Research Journal of Applied Sciences, Engineering and Technology 7(10): 2027-2032, 2014
DOI:10.19026/ajfst.7.495
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: June 22, 2013 Accepted: July 08, 2013 Published: March 15, 2014

Research Article

Web Crime Mining by Means of Data Mining Techniques

[1]Javad Hosseinkhani, [1]Suhaimi Ibrahim, [1]Suriayati Chuprat and [2]Javid Hosseinkhani Naniz
[1]Advanced Informatics School (AIS), Universiti Technologi Malaysia (UTM), Kuala Lumpur, Malaysia
[2]Department of Computer Engineering, Islamic Azad University, Kerman Branch, Kerman, Iran

Abstract: The purpose of this study is to provide a review to mining useful information by means of Data Mining. The procedure of extracting knowledge and information from large set of data is data mining that applying artificial intelligence method to find unseen relationships of data. There is more study on data mining applications that attracted more researcher attention and one of the crucial field is criminology that applying in data mining which is utilized for identifying crime characteristics. Detecting and exploring crimes and investigating their relationship with criminals are involved in the analyzing crime process. Criminology is a suitable field for using data mining techniques that shows the high volume and the complexity of relationships between crime datasets. Therefore, for further analysis development, the identifying crime characteristic will be the first step and obtained knowledge from data mining approaches is a very useful tool to help and support police forces. This research aims to provide a review to extract useful information by means of Data Mining, in order to find crime hot spots out and predict crime trends for them using crime data mining techniques.

Keywords: Crime data mining techniques, forensics analysis, web crime mining, web mining

INTRODUCTION

Criminal web data always offer valuable and appropriate information for Law administration. The evaluation of the different capacities of widespread criminal web data is very difficult all the time so it is one of the most noteworthy tasks for law administration. Crimes may be as extreme as murder and rape where advanced analytical methods are required to extract useful information from the data Web mining comes in as a solution (Hosseinkhani *et al*., 2012b; Fayyad and Uthurusamy, 2002).

Definitely, one of influential factors that encounter a crime phenomenon is the humans' social life circumstances so the crime analysis knowledge is needed as an efficient combating tool. It also comprises of leveraging a systematic approach for discovering, identifying and predicting crime incidents and its input is contained assigned information and data in crime variables and the output contains the answer to knowledge extraction, analytical and investigative questions and the visualization of the results. Due to the criminality-related data and crime complexity and also the existence of intangible relations between them, data mining a rapidly made in growing field among criminologists. In the police departments, large volumes of crime-related data are existed. Due to the crime-related complexity relationships, the traditional methods of crime analysis are out-of-date that consume many time and human resources. Moreover, these methods are not able to obtain all influential parameters because of their high amount of human interference, therefore, using an intelligent and systematic approach for crime analysis more than ever. However, the data mining techniques can be the key solution (Keyvanpour *et al*., 2011).

Areas of concentrated crime are often referred to as hot spots. Researchers and police use the term in many different ways. Like researchers, crime analysts look for concentrations of individual events that might indicate a series of related crimes. They also look at small areas that have a great deal of crime or disorder, even though there may be no common offender. Analysts also observe neighborhoods and neighborhood clusters with high crime and disorder levels and try to link these to underlying social conditions.

Nowadays, the accessible data sources are provided by the rapid growth of the Web that has many specific characteristics. In fact, these characteristics make the mining useful knowledge and information a challenging task. It is necessary to know data mining in order to discover information mining on the Web that is exist in many Web mining tasks. Though, Web mining is not completely the application of data mining (Hosseinkhani *et al*., 2012a).

Data mining is defined as the process of discovering, extracting and analyzing meaningful patterns, structure, models and rules from large

Corresponding Author: Javad Hosseinkhani, Advanced Informatics School (AIS), Universiti Technologi Malaysia (UTM), Kuala Lumpur, Malaysia

quantities of data. Data mining is emerging as one of the tools for crime detection, clustering of crime location for finding crime hot spots, criminal profiling, predictions of crime trends and many other related applications.

The aim of web mining is to extract appropriate information from the page content, Web hyperlink structure and usage data. Although Web mining uses many data mining techniques, it is not purely an application of traditional data mining due to the heterogeneity and semi-structured or unstructured nature of the Web data (Hosseinkhani *et al*., 2012a).

The user is interested to identify crime hot spots of a particular region on certain crime types for a specific period. In order to fulfill such a requirement, a user interactive query interface is needed. Kumar *et al*. (2009) has presented an interactive media system with capability to adapt to various conditions from user preferences and terminal capabilities to network constraints. Newsome *et al*. (1997) has proposed Hyper SQL as a web-based query interfaces for biological databases. The design of query interfaces to biological database has also been presented by Che *et al*. (1999). However, no online adaptive query interface has been designed for mining crime data. The purpose of the research is to design an adaptive query interface for mining crime data or similar kind of problems. The proposed query interface provides a tool for making an online query and helps in identifying crime hot spots; predict crime trends for the crime hot spots based on the query.

Criminal web data provide unknown and valuable information for Law enforcement agencies continuously. The analysis of vast capacities of comprehensive criminal web data is very complicated in an area over periods of time and that is one of the most significant tasks for law enforcement. Crime database consists of various relational tables which contains the information about crime details in a region under various crime heads such as murder, rape etc. at different time points. Advanced analytical methods are required to extract useful information from large amount of crime data. Data mining is looked as a solution to such problems (Hosseinkhani *et al*., 2012a).

Many scientific researchers have been done on the importance of crime data mining and their results are revealed in the new software applications to analysis and detecting the crime data.

A framework has been developed by Hosseinkhani *et al*. (2012a) for crime web mining consists of two parts. In the first part, some pages which are concerned with the targeted crime are fetched. In the second part, the content of pages are parsed and mined. In fact, a crawler fetches some pages which are associated with the crimes. Previously, pages were fetched by crawler at a time, which was inefficient since the resource was wasted. The proposed model intends to promote efficiency by taking advantage of multiple processes, threads and asynchronous access to resources.

According to research by Hosseinkhani *et al*. (2013) the aim was suggesting a framework by using concurrent crawler to show the process of exploring the criminal accused of legal data evaluation which insures the reliability gap.

WEB CRIME MINING

All intelligence-gathering and law-enforcement organizations major challenge is facing to the efficient and correct evaluating of the crime data growing volumes. One of the examples of this can be complex conspiracies that are often hard to undo since the knowledge of suspects can be geographically span and diffuse in the long time. Detecting cybercrime can be very hard as well, because of frequent online transactions and busy network traffic which create huge amounts of data and just a portion of which relates to illegal activities (Hosseinkhani *et al*., 2012a).

In the last decade, through the rapid growth of the Web and through the many unique characteristics, in the following some of them are shown that causes of attracting and challenging for mining the useful information and knowledge (Baldi *et al*., 2003):

- Facing to the huge amount of information on the Web that is very wide and diverse so any user can find information on almost anything on the Web.
- Huge amount of data from all types are exist in unstructured texts, semi-structured Web pages structured tables and multimedia files.
- The diversity of the information on the. Multiple pages show similar information in different words or formats based on the diverse authorship of Web pages that make the integration of information from multiple pages as a challenging problem.
- An association is exist on the significant amount of information of the Web. Hyperlinks are in Web pages across different sites and within a site. Hyperlinks are implicit conveyance of authority to the target pages in across different sites. And hyperlinks serve as information organization mechanisms within a site.
- The information on the Web is noisy that is comes from two main sources. The first one is that a typical Web page involves many pieces of information for instance the navigation links, main content of the page, copyright notices, advertisements and privacy policies. Only part of the information is useful for a particular application but the rest is considered noise. For performing a fine-grain, the data mining and Web information analysis, the noise should be removed. The second one is due to the fact that the Web does not have quality control of information, for example, a large

amount of information on the Web is of low quality because any one can write everything.

- The Web is about services for example most commercial Web sites allow the users to perform useful operations at their sites such as paying bills, purchasing products and filling the forms.
- The Web pages are dynamic that is the information is changes constantly. Copping the changes and monitoring them is an important issue for many applications.
- The Web is a virtual society that is not only information, data and services; it also is the organizations, the interactions of people and automated systems. Any user can communicate with people anywhere in the world easily and express his/her views on anything in Internet blogs, forums and review sites (Bing, 2007).

All these characteristics present both challenges and opportunities for mining and discovery of information and knowledge from the Web. This research only focuses on mining textual data. For mining of images, videos and audios please refer to Djeraba *et al*. (2007) and Perner (2003). To explore information mining on the Web, it is necessary to know data mining, which has been applied in many Web mining tasks. However, Web mining is not entirely an application of data mining. Due to the richness and diversity of information and other Web specific characteristics discussed above, Web mining has developed many of its own algorithms (Bing, 2007).

The Web mining process and the data mining process are very similar to each other and their difference is just in their data collection. In traditional method of data mining, the data is gathered and stored in a data warehouse and the other hand, in Web mining, the data gathered is a substantial task that includes crawl of the large number of target Web pages (Duda *et al*., 2001).

Web pages are also quite different from conventional text documents used in traditional IR systems. First, Web pages have hyperlinks and anchor texts, which do not exist in traditional documents (except citations in research publications). Hyperlinks are extremely important for search and play a central role in search ranking algorithms. Anchor texts associated with hyperlinks too are crucial because a piece of anchor text is often a more accurate description of the page that its hyperlink points to. Second, Web pages are semi-structured. A Web page is not simply a few paragraphs of text like in a traditional document. A Web page has different fields, e.g., title, metadata, body, etc. The information contained in certain fields (e.g., the title field) is more important than in others. Furthermore, the content in a page is typically organized and presented in several structured blocks (of rectangular shapes). Some blocks are important and some are not (e.g., advertisements, privacy policy, copyright notices, etc.). Effectively detecting the main content block (s) of a Web page is useful to Web search because terms appearing in such blocks are more important (Hosseinkhani *et al*., 2012b).

A criminal might either give a deceptive identity or falsely use an innocent person's identity. There are currently two ways law enforcement officers can determine false identities. First, police officers can sometimes detect a deceptive identity during interrogation and investigation by repeated and detailed questioning, such as asking a suspect the same question ("What is your Social Security number?") over and over again. The suspect might forget his or her false answer and eventually reply differently. Detailed questioning may be effective in detecting lies, such as when a suspect forgets detailed information about the person whose identity he or she is impersonating. However, lies are difficult to detect if the suspect is a good liar. Consequently, there are still many deceptive records existing in law enforcement data. Sometimes a police officer must interrogate an innocent person whose identity was stolen, until the person's innocence is proven (Hosseinkhani *et al*., 2012b). Second, crime analysts can detect some deceptive identities through crime analysis techniques, of which link analysis is often used to construct criminal networks from database records or textual documents. Besides focusing on criminal identity information, link analysis also examines associations among criminals, organizations and vehicles, among others. However, in real life crime analysis usually is a time consuming investigative activity involving great amounts of manual information processing (Hosseinkhani *et al*., 2012b).

CRIME DATA MINING TECHNIQUES

The traditional data mining techniques just classify the patterns in structured data for example, classification and prediction, association analysis, outlier analysis and cluster analysis. On the other hand, the newer techniques identify patterns from unstructured and structured data (Han and Kamber, 2010). Crime data mining increases the privacy concerns like the other forms of data mining (Kargupta *et al*., 2003). However, the researchers' effort to promote the various automated data mining techniques for national security applications and local law enforcement. Particular patterns are identifies by Entity extraction from data such as images, text, or audio materials that has been utilized to automatically identify addresses, persons, vehicles and personal characteristics from police narrative reports (Chau *et al*., 2007) In computer forensics, the extraction of software metrics (Gray *et al*., 1997) which includes the data structure, program flow, organization and quantity of comments and use of variable name scan facilitate further

investigation by, for example, grouping similar programs written by hackers and tracing their behavior. Entity extraction provides basic information for crime analysis, but its performance depends greatly on the availability of extensive amounts of clean input data.

The main techniques of the crime data mining are clustering (Jain *et al*., 1999), association rule mining (Agrawal *et al*., 1993), classification (Han and Kamber, 2009) and sequential pattern mining (Agrawal and Srikant, 1995). Although all of these efforts, the crime Web mining still is a highly complex task:

- Clustering techniques group data objects into classes by similar characteristics to minimize or maximize interclass similarity for instance, to identify suspects that bearing the crimes in similar ways or discriminate among groups belonging to different gangs. These techniques do not have a set of predefined classes for assigning items. Some researchers use the statistics-based concept space algorithm to automatically associate different objects such as persons, organizations and vehicles in crime records (Hauck *et al*., 2002). Using link analysis techniques to identify similar transactions, the Financial Crimes Enforcement Network AI System (Senator *et al*., 1995) exploits Bank Secrecy Act data to support the detection and analysis of money laundering and other financial crimes. Clustering crime incidents can automate a major part of crime analysis but is limited by the high computational intensity typically required.
- Association rule mining determines frequently occurring item sets in a database and offerings some patterns as rules that been used in network intrusion detection to develop the connection rules from users' interaction history. Investigators also can apply this technique to network intruders' profiles to help detect potential future network attacks (Lee *et al*., 1999). Similar to association rule mining, sequential pattern mining finds frequently occurring sequences of items over a set of transactions that occurred at different times. In network intrusion detection, this approach can identify intrusion patterns among time-stamped data. Showing hidden patterns benefits crime analysis, but to obtain meaningful results requires rich and highly structured data.
- Deviation detection utilizes the particular measures to study data that differs noticeably from the rest of the data. Also called outlier detection, investigators can apply this technique to fraud detection, network intrusion detection and other crime analyses. However, such activities can sometimes appear to be normal, making it difficult to identify outliers.
- Classification finds mutual properties between various crime entities and arranges them into predefined classes that have been applied for identifying the source of e-mail spamming according to the sender's structural features and linguistic patterns (Vel *et al*., 2001). Often used to predict crime trends, classification can reduce the time required to identify crime entities. However, the technique requires a predefined classification scheme. Classification also requires reasonably complete training and testing data because a high degree of missing data would limit prediction accuracy.
- String comparator techniques that show the relation the textual fields in pairs of database records and calculate the correspondence among the records that can detect deceptive information in criminal records for instance the name and address (Wang *et al*., 2004). The researchers can utilize string comparators to evaluate textual data that often need intensive computation.

  String comparison is the interesting field for computer scientists that whether string matching or string distance measures. Levenshtein define a usual measure of similarity between two strings as "edit distance" (Levenshtein, 1966) so, the minimum number of, deletions, single character insertions and substitutions need to transform one string into the other. Jaro's method is the edit distance measure of outperforms since it can manage all kinds of string patterns and it does not detect phonetic errors because this method is designed to detect the spelling differences of two strings.
- A description of the nodes role in a conceptual network is Social network analysis. Investigators can use this technique to construct a network that illustrates criminals' roles, the flow of tangible and intangible goods and information and associations among these entities. Further analysis can reveal critical roles and subgroups and vulnerabilities inside the network. This approach enables visualization of criminal networks, but investigators still might not be able to discover the network's true leaders if they keep a low profile.

CONCLUSION

The majority of digital evidence is collected from textual data such as blogs, as e-mails, web pages, text documents and chat logs. The researcher uses some search tools to explore and extract the useful information from the text because the nature of textual data is unstructured and then for further investigation, enter the appropriate pieces into a well-structured database manually which will be boring and error prone. Therefore, the investigators expertise and experience is very important in search and the quality of an analysis. If a criminal hide some essential information, it may be missed.

In this review all preliminary concepts such as Web Mining, Criminal Identities and Crime Data Mining Techniques are described. The vision of the Web Mining is to provide a Web where all published material is understandable by software agents. Moreover, Data Mining defined as the process of discovering useful patterns or knowledge from data sources, e.g., databases, texts, images, the Web, etc. Web mining aims to discover useful information or knowledge from the Web hyperlink structure, page content and usage data. Inspection of files involves searching content for information that can be used as evidence or that can lead to other sources of information that may assist the investigation process and analysis of the retrieved information. It is typically up to the investigator on what and how to search for evidence, depending on the case.

Therefore, we evaluated State-of-the-Art approaches for extracting useful information by means of Data Mining, in order to find crime hot spots out and predict crime trends for them using crime data mining techniques.

ACKNOWLEDGMENT

This Project is sponsored by Ministry of Higher Education (MOHE) in corporation with Universiti Teknologi Malaysia, Research Management Centre (UTM-RMC) under Vote No: 03H74. The authors also would like to thanks those who are directly or indirectly involved in this project.

REFERENCES

Agrawal, R. and R. Srikant, 1995. Mining sequentiel motifs. Proceeding of the 11th Intrnational Conference on Data Engineering.

Agrawal, R., T. Imielinski and A.N. Swami, 1993. Mining association rules between sets of items in large databases. Proceedings of the ACM SIGMOD International Conference on Management of Data.

Baldi, P., P. Frasconi and P. Smyth, 2003. Modeling the Internet and the Web: Probabilistic Methods and Algorithms, Wiley, Chichester.

Bing, L., 2007. Web Data Mining: Exploring Hyperlinks, Contents and Usage Data. Springer Verlag, NY.

Chau, M., J.J. Xu and H. Chen, 2007. Extracting meaningful entities from police narrative reports. Proceeding of the National Conference on Digital Government Research. Digital Government Research Center, pp: 271-275.

Che, D., K. Aberer and Y. Chen, 1999. The design of query interfaces to the GPCRDB biological database. Proceedings of the User Interfaces to Data Intensive Systems.

Djeraba, C.O., R. Zaiane and S. Simoff, 2007. Mining Multimedia and Complex Data. Springer, New York.

Duda, R.O., P.E. Hart and D.G. Stork, 2001. Pattern Classification. 2nd Edn., John Wiley and Sons Inc., New York.

Fayyad, U.M. and R. Uthurusamy, 2002. Evolving data into mining solutions for insights. Commun. ACM Evol. Data Min. Solut. Insights, 45(8): 28-31.

Gray, A., P. Sallis and S. MacDonell, 1997. Software forensics: Extending authorship analysis techniques to computer programs. Proceeding of the 3rd Biannual Conference International Association Forensic Linguistics. International Association of Forensic Linguistics, pp: 1-8.

Han, J. and M. Kamber, 2009. Data Mining: Concepts and Techniques. Morgan Kaufmann, San Francisco.

Han, J. and M. Kamber, 2010. Data Mining: Concepts and Techniques. Morgan Kaufmann, San Francisco.

Hauck, R.V., H. Atabakhsb, P. Ongvasith and H. Gupta, 2002. Using coplink to analyze criminal-justice data. Computer, 35(3): 30-37.

Hosseinkhani, J., S. Chuprat and H. Taherdoost, 2012a. Criminal network mining by web structure and content mining. Proceeding of the 11th WSEAS International Conference on Information Security and Privacy (ISP '12). Prague, Czech Republic, September 24-26.

Hosseinkhani, J., S. Chuprat, H. Taherdoost and S.M. Amin, 2012b. Propose a framework for criminal mining by web structure and content mining. Int. J. Adv. Comp. Sci. Info. Technol., 1(1): 1-13.

Hosseinkhani, J., H. Taherdoost and S. Chuprat, 2013. Discovering criminal networks by web structure mining. Proceeding of the 7th International Conference on Computing and Convergence Technology. Seoul, South Korea, December 3-5 (In Press).

Jain, A.K., M.N. Murty and P.J. Flynn, 1999. Data clustering: A review. ACM Comput. Surv., 31(3): 264-323.

Kargupta, H., K. Liu and J. Ryan, 2003. Privacy-sensitive distributed data mining from multi-party data. Proceeding of the 1st NSF/NIJ Symposium on Intelligence and Security Informatics. LNCS 2665, Springer-Verlag, pp: 336-342.

Keyvanpour, M., M. Javideh and M. Ebrahimi, 2011. Detecting and investigating crime by means of data mining: A general crime matching framework. Proc. Comput. Sci., 3(2011): 872-880.

Kumar, M., A. Gupta and S. Saha, 2009. Approach to adaptive user interfaces using interactive media systems. Proceedings of the 11th International Conference on Intelligent User Interfaces.

Lee, W., S.J. Stolfo and W. Mok, 1999. A data mining framework for building intrusion detection models. Proceeding of the 1999 IEEE Symposium on Security and Privacy. Oakland, CA, pp: 120-132.

Levenshtein, V.L., 1966. Binary codes capable of correcting deletions, insertions and reversals. Soviet Phys. Doklady, 10: 707-710.

Newsome, M., C. Pancake and J. Hanus, 1997. HyperSQL: Web-based query interfaces for biological databases. Proceedings of the 13th Hawaii International Conference on System Sciences.

Perner, P., 2003. Data Mining on Multimedia Data. Springer, New York.

Senator, T., H. Goldberg, J. Wooton, A. Cottini, A. Umar, C. Klinger, W. Llamas, M. Marrone and R. Wong, 1995. The FinCEN artificial intelligence system: Identifying potential money laundering from reports of large cash transactions. AI Mag., 16(4): 21-39.

Vel, O. de., A. Anderson, M. Corney and G. Mohay, 2001. Mining e-mail content for author identification forensics. SIGMOD Record, 30(4): 55-64.

Wang, G., H. Chen and H. Atabakhsh, 2004. Automatically detecting deceptive criminal identities. Commun. ACM, 47(3): 70-76.