Criminal Network Mining by Web Structure and Content Mining

JAVAD HOSSEINKHANI[1], SURIAYATI CHAPRUT[1], HAMED TAHERDOOST[2]
[1]Advanced Informatics School
Universiti Teknologi Malaysia, Kuala Lumpur, Malaysia
[2]Department of Computer Engineering
Islamic Azad University, Semnan Branch, Semnan, Iran
[1]jhkhani@gmail.com, [1]suria@ic.utm.my, [2]hamed.taherdoost@gmail.com

Abstract: - Criminal web data provide unknown and valuable information for Law enforcement agencies continuously. The digital data which is applied in forensics analysis includes pieces of information about the suspects' social networks. However, there is challenging issue with regard to analysing these pieces of information. It is related to the fact that an investigator has to manually extract the useful information from the text in website and then establish connection between different pieces of information and categorise them into a structured database with which the set becomes ready to use various criminal network analysis tools for examination. It is believed that such process of preparing data for analysis which is done manually is not efficient because it is likely to be affected by errors. Besides, since the quality of resulted analysed data depends on the experience and expertise of the investigator, its reliability is not constant. In fact, the more experienced is an operator, the better result is gained. The main objective of this paper is to address the procedure of investigating the criminal suspects of forensic data analysis which cover the reliability gap by proposing a framework.

Key-Words: - Crime Web Mining, Terrorist Network, Criminal Network, Social Network, Forensics Analysis, Framework.

1 Introduction

Unknown and valuable information are always provided by criminal web data for Law enforcement agencies. 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. 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 [1, 2].

In many illegal situations, suspects have possession of computers including notebooks, desktops and smart phones which are the main aim of criminal attack and have important information about social networks of the suspect.

FBI Regional Computer Forensics Laboratory (RCFL) has been done 6000 researched from 689 law execution organizations against the United States through a year in the United States. In 2009, the amount data of these researches reached to 2334 Terabytes (TB) that is two times more than the amount in 2007. However, better resources are required to promote and increase demands and help the investigators process to collect data legally [15].

September 11th has called the attention of the American public for instance on the value of information collected from within terrorist cells. At least, a portion of these terroristic activities is online [4].

Most collected digital evidence is often textual such as e-mails, chat logs, blogs and web pages. The data is usually unstructured, demanding the investigator to use novel techniques to extract information from them. The task of data entry is manual which becomes laborious. Depending on the collector's expertise the completeness of information may vary and usually the criminal can hide whatever information he may desire [15].

There are many applications for crawling on the Web. One is surfing on the Internet and visiting web sites, it can help a user to notify when new information updated. Wicked applications are also exist for crawlers such as the spammers or theft attackers who use the email addresses to collect personal information. However, supporting the search engines are the most common use of crawlers. Actually, the main clients of Internet bandwidth are crawlers that help search engines to gather pages and build their indexes for example, proficient universal crawlers designed for research engines such as Google, Yahoo and MSN to collect all pages regardless to the content. Other crawlers are called preferential crawlers who are attempt to download only pages of certain types or topics and they are more targeted. Proposed framework applies preferential crawlers for crime web mining. Preferential crawlers are crawlers which fetch pages on the web based on ranked pages [5, 6].

In this paper, a framework is proposed for crime web mining which contains two sections. The former section, crawl the web based on ranked pages and the later section is for mining criminal networks by content mining.

2 Knowledge Discovery on the Web

Information detection and data mining are achieved by World Wide Web that are also presents an interesting opportunity for them. The growth of this area is very fast as business activity and research topic. The internet has effect on every aspect of daily life such as the way of learning, it means that internet can places anyone with a computer, and can prepare variety answers to any question.

2.1 Web Mining

The process of realizing, taking out and analyzing important structure, models, patterns, methods and rules from large amounts of web data is web mining. The rapid growth of the Web in the last decade makes it the largest publicly accessible data source in the world, for which reason also ironically; most of the information online is false and erroneous, since anyone can upload anything into the web. This makes web mining a challenging task [3].

Web mining aims to extract useful information or knowledge from the Web hyperlink structure, page content and usage data. There are differences between web mining and data mining. For example online data are heterogeneous and semi or unstructured for the mining of which a number of algorithms have been proposed over the past decade. Based on the types of primary data used in the mining process, Web mining tasks are categorized into three classes: Web structure mining, Web content mining and Web usage mining. Another difference is that in Web mining, data collection involves crawling a large number of target Web pages [5].

2.2 Data Processing Challenges of Criminal Network

Mining law application data contains many difficulties in other areas of data mining applications for example; incorrect, incomplete or inconsistent data which are described below. Moreover, these characteristics of criminal networks create difficulties not common in other data mining applications:

Incompleteness

Criminal networks are covert [17]. Criminals may minimize interactions to avoid the attention of the law with activities hidden behind various illicit acts. Subsequently data on criminals and their interactions and associations become incomplete, causing missing nodes and links in the network [16].

Incorrectness

Incorrect data on criminal characters that is physical features comes from accidental data entry errors or criminals active fraud. Many criminals option to misrepresentation under questioning.

Inconsistency

Information on a criminal with case history may enter into law enforcement databases under various instances; but these proceedings may not essentially relate. Various proceedings could make a single criminal which has diverse characters. In this situation, the information may be deceptive when they have apparently different individuals in a network under study. Criminal network analysis has specific problems such as data transformation, fuzzy boundaries, and network dynamics:

Data transformation

Network analysis need that data to be accessible in a precise design that nodes represent to network members, and their relations or connections are signified by links. On the other hand, usually the information of criminal relations is not clear in raw data. The process of pulling out criminal relations from raw data and changing them to the requisite design can be properly hard and time-consuming.

Fuzzy boundaries

Boundaries of criminal networks are most likely ambiguous making it difficult for the analyst to decide on the inclusion of the individual targets in a network under study [16].

Network dynamics

Criminal networks are changed during time it means that to access the dynamics of criminal networks, new data and collection methods may be essential [16].

Some developed techniques address these problems, for instance many experimental techniques are using in the FinCEN system at the U.S to develop data accuracy and reliability. Department of Treasury is established to disambiguate and combine financial businesses into exclusively known entities in the system [17]. Other methods such as the concept space [16] can change crime occurrence data into a networked design [17]. Criminal network mining is on the rise as a tool for crime detection, clustering of crime locations in search of hot spots, criminal profiling, crime trend prediction and many other related applications.

3 Related works

Many researchers have great attention to criminal network analysis. The previous works [7] have shown an effective use of data mining techniques to show the criminal associations from a large volume of incident reviews by police departments. These use co-occurrence frequencies to determine correlations between pairs of criminals [8] shows a method to pull out criminal networks from websites which is delivered blogging services all over a topic-specific investigation devices. In addition, they classify the performers in the network in their approach by utilizing web crawlers that examine blog subscribers. Blog subscribers are contributed in a discussion associated to some criminal topics. When the network is built, some text organization techniques are utilized to evaluate the content of the documents. Therefore, a visualization of the network is suggested to social network view or concept network view.

Al-Zaidy et al [15] proposed a work that is different in three aspects. Initially, their study focuses on unstructured textual data obtained from a suspect's hard drive rather than a well-structured police database. This method in turn, can discover prominent communities of indefinite size i.e. not limited to pairs of criminals. In addition, while most previous works identify direct relationships, the latter's methods also identify indirect relationships.

A social network paradigm is followed by criminal network. Therefore, the recommended method for social network analysis can be used in criminal networks. Many researchers have been conducted on the different methods which can be used to build a social network from text documents. Jin et al [9] proposed a framework to extract social networks from text documents available on the web. A method has been stated by [10] to rank companies based on the social networks extracted from WebPages. Mainly, these approaches are dependent on web mining techniques that are searched for the actors in the social networks from web documents. Other social network studies are focused on some type of text documents such as e-mails. Zhou et al [11] suggested a probabilistic approach that not only identifies communities in email messages but also extracts the relationship information using semantics to label the relationships. However, the method is only applicable to e-mails and the actors in the network are limited to the authors and recipients. Researchers in the field of knowledge discovery have proposed methods to analyze relationships between terms in text documents in a forensic context. Jin et al [12] introduced a concept association graph-based approach to search for the best evidence trail across a set of documents that connects two given topics. [13] The suggestions of the open and closed finding algorithms is to find and show evidence pathways that are between two topics, these two can be take place in the document set and it is not essentially to be in the same document. [14] In order to search for keywords that the users need, the open finding approach are used and bring back documents comprising related topics. Moreover, they utilize clustering techniques to evaluate the results and give the operator clusters of new information, this new information are related in concept of the initial request terms. Therefore, in order to improve the results of web queries, this open discovery approach explore for new links between concepts. In contrast, this paper focuses on extracting web published textual documents and information from criminal network sites for investigation

4 Proposed Framework

The proposed framework of 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 case of the initial part, it can be seen in Fig. 1 that a sequential crawler occurs. 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.

The whole process starts with having kept a list of unvisited URLs called the frontier. In fact, Frontier is considered as a priority queue which is applied in ranking pages because of its sensitivity. The list of URALS comes from the seed URLs which can be prepared by a user. Having prepared the URLS gives the chance that in each main loop, URL be picked from the frontier by crawler. Then, the page related to the URL is fetched by means of HTTP. Having fetched the page, the retrieved page is parsed, with which the URLs is extracted and after that newly discovered URLs is added to the frontier. It should be noted that the page or other extracted information not related to the targeted terms are stored in a local disk repository.

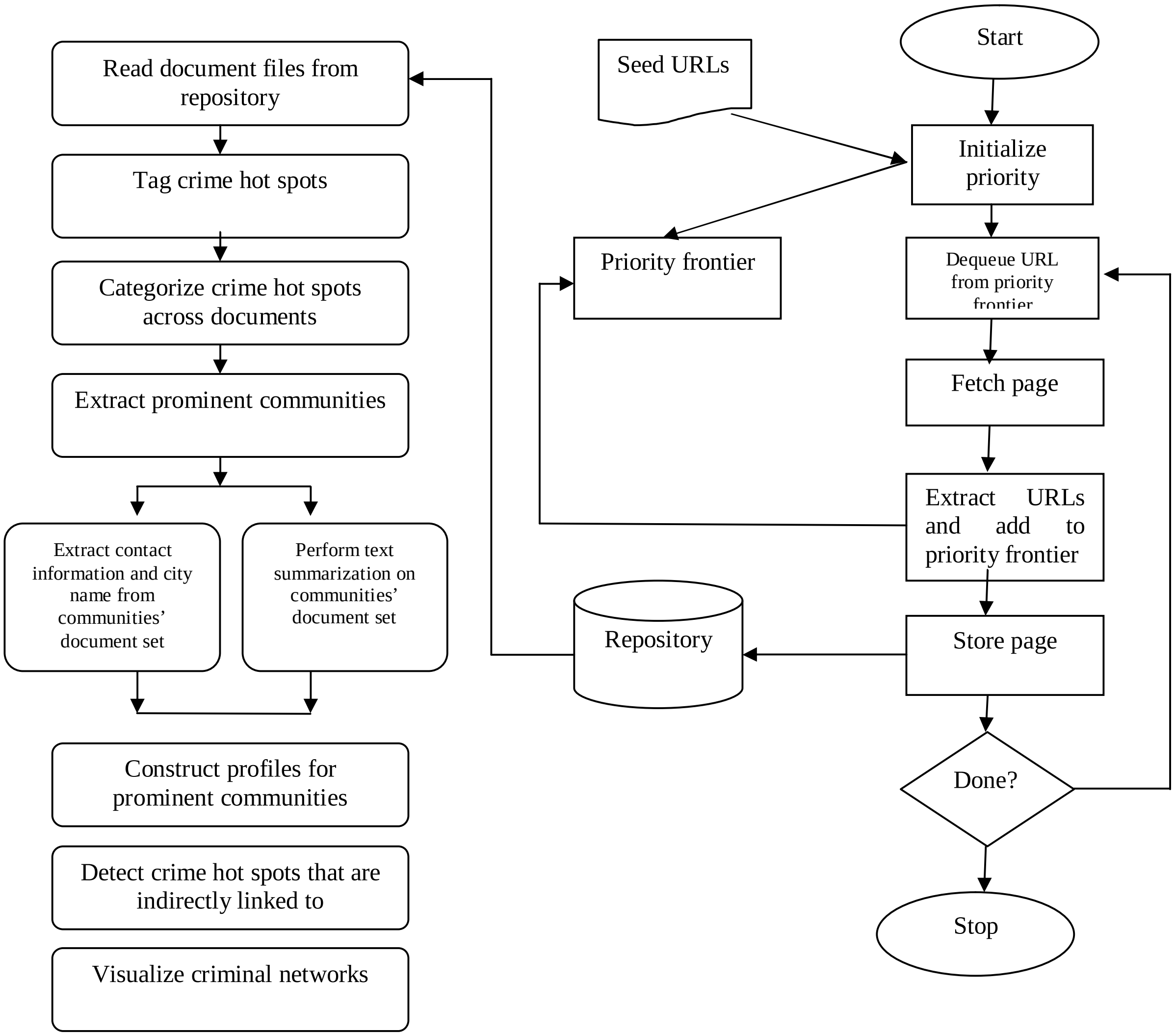
Read document files from repository
Tag crime hot spots
Categorize crime hot spots across documents
Extract prominent communities
Extract contact information and city name from communities' document set
Perform text summarization on communities' document set
Construct profiles for prominent communities
Detect crime hot spots that are indirectly linked to
Visualize criminal networks
Seed URLs
Start
Initialize priority
Priority frontier
Dequeue URL from priority frontier
Fetch page
Extract URLs and add to priority frontier
Repository
Store page
Done?
Stop

Fig. 1, Combined Websites and Textual Document Framework (CWTDF) for Criminal Network Mining

Termination of crawling can be done in several forms. In one case, the crawling ends once the intended number of pages is crawled. Besides that, the process can be pushed to be ended due to the frontiers' getting empty. However, this condition is not likely to happen because of the high average number of links.

Crawlers considered as the graph search algorithms can be applied in webs since webs can be regarded as a large graph, in which pages can be seen as its nodes and hyperlinks can be taken as its edges. In the process of searching in web, a crawler first searches a few of the nodes (seeds) and then by going after the edges, it reaches other nodes. This process of fetching a page and extracting the links is similar to expanding a node in graph search. The whole process is based on the frontier as its base data structure, which provides the crawler with the list of URLs unvisited pages. In order to enhance the efficiency of the process, the frontier is stored in the main memory by crawlers. However, the following points should be considered. It should be noted that a large frontier size is observed due to declining price of memory. Thus, the crawler designer should identify the URLs with low priority to be eliminated when the frontier is filled. Besides, it should be taken into consideration that the frontier is likely to fill quickly in case the size is maximum. Yet, other important point is that the sequence of extracting the URLs should be pre identified. In fact, the algorithm should be capable of specifying the order of appearances of URLs

The following steps lay out the procedure for the second part of the proposed model on parsing the contents of ranked pages. First, the text documents are investigated to extract the crime hot spot. Then, crime hot spots are records related to target crimes. Next, the normalization process is followed to remove the probability of unwanted crime hot spot duplication. Following this outstanding criminal communities are identified from the extracted crime hot spots. Having identified the crime community, the profile information useful to investigators including the contact information is provided. After that, the indirect relationships between the criminals across the document are established. Finally, a total scheme is prepared, in which visual representation of the prominent communities, their related information, and the indirect relationships are presented.

5 Conclusion

This paper presents an integrated framework for investigating the criminal suspect forensic data analysis. Previous studies on criminal network analysis mainly focus on analysing links between criminals in structured data or mere text documents. This paper has introduced the framework in two parts. The first part extracts pages related to crimes, and the second part parse and mines contents of prominent pages. For future works we recommend the presentation of some effective algorithms for various components of the framework indicated in this article for example, an algorithm for the prioritization of the URLs for the new frontier queue and even the differentiation of the techniques to identify criminal networks as opposed to existing ones.

References:

[1] U.M. Fayyad and R. Uthurusamy, “Evolving Data Mining into Solutions for Insights,” Comm. ACM, Aug. 2002, pp. 28-31.

[2] W. Chang et al., “An International Perspective on Fighting Cybercrime,” Proc. 1st NSF/NIJ Symp. Intelligence and Security Informatics, LNCS 2665, Springer-Verlag, 2003, pp. 379-384.

[3] Kaur, P. G., Raghu ; Singh, Ravinder ; Singh, Mandeep (2012). Research on the application of web mining technique based on XML for unstructured web data using LINQ. 2011 7th International Conference on MEMS, NANO and Smart Systems, ICMENS 2011. Kuala Lumpur, Malaysia, Trans Tech Publications, P.O. Box 1254, Clausthal-Zellerfeld, D-38670, Germany. 403-408: 1062-1067.

[4] Xu, J.J., Chen, H.: CrimeNet Explorer: A framework for criminal network knowledge discovery. ACM Transactions on Information Systems 23(2), 201–226 (2005)

[5] Peng Tao, "Research on Topical Crawling Technique for Topic- Specific Search Engine," Doctor degree thesis of Jilin University, 2007.

[6] Jiang Peng and Song Ji-hua, "A Method of Text Classifier for Focused Crawler," JOURNAL OF CHINESE INFORMATION PROCESSING, vol. 26, pp. 92-96 Nov. 2010.

[7] Chen H, Chung W, Xu JJ, Wang G, Qin Y, Chau M. Crime data mining: a general framework and some examples. Computer 2004;37(4):50–6.

[8] Yang CC, Ng TD. Terrorism and crime related weblog social network: link, content analysis and information visualization. In: IEEE international conference on intelligence and security informatics (ISI);2007. p. 55–8.

[9] Hope T, Nishimura T, Takeda H. An integrated method for social network extraction. In: Proc. Of the 15th international conference on world wide web (WWW); 2006. p. 845–6.

[10] Jin W, Srihari RK, Ho HH. A text mining model for hypothesis generation. In: Proc. Of the 19th IEEE international conference on tools with artificial intelligence ICTAI; 2007. p. 156–62.

[11] Zhou D, Manavoglu R, Li J, Giles CL, Zha H. Probabilistic models for discovering e-communities. In: Proc. of the 15th international conference on world wide web (WWW); 2006. p. 173–82.

[12] Jin Y, Matsuo Y, Ishizuka M. Ranking companies on the web using social network mining. In: Ting IH,Wu HJ, editors.Web mining applications in e-commerce and e-services. Studies in computational intelligence, vol. 172. Berlin/Heidelberg: Springer; 2009. p. 137–52.

[13] Srinivasan P. Text mining: generating hypotheses from medline.Journal of the American Society for Information Science and Technology 2004; 55:396–413.

[14] Skillicorn DB, Vats N. Novel information discovery for intelligence and counterterrorism. Decision Support Systems 2007;43(4): 1375–82.

[15] Al-Zaidy, R. F., Benjamin C.M.; Youssef, Amr M ; Fortin, Francis (2012). "Mining criminal networks from unstructured text

documents." Concordia Institute for Information Systems Engineering, Concordia University, 1455 De Maisonneuve Blvd. West, CIISE (EV7.640), Montreal, QC H3G 1M8, Canada 8: 147-160.

[16] Sparrow, M.K. The application of network analysis to criminal intelligence: An assessment of the prospects. Social Networks 13 (1991), 251–274.

[17] Krebs, V. E. Mapping networks of terrorist cells. Connections 24, 3 (2001), 43–52.

[18] Bowman, M., Debray, S. K., and Peterson, L. L. 1993. Reasoning about naming systems. *ACM Trans. Program. Lang. Syst.* 15, 5 (Nov. 1993), 795-825. DOI= http://doi.acm.org/10.1145/161468.16147.

[19] Ding, W. and Marchionini, G. 1997. *A Study on Video Browsing Strategies*. Technical Report. University of Maryland at College Park.

[20] Fröhlich, B. and Plate, J. 2000. The cubic mouse: a new device for three-dimensional input. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (The Hague, The Netherlands, April 01 - 06, 2000). CHI '00. ACM, New York, NY, 526-531. DOI= http://doi.acm.org/10.1145/332040.332491.

[21] Tavel, P. 2007. *Modeling and Simulation Design*. AK Peters Ltd., Natick, MA.

[22] Sannella, M. J. 1994. *Constraint Satisfaction and Debugging for Interactive User Interfaces*. Doctoral Thesis. UMI Order Number: UMI Order No. GAX95-09398., University of Washington.

[23] Forman, G. 2003. An extensive empirical study of feature selection metrics for text classification. *J. Mach. Learn. Res.* 3 (Mar. 2003), 1289-1305.

[24] Brown, L. D., Hua, H., and Gao, C. 2003. A widget framework for augmented interaction in SCAPE. In *Proceedings of the 16th Annual ACM Symposium on User Interface Software and Technology* (Vancouver, Canada, November 02 - 05, 2003). UIST '03. ACM, New York, NY, 1-10. DOI= http://doi.acm.org/10.1145/964696.964697.

[25] Yu, Y. T. and Lau, M. F. 2006. A comparison of MC/DC, MUMCUT and several other coverage criteria for logical decisions. *J. Syst. Softw.* 79, 5 (May. 2006), 577-590. DOI= http://dx.doi.org/10.1016/j.jss.2005.05.030.

[26] Spector, A. Z. 1989. Achieving application requirements. In *Distributed Systems*, S. Mullender, Ed. ACM Press Frontier Series. ACM, New York, NY, 19-33. DOI= http://doi.acm.org/10.1145/90417.90738.