accepts a user's policy in condition 3 or after negotiation process, PAAC provides the SP with the finally selected attributes. Thereafter, the user can be offered something from the SP.

3 Discussion and Conclusion

According to [6], privacy involves the right to control one's own personal information, and the ability to determine if and how one's information should be collected and used. As more and more of personal belongings are stored in persistent media, the right of informational self-control is merging with the right of self-determination to a new meaning of freedom. The crux of the matter is that the decision whether we would like to share personal information with others is up to us. If we do not want to reveal personal data, we do not have to. If we wish to remain anonymous, we should be capable of doing so [21].

One of the most important properties of our method N-PAC is self-determination and self-control of personal information. One of the 4 levels for each information is selected by the user himself. According to the classified level, the different encryption module is applied with 2 kinds of keys. But, the data of *level 1* don't need to be encrypted. Because of this different encryption module, even PAAC cannot negotiate or disclose on its own authority. Negotiation is only allowed to level 3 data. Only under user's consent, it is also possible to release data through negotiation because the data are encrypted twice with K_2 and K_3. K_3 is the key which the user only knows and PAAC does not know. The data of level 4 must not be disclosed by no means so they are encrypted with K_3.

Another specialty of this paper is PAAC as a kind of TTP. All accesses into users' data should pass through PAAC. PDM only has to manage and implement the requests from PAAC. PAAC is an intermediary between a user and a SP/PIR in negotiation process. This privacy policy negotiation process enables users to self-determine and self-control their personal information.

In this paper, we propose a new method to preserve users' privacy in the forthcoming daily life service. We classify all data into 4-level based on sensitiveness and encrypt them with different encryption module. All of these are done by a user himself. Based on the data classification, negotiation is processed by PAAC. Through these processes, users can accomplish self-determination and self-control of their information. Finally, N-PAC enables users' data protection to accomplish FIPs (fair information practices).

Acknowledgments. This work was supported by the Korea Research Foundation Grant funded by the Korean Government(MOEHRD)" (KRF-2007-612-D00132 (I00051)).

References

1. Agrawal, R., Kiernan, J., Srikant, R., Xu, Y.: Hippocratic databases. In: The 28th International Conference on Very Large Databases (VLDB) (2002)
2. Ardagna, C.A., Damiani, E., Cremonini, M., De Capitani di Vimercati, S., Samarati, P.: The architecture of a privacy-aware access control decision component. In: Barthe, G., Grégoire, B., Huisman, M., Lanet, J.-L. (eds.) CASSIS 2005. LNCS, vol. 3956. Springer, Heidelberg (2006)
3. Ashley, P., Hada, S., Powers, C., Schunter, M.: Enterprise Privacy Authorization Language (EPAL). IBM Research (2003)
4. Byun, J., Bertino, E., Li, N.: Purpose-based access control for privacy protection in relational database systems. Technical Report 2004-52, Purdue University (2004)
5. Byun, J., Bertino, E., Li, N.: Purpose based access control of complex data for privacy protection. In: Symposium on Access Control Models and Technologies Proceedings of the tenth ACM symposium on Access control models and technologies, pp. 102–110 (2005)
6. Cavoukian, A.: Genetic Privacy: the right "not to know", Notes for Remarksin 10th World Congress on Medical Law (1994)
7. Mont, M.C., Pearson, S., Bramhall, P.: An Adaptive Privacy Management System For Data Repositories, HPL-2004-211 (November 18, 2004)
8. Wu, C., Potdar, V., Chang, E.: A conceptual framework for privacy policy negotiation in web services. In: Furnell, S.M., Dowland, P.S. (eds.) Sixth International Network Conference (INC), pp. 195–202 (2006)
9. Eldin', Wagenaar, R.: IEEE International Conference on Towards users driven privacy control 2004, vol. 5, pp. 4673–4679 (2004)
10. Hommel, W.: An Architecture for Privacy-Aware Inter-domain Identity Management. In: Schönwälder, J., Serrat, J. (eds.) DSOM 2005. LNCS, vol. 3775, pp. 49–60. Springer, Heidelberg (2005)
11. Hatakeyama, M., Gomi, H.: Privacy Policy Negotiation Framework for Attribute Exchange. In: W3C Workshop on Languages for Privacy Policy Negotiation and Semantics-Driven Enforcement (2006)
12. El-Khatib, K.: A Privacy Negotiation Protocol for Web Services. In: Workshop on Collaboration Agents: Autonomous Agents for Collaborative Environments Halifax (2003)
13. LeFevre, K., Agrawal, R., Ercegovac, V., Ramakrishnan, R., Xu, Y., DeWitt, D.: Disclosure in Hippocratic databases. In: The 30th International Conference onVery Large Databases (VLDB) (August 2004)
14. Mun, H.J., Lee, K.M., Lee, S.H.: Person-Wise Privacy Level Access Control for Personal Information Directory Services. In: Sha, E., Han, S.-K., Xu, C.-Z., Kim, M.-H., Yang, L.T., Xiao, B. (eds.) EUC 2006. LNCS, vol. 4096, pp. 89–96. Springer, Heidelberg (2006)
15. Ni, Q., Lin, D., Bertino, E., Lobo, J.: Conditional Privacy-Aware Role Based Access Control. In: Biskup, J., López, J. (eds.) ESORICS 2007. LNCS, vol. 4734, pp. 72–89. Springer, Heidelberg (2007)
16. Al-Fedaghi, S.S.: Beyond Purpose-Based Privacy Access Control. In: Bailey, J., Fekete, A. (eds.) Australasian Database ADC 2007. CRPIT 63, Ballarat, Australia, pp. 23–32. ACS (2007)
17. P3P. The Platform for Privacy Preferences 1.0 (P3P1.0) Specification, The World Wide Web Consortium (2002) (April 16, 2002), `http://www.w3.org/p3p/`
18. http://www.istmobilife.org/index.php?option=com_content&task=view&id=41&Itemid=51
19. http://www.newsfactor.com/perl/story/20064.html
20. http://www.ist-mobilife.org/images/stories/architecture%20wp5.pdf
21. http://www.acm.org/crossroads/xrds11-2/spa_article.html

Author Index