Chapter 1

DEALING WITH TERABYTE DATA SETS IN DIGITAL INVESTIGATIONS

Nicole Beebe and Jan Clark

Abstract Investigators and analysts are increasingly experiencing large, even terabyte sized data sets when conducting digital investigations. State-of-the-art digital investigation tools and processes are efficiency constrained from both system and human perspectives, due to their continued reliance on overly simplistic data reduction and mining algorithms. The extension of data mining research to the digital forensic science discipline will have some or all of the following benefits: (i) reduced system and human processing time associated with data analysis; (ii) improved information quality associated with data analysis; and (iii) reduced monetary costs associated with digital investigations. This paper introduces data mining and reviews the limited extant literature pertaining to the application of data mining to digital investigations and forensics. Finally, it provides suggestions for applying data mining research to digital forensics.

Keywords: Digital forensics, digital investigation, large data sets, data mining

1. Introduction

The digital forensic discipline is experiencing heightened importance and attention at a time when data storage requirements are increasing exponentially. Enterprise storage needs are predicted to increase seven-fold between 2003 and 2006, and email volume is increasing at an annual rate of 29% [12]. Hinshaw [18] reports that data rates are doubling every nine months – twice as fast as Moore's Law. Because of this exponential growth, it is not uncommon for larger corporations and law enforcement agencies to face digital investigations with data sets as large, or larger than a terabyte [24, 28].

Current digital investigation tools cannot handle terabyte-sized data sets in an efficient manner [25]. Their overall efficiency is constrained

by the employment of simple hashing and indexing algorithms. Current digital forensics tools and processes are simply not scalable to large data sets [7, 14, 26]. Even with moderately large data sets (i.e., 200 gigabytes), data extraction and analytic activities become inordinately slow and inefficient. Processing times for limited keyword searches (10-20 keywords) can take days, and the human analyst is overwhelmed with the number of "hits" to review.

Digital investigations are also hindered by the limited processing capabilities of human analysts. As data sets increase in size, the amount of data required for examination and analysis also increases. This obviates the digital investigator's ability to meticulously review all keyword search "hits," files by file type, or all applicable system logs. It is therefore imperative that the digital investigation process be improved.

Currently, digital investigation processes and tools underutilize computer processing power through continued reliance on simplistic data reduction and mining algorithms. In the past, when human labor was cheap and computers were expensive, the analytical burden was shifted to analysts. For quite some time, the roles have been reversed, yet the digital forensics field has continued to levy the preponderance of its analytical burden on the human analyst.

Digital forensics is not the only discipline faced with the task of sifting through massive volumes of data. Other disciplines have employed data mining techniques to resolve this problem. However, little research has focused on applying these techniques to criminal forensics, and even less on digital forensics. The purpose of this paper is to increase awareness of data mining techniques within the digital forensic community and to show how they can be utilized to solve large data set challenges.

The following section provides a brief tutorial of data mining. Next, Section 3 surveys data mining research as applied to digital forensics. Section 4 suggests ways in which data mining techniques can be applied to digital forensics. This is followed by a discussion in Section 5 of the benefits and limitations of extending data mining to digital investigations. Concluding remarks are provided in Section 6.

2. Data Mining

Data mining embodies a multi-disciplinary approach to finding and retrieving information, and relies on several reference disciplines that enjoy long, rich research streams, including mathematics, statistics, computer science, and information science. Techniques developed and/or used within these disciplines (e.g., artificial intelligence, machine learning, pattern recognition, data visualization, and database processes) are

utilized to develop data models, identify patterns, detect anomalies, and retrieve information.

Data mining processes, methods and techniques can be divided into three major classes: descriptive modeling, predictive modeling and content retrieval. Descriptive modeling summarizes or discriminates data, whereas predictive modeling identifies characteristics that can help predict future observations. Both modeling techniques largely necessitate structured data (i.e., databases, XML documents and graphs). Content retrieval data mining is perhaps the most complex type of data mining, particularly because it extracts information from complex and/or semi-structured/unstructured data sets. Content retrieval techniques are typically directed toward text data, multimedia data (e.g., image, video and audio data), World Wide Web data, spatial data, time-series or sequential data and complex objects (containing more than one data type). Han and Kamber [15] and Hand, *et al.* [16] provide excellent discussions regarding each of these classes.

2.1 Descriptive Data Modeling

Descriptive data mining relies on data generalization and conceptualization to generate descriptions that facilitate both characterization (summarization) and comparison (discrimination). Characterization techniques tend to overlap with data warehouse techniques, which cleanse, transform and summarize large data sets into smaller data sets with aggregate level information. Since aggregation inherently results in data loss, characterization data mining techniques will likely have limited utility in digital forensic investigations, but may prove very helpful in non-forensic digital investigations, such as internal corporate investigations and military operations.

Comparison techniques, also known as discrimination, produce a set of rules for comparing the general features of objects found within two or more data collections. As a generic example, one may want to determine the general characteristics of customers who frequently make on-line purchases, compared with those who rarely make on-line purchases. The techniques used for data discrimination are very similar to those used for data characterization, except they include some form of comparative measure. The aggregation function associated with characterization techniques is not necessarily applicable to comparison techniques, thus comparison techniques may have greater potential in digital forensic investigations than characterization techniques.

2.2 Predictive Data Modeling

Predictive data mining builds on descriptive data mining, in that the first step is inherently descriptive. However, the ultimate goal is to anticipate and/or categorize future observations (data). There are three primary sub-classes of predictive data modeling: association rule-based data analysis, regression and classification. Association rule-based data mining attempts to identify relationships or "associations" between items or item features. Given a set of items, association analysis establishes rules that can predict the occurrence of an item, based on the occurrence of other items in the transaction. Thus, they capture the probability that two data items co-occur, facilitating the profiling of co-occurring items with high frequencies and probabilities. Similarly, anomalies can be detected by finding co-occurring items within the data set that have a very low probability of co-occurring.

Regression-based data mining is used when data observations (response or dependent variable) can be modeled, and therefore predicted, by a mathematical function using a given set of data characteristics (predictor or independent variables). The mathematical function may be linear or non-linear. For example, weight (Y) can be modeled as a linear function of height $(X) : Y = \alpha + \beta X$.

Classification techniques represent the third and final sub-class of predictive data mining techniques. While introduced last, these techniques are potentially the most relevant to the present discussion. Classification techniques are used for both descriptive and predictive data mining – the primary difference being the purpose for which they are employed. Classification data mining techniques employ a whole host of methods to classify data, including: decision tree induction, Bayesian classification/belief networks, neural networks, nearest neighbor classification, genetic algorithms, case-based reasoning, rough sets and fuzzy logic. Each of these methods offers different ways to develop classification schemes to describe data and subsequently classify future observations (data). Classification rules, decision trees and/or mathematical formulae may be generated to facilitate prediction.

Cluster analysis (or clustering) is a frequently used classification technique, and thus bears specific mention. All classification techniques employ a training (or learning) phase and a validation phase. During the training phase, data is analyzed to support the development of the classification scheme. The learning may be "supervised" or "unsupervised." In supervised learning, the class labels are predefined in the data set, and thus the classes as well as the number of classes are predefined. In unsupervised learning, nothing is predetermined – the classification

technique itself establishes classes and number of classes based on similarities and dissimilarities found in the data. Cluster analysis employs unsupervised learning via partitioning methods, hierarchical methods, density based methods, grid-based methods, and model-based methods. Cluster analysis is often used for anomaly and outlier detection, and is applicable to intrusion detection and fraud detection.

2.3 Content Retrieval Data Mining

The third major class of data mining techniques is content retrieval. Whereas descriptive and predictive data mining techniques largely leverage mathematical and statistical reference disciplines, content retrieval depends heavily on research in information science and computer science, particularly in the areas of information retrieval, artificial intelligence, machine learning and natural language processing. Content retrieval methodologies are geared toward retrieving content from unstructured or semi-structured data sets (e.g., a text document or set of images, as opposed to a structured database). The primary sub-classes of content retrieval are: information (text) retrieval, multimedia data mining, web mining, complex data object mining, spatial data mining and time-series/sequential data mining.

Text retrieval is often referred to as information retrieval, simply because information retrieval goals and objectives have historically been text related – it is only recently (relatively speaking) that the desire to mine other types of content has emerged. The goals of information retrieval are usually to compare documents, rank importance or relevance of documents, or find patterns/trends across multiple documents. Common information retrieval techniques fit into two major categories: keyword-based (similarity-based using terms) and indexing-based. Index-based approaches such as latent semantic indexing are more prevalent, due to current limitations of natural language processing algorithms,[1] and the inherent ability of indexing approaches to be more conceptually based.

Multimedia data mining techniques are particularly relevant to digital forensics in the realm of image retrieval. Images can be analyzed and retrieved using description-based retrieval systems that use keywords, captions, size, creation time, etc., or using content-based retrieval systems that use color histograms, wavelet transformations, or measures of texture, shape, or objects. In content-based retrieval systems, image representations are created using a variety of methods, including: (i) feature representation via abstract pixel data; (ii) 3-D color feature vectors spatially averaged over the entire image; (iii) k-dimensional color his-

tograms with subsequent partitioning using clustering algorithms; (iv) 3-D texture vectors using coarseness/scale, directionality and contrast; and (v) 20-dimensional shape feature vectors using area, circularity, eccentricity and axis orientation.

Using such techniques, images can be classified as containing humans, buildings, etc. and retrieved accordingly. Other content retrieval techniques are directed at web data, complex data objects, spatial data and time-series/sequential data. Each has potential application to digital forensics, such as the applicability of time-series/sequential data mining techniques to network log data in network forensics cases and the applicability of web mining to retrieve content, structure and usage data from World Wide Web (WWW) based data.

3. Data Mining and Digital Investigations

A basic understanding of data mining illuminates its potential application to digital investigations. Data mining techniques are specifically designed for large data sets – attempting to find and retrieve data and otherwise hidden information amongst voluminous amounts of data. The data may or may not be structured, noisy or from the same source. In digital forensics, data sources are both structured and unstructured; noisy and not noisy; and from both homogeneous and heterogeneous sources-particularly in large data set cases.

Large data set research, specifically research related to data mining, has yet to be extended to digital forensics to any appreciable degree. We argue that the lost potential associated with this continued research void is analogous to the Internet without *Google* (data indexing and querying), the credit card industry without fraud detection (data mining) algorithms, and Wal-Mart without its 500 terabyte data warehouse to facilitate customer relationship management (CRM) and retailer-supplier decision support. As we collectively strive to strengthen the science of digital forensic and investigations [22], it is imperative that researchers begin to leverage and extend large data set research – particularly in the area of data mining.

Large data set research has only been extended to digital forensics in a handful of instances. The following subsections describe the various data mining techniques employed to date.

3.1 Predictive Data Modeling

3.1.1 Classification via Cluster Analysis. de Vel, *et al.* [13] utilized a Support Vector Machine learning algorithm to mine e-mail content and positively identify its authorship from a set of exemplars

from known authors. This procedure is intuitively akin to handwriting analysis, although its accomplishment requires much more complex processing. A Support Vector Machine learning algorithm is a classification-based data mining algorithm that seeks to categorize data based on certain key features of the data. In this instance, categories refer to different authors. Distinguishing features of the e-mail content and headers are used to classify each e-mail in the proper category according to its author.

3.1.2 Classification via Discriminant Analysis. Carney and Rogers [6] demonstrated how stepwise discriminant analysis can be used to determine the probability of intentionality associated with downloading contraband images (i.e., child pornography). Their motivation was to provide a mechanism for event reconstruction with calculable accuracy probability to help investigators investigate the "Trojan defense."[2] They examined seven different characteristics (variables or features) of the data and empirically determined that a single model using two features (average difference between file creation times and median difference between file creation times) can be developed and used to ascertain user intentionality associated with the incidence of contraband stored on digital media.

3.1.3 Association Rule Mining. de Vel collaborated with Abraham [1] and Kling [2] to profile user behavior and identify behavioral irregularities using system activity logs. These researchers applied association rule data mining to network and system data to determine association rules based on system interaction history of the user. They developed activity-based and event-based association rules that related user role (e.g., system administrator vs. financial analyst) to typical system activities (e.g., scan the internal network vs. review company financial statements). In doing so, they were able to develop behavioral profiles of typical users, and thereby mine subsequent log data sets for anomalies (e.g., someone using the account of a financial analyst to scan the internal network).

3.1.4 Content Retrieval Incorporating Text Mining. Text mining (also referred to as "information retrieval") is an outgrowth of the information science discipline that has enjoyed several decades of research advances in computational linguistics, which facilitate text categorization, semantic extraction and content summarization [30]. Text mining has received expanded research attention in recent years due to significant increases in business intelligence demands and data avail-

ability [30]. Shannon [27] developed a text mining technique called the Forensic Relative Strength Scoring (FRSS), which is basically a data categorization algorithm applicable to data reduction and extraction activities. The FRSS uses a scoring system with two measures: ASCII proportionality and entropy score (ASCII data tends to have less entropy and non-ASCII data). Using FRSS, Shannon [27] demonstrated how these two measurers could be used for data reduction and extraction of ASCII data. Shannon recommended further research be conducted using N-gram-based text categorization techniques [8] to subsequently identify the language and category (email, news posting, technical document, etc.) of the extracted ASCII data.

3.1.5 Peripheral Data Mining Research. Other data mining research has been applied peripherally to digital forensics and digital investigations. A large research stream has developed regarding the application of data mining techniques to intrusion detection, with particular emphasis on anomaly detection versus signature-based intrusion detection. A few recent examples include research by Barbara, *et al.* [3], Mukkamala and Sung [21], and Stolfo, *et al.* [29].

Data mining techniques have also been extended to image analysis (counterfeit detection [23] and steganography detection [19]), as well as data visualization to facilitate link analysis. Finally, data mining techniques have been extended to crime data analysis to profile criminals and identify criminal networks [9–11, 17, 31, 32]. These extensions of data mining theory demonstrate the vast potential data mining has for the digital forensics and digital investigation disciplines.

4. Data Mining in Digital Investigations

We are urging greater investigation into the use of data mining techniques to aid digital investigations in two primary areas: crime detection and crime investigation. Descriptive, predictive and content retrieval data mining techniques can be loosely mapped to each of these areas. We purport that such mapping and application will result in: (i) reduced system and human processing time associated with data analysis; (ii) improved analytical effectiveness and information quality; and (iii) reduced monetary costs associated with digital investigations.

4.1 Crime Detection

Crime detection activities involve behavioral profiling and anomaly detection (both behavioral and technological). Descriptive modeling and predictive modeling techniques are thus applicable, while content

retrieval data mining techniques are not. Descriptive modeling incorporating characterization (summarization) techniques are applicable, for example, in determining conformability of a data set to Benford's Law [20], thereby supporting economic fraud detection upon analyzing electronic data. Characterization techniques could also be used to better focus limited investigative resources. Such techniques, for example, could show which computers were used more for certain types of activity, such as electronic communication.

Descriptive modeling incorporating comparison (discrimination) techniques can be used to determine the similarity of two non-identical objects (with non-matching hashes), such as images in steganography detection, or source code in intellectual property theft detection. Comparison techniques are also applicable when comparing user (account) data from network logs. Stark dissimilarity where none is expected might be indicative of unauthorized activity.

While the characterization and comparison descriptive modeling techniques described above are applicable to crime detection data mining, predictive modeling techniques, e.g., association-based rule mining and classification data mining, are more commonly applied. Association-based rule mining and classification data mining are inherently designed to describe similarities amongst data observations and occurrences. As a result, dissimilarities (or anomalies) clearly emerge. Anomaly detection based on association rule mining has obvious applications in the areas of network intrusion detection, fraud detection, and unauthorized use detection (i.e., in espionage cases), and have already been introduced.

4.2 Crime Investigation

Crime investigation activities map to the Data Analysis Phase of the digital investigations process. The Data Analysis Phase consists of three sub-phases: data surveying, data extraction, and data examination [4]. Data mining techniques can assist with all three sub-phases, and the applicability of data mining to the Data Analysis Phase may often result in a natural progression through the data mining classes (descriptive $\rightarrow$ predictive $\rightarrow$ content retrieval).

During the data survey sub-phase, descriptive (characterization) modeling techniques can be employed to profile use and activity (e.g., percent free space, percent free space wiped, percent ASCII vs. binary and percent data by file type). During the data extraction sub-phase, classification data mining techniques can be used to reduce the amount of data for analysis. For example, before information retrieval techniques are employed, a data set can be reduced to ASCII data. It is important to

emphasize that data reduction techniques should be classification-based as described, as opposed to descriptive characterization (summarization) techniques, due to data loss associated with the latter.

The applicability of data mining techniques to the data extraction sub-phase continues and overlaps with the data examination sub-phase. Key crime investigation activities include: entity extraction [11], content retrieval [15, 16, 30], and crime data analysis (or link analysis) [5, 17, 32]. Entity extraction refers to predictive modeling based classification techniques that are geared toward identification. Identification may be a function of person, user account, email account, authorship, personal characteristics, etc. This technique can be a useful mechanism for investigators seeking attribution.

Content retrieval has clear and extensive applicability to digital investigations, such as mining large data sets for text documents containing specific content or involving particular individuals, or mining large data sets for contraband graphic images (e.g., child pornography, counterfeit currency). Taking a closer look at the former example, the goal of text (information) retrieval is usually to compare documents, rank importance or relevance of documents, or find patterns/trends across multiple documents. Each of these goals is extensible to digital investigations – particularly the latter two. Ranking the importance or relevance of documents relative to investigative objectives, criminal allegations, or target content facilitates data extraction during the Data Analysis Phase and minimizes, as well as prioritizes, the "hits" an investigator or analyst has to review. This is critical when dealing with large data sets. Finding patterns and trends across multiple documents assists an investigator in profiling users and uncovering evidence for which exact keywords are unknown.

Content retrieval data mining is also of use in the areas of multimedia mining, web mining and time/spatial data mining. Multimedia mining is particularly useful from the standpoint of image detection involving contraband, counterfeit and steganographic images. Web mining classifies and retrieves content, structure and usage data from World Wide Web (WWW) based data. This technique could be very useful for digital forensic investigators. In addition to being voluminous, web data is exceptionally "noisy" relative to the investigative objectives at hand. The sheer volume and "noisiness" of this data is absolutely overwhelming and incompatible with manual data analysis techniques. Finally, time/spatial data mining may have applicability in chronology analyses in network and media investigations.

The last key crime investigation activity – crime data (link) analysis – demonstrates the data mining class progression that often occurs

during the Data Analysis Phase. The goal of crime data analysis is to identify and visualize associations amongst social (criminal) networks. Three primary steps are associated with crime data analysis: transformation, sub-group detection, and association/pattern visualization. During data transformation, data from disparate sources is converted via characterization based descriptive data mining techniques to develop a basic "concept space" for the data. The results are then fed into the sub-group detection step, wherein cluster analysis (predictive) data mining techniques are used to identify groups of criminal actors, e.g., those that communicate with one another. Associations between sub-groups are then identified, including network form (e.g., star shaped) and node (criminal actor) centrality/criticality. Predictive data mining techniques (social network analysis approaches, such as block-modeling) are employed to accomplish this [11]. Content retrieval techniques may or may not play a role in crime data (link) analysis, but the progressive application of multiple data mining classes and techniques is evident.

5. Discussion

Employing data mining techniques to aid digital forensic investigations has many potential advantages. If applied properly it meets the three-fold goal of (i) reducing system and human processing time; (ii) improving the effectiveness and quality of the data analysis; and (iii) reducing cost. Other benefits include better utilization of the available computing power, as well as improved ability to discover patterns/trends normally hidden to the human analyst. There are, however, potential limitations of employing data mining techniques to aid in digital forensic investigations. First and foremost, the techniques are virtually untested within the digital forensic discipline. Since data mining has enjoyed previous success, however, there is reason to believe it will be effective in this discipline. Second, there is a lack of understanding of data mining techniques by the digital forensics and digital investigations community.

Other limitations include: (i) evaluation of content retrieval algorithms is inherently subjective, thus different analysts may view success differently; (ii) data mining inherently converts data to a much higher level of abstraction, therefore uncertainty and error calculations are particularly relevant; and (iii) the application of data mining techniques may require advanced knowledge and training of analysts and investigators. We argue that these limitations do not limit the potential of extending data mining research to digital forensics and digital investigations. Instead, these limitations merely reflect the challenges associated with doing so.

6. Conclusions

Clearly, data mining techniques can support digital investigations in myriad ways. However, much work needs to be done before these techniques can be successfully applied and disseminated throughout the digital investigation community. Suggestions for promoting these techniques include: increasing the awareness and understanding of data mining techniques, training digital investigators in the use of these techniques, and creating a framework for using these techniques in digital investigations.

Originally, it was our intention to create such a framework, but it soon became clear that research and awareness of data mining techniques, as applied to digital investigations, is in its infancy. We therefore encourage other researchers and practitioners to assist us in improving awareness and skills in this area. Terabyte-sized data sets are already challenging analysts and investigators. Therefore, an active stream of research extending data mining research to digital forensics and digital investigations is desperately needed.

Notes

1. Natural language processing (NLP) techniques have trouble overcoming polysemy (multiple meanings for the same term) and synonmy (multiple terms with the same meaning).

2. The "Trojan defense" is a criminal defense that argues the defendant did not intentionally engage in the illegal activity, but rather that a Trojan, virus, or hacker was responsible for the illegal activity.

References

[1] T. Abraham and O. de Vel, Investigative profiling with computer forensic log data and association rules, *Proceedings of the IEEE International Conference on Data Mining*, pp. 11-18, 2002.

[2] T. Abraham, R. Kling and O. de Vel, Investigative profile analysis with computer forensic log data using attribute generalization, *Proceedings of the Fifteenth Australian Joint Conference on Artificial Intelligence*, 2002.

[3] D. Barbara, J. Couto, S. Jajodia and N. Wu, ADAM: A testbed for exploring the use of data mining in intrusion detection, *ACM SIGMOD Record*, vol 30(4), pp. 15-24, 2001.

[4] N. Beebe and J. Clark, A hierarchical objectives-based framework for the digital investigations process, to appear in *Digital Investigation*, 2005.

[5] D. Brown and S. Hagen, Data association methods with applications to law enforcement, *Decision Support Systems* vol. 34, p. 10, 2002.

[6] M. Carney and M. Rogers, The Trojan made me do it: A first step in statistical based computer forensics event reconstruction, *Digital Evidence*, vol. 2(4), p. 11, 2004.

[7] E. Casey, Network traffic as a source of evidence: Tool strengths, weaknesses and future needs, *Digital Investigation*, vol. 1, pp. 28-43, 2004.

[8] W. Cavnar and J. Trenkle, N-gram-based text categorization, *Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval*, pp. 161-175, 1994.

[9] M. Chau, J. Xu and H. Chen, Extracting meaningful entities from police narrative reports, *Proceedings of the National Conference for Digital Government Research*, pp. 271-275, 2002.

[10] H. Chen, W. Chung, Y. Qin, M. Chau, J. Xu, G. Wang, R. Zheng and H. Atabakhsh, Crime data mining: An overview and case studies, *Proceedings of the National Conference for Digital Government Research*, p. 4, 2003.

[11] H. Chen, W. Chung, J. Xu, G. Wang, Y. Qin and M. Chau, Crime data mining: A general framework and some examples, *IEEE Computer*, vol. 37(4), pp. 50-56, 2004.

[12] Connected Corporation, Storage reduction facts and figures (www.connected.com/downloads/Items_for_Downloads/Storage Facts_Figures.pdf).

[13] O. de Vel, A. Anderson, M Corney and G. Mohay, Mining e-mail content for author identification forensics, *ACM SIGMOD Record*, vol. 30(4), pp. 55-64, 2001.

[14] J. Giordano and C. Maciag, Cyber forensics: A military operations perspective, *Digital Evidence*, vol 1(2), pp. 1-13, 2002.

[15] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Academic Press, San Diego, California, p. 550, 2001.

[16] D. Hand, H. Mannila and P. Smyth, *Principles of Data Mining*, MIT Press, Cambridge, Massachusetts, 2001.

[17] R. Hauck, H. Atabakhsh, P. Ongvasith, H. Gupta and H. Chen, Using Coplink to analyze criminal justice data, *IEEE Computer*, vol. 35, pp. 30-37, March 2002.

[18] F. Hinshaw, Data warehouse appliances: Driving the business intelligence revolution, *DM Review Magazine*, September, 2004.

[19] J. Jackson, G. Gunsch, R. Claypoole and G. Lamont, Blind steganography detection using a computational immune system: A work in progress, *Digital Evidence*, vol. 1(4), pp. 1-19, 2003.

[20] G. Moore and C. Benjamin, Using Benford's Law for fraud detection, *Internal Auditing*, vol. 19(1), pp. 4-9, 2004.

[21] S. Mukkamala and A. Sung, Identifying significant features for network forensic analysis using artificial intelligence techniques, *Digital Evidence*, vol. 1(4), pp. 1-17, 2003.

[22] G. Palmer, A Road Map for Digital Forensics Research: Report from the First Digital Forensics Research Workshop, Technical Report DTR-T001-01 Final, Air Force Research Laboratory, Rome, New York, 2001.

[23] F. Petitcolas, R. Anderson and M. Kuhn, Information hiding: A survey, *Proceedings of the IEEE*, vol. 87(7), pp. 1062-1078, 1999.

[24] D. Radcliff, Inside the DoD's Crime Lab, *NetworkWorldFusion*, pp. 1-5, March 8, 2004.

[25] V. Roussev and G. Richard III, Breaking the performance wall: The cases for distributed digital forensics, *Proceedings of the Digital Forensics Research Workshop*, pp. 1-16, 2004.

[26] M. Schwartz, Cybercops need better tools, *Computerworld*, p. 1, July 31, 2000.

[27] M. Shannon, Forensics relative strength scoring: ASCII and entropy scoring, *Digital Evidence*, vol. 2(4), pp. 1-19, 2004.

[28] P. Sommer, The challenges of large computer evidence cases, *Digital Investigation*, vol. 1, pp. 16-17, 2004.

[29] S. Stolfo, W. Lee, P. Chan, W. Fan and E. Eskin, Data mining based intrusion detectors: An overview of the Columbia IDS Project, *ACM SIGMOD Record*, vol. 30(4), pp. 5-14, 2001.

[30] D. Sullivan, *Document Warehousing and Text Mining: Techniques for Improving Business Operations, Marketing and Sales*, John Wiley, New York, p. 542, 2001.

[31] G. Wang, H. Chen and H. Atabakhsh, Automatically detecting deceptive criminal identities, *Communications of the ACM*, vol. 47(3), pp. 71-76, 2004.

[32] J. Xu and H. Chen, Fighting organized crimes: Using shortest-path algorithms to identify associations in criminal networks, *Decision Support Systems*, vol. 38, pp. 473-487, 2004.