International Journal of Advanced Computer Science and Information Technology (IJACSIT)
Vol. 2, No. 2, April 2013, Page: 13-22, ISSN: 2296-1739

www.elvedit.com

Criminal Communities Mining on the Web

Authors

Javad Hosseinkhani *jhkhani@gmail.com*
Department of Computer, Science and Research Branch, Islamic Azad University, Zahedan, Iran *Zahedan, Iran*

Suriayati Chuprat *suria@ic.utm.my*
Advanced Informatics School, University Technology Malaysia, Kuala lumpur, Malaysia *Kuala Lumpur, Malaysia*

Hamed Taherdoost *hamed.taherdoost@gmail.com*
Department of Computer Engineering/ Islamic Azad University, Semnan Branch *Semnan, Iran*

Morteza Harati Cool *Morteza_Harati67@yahoo.com*
Department of Computer, Science and Research Branch, Islamic Azad University, Zahedan, Iran *Zahedan, Iran*

Sadegh Emami Korani *Sadegh_2008Emami@yahoo.com*
Department of Computer, Science and Research Branch, Islamic Azad University, Zahedan, Iran *Zahedan, Iran*

Abstract

Criminal web data always offer novel and useful knowledge and information for Law administration. The used digital data in legal assessments are involved parts of information about the accused' social networks. Thus, evaluation of these parts of information is challenging. Therefore, an investigator has to pull out the appropriate information from the text manually, these information are on the website. An investigator also makes the relationship between various parts of information and classify them in to a structured database so the category is ready to use in different criminal network evaluation tools for analysis. On the other hand, these manual processes are not adequate in the case that it has many errors. Moreover, as the quality of resulted evaluation depends on the investigator's proficiency, the reliability is not stable. On another word, the more proficient operator, the better result achieved. The purpose of this paper is 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.

Key Words

Crime Web Mining, Concurrent Crawler, Terrorist Network, Social Network, Forensics Analysis.

I. 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 [1, 27].

Doubts have controlled the computers such as desktops smart phones notebooks, and in many suspect situations. Computers are the main purpose 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 demand 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 are 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. Wicket applications 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 is 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 of the content. Other crawlers are called preferential crawlers who are attempting to download only pages of certain types or topics and they are more targeted. A 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].

A crawler uses three main properties: CPU, network, and disk that each one is a bottleneck with boundaries enforced by CPU speed, bandwidth, and disk transfer times. [27] Explained the simple order crawler to create a very incompetent use of these resources for the reason that each time, the crawler goes through the third, two of them are idle.

This paper, suggested a framework for crime web mining that is comprised of two sections. The first one, crawl the web depends on the ranked pages by using concurrent crawler. The second one section is content mining that is for mining criminal networks.

II. WEB MINING

Data mining and information investigation are obtained by World Wide Web. The improvement in this area is very fast like research topic and business activity. The internet has influence on every part of daily life, it also changes the way of learning. On the other word, it can replace anyone with a computer, and can prepare several appropriate answers to any question.

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, 27].

Web mining aims to extract useful information or knowledge from the Web hyperlink structure, page content and usage data. Data mining is different from web mining, for instance in 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].

Crawlers or spiders are the name of programs, which is downloaded Web pages automatically. The knowledge and information is distributed on the Web among billions of pages that assisted by millions of servers all around the world. The users browse the Web and track hyperlinks to right to use information; they surf on the Internet from one page to the next. For gathering information, a crawler visit many sites for evaluation and mined in a central location, it means off-line or online. Where the Web gathered pages, for crawling, we use little long term that is when all the pages are fetched and saved in a sources. On the other hand, the Web is a dynamic unit developing at rapid rates. Therefore, crawlers need to help requests stay current whenever links and pages are moved, deleted, added or modified. The most common use of crawlers is in the provision of search engines. Actually, the main consumers of \Internet bandwidth are crawlers that build their indexes by collecting pages for search engines [27].

Crawlers considered as the graph search algorithms can be applied in webs since webs can be regarded as a large group, in which pages can be seen as its nodes and hyperlinks can be taken as its edges. In the process of searching on the web, a crawler first searches a few of the nodes (seeds) and then by going over 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 the 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 maximized. Yet, another 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.

Generally, the second section of Web data mining is Web content mining, or text mining that is the mining and scanning of pictures, graphs, and text on a Web page to define the relevance of the content to the search request. The lack of structure that floods much of the ever growing information sources on the World Wide Web, such as hypertext documents, makes automated organization, and search and indexing tools on the Internet. The World Wide Web offers some ease to their user, but generally they do not offer structural information or filter, categorize or interpret documents. Recently, these factors have stimulated researchers to improve more intellectual tools for information recovery along with extending the database and data mining techniques to be for a higher level of organization for semi-structured data on the web.

III. DATA PROCESSING CHALLENGES OF CRIMINAL NETWORK

Mining law application data comprises of many problems in other areas of data mining applications such as incomplete, inconsistent, incorrect data that explained below. In addition, these features of criminal networks make some difficulties that are not common in other data mining applications:

A. 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].

B. Incorrectness

In criminal features, the incorrect data have physical features that are from unplanned criminals active fraud or data entry errors and most of the criminal possibilities to misrepresentation under questioning.

C. 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:

D. Data transformation

Network analysis uses that data to be available in an accurate design that the nodes exemplify to network members, and their connections or associations are indicated by the links. However, in raw data, the information of criminal associations is not clear. The procedure of extracting criminal associations from raw data and changing them to the essential design can be accurately hard and wasting of time.

E. 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].

F. Network dynamics

Through the time, criminal networks are changed it means that in order to achieve the dynamics of criminal networks, collection methods and new data is essential [16].

Some developed techniques address these problems, for instance many experimental techniques are used in the FinCEN system in 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.

IV. Related Works

Criminal network evaluation attracted the attention of many researchers. The previous studies [7] have been done on the actual use of data mining techniques to represent the criminal relations from a large amount of event 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 contributing in a discussion associated with 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.

Hosseinkhani et al [27] proposed a 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 is 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.

A social network paradigm is followed by a 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 Web Pages. 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 focus 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 essential to be in the same document. [14] In order to search for keywords that the users need, the open finding approach is 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 explores for new links between concepts. In contrast, this paper focuses on extracting web published textual documents and information from criminal network sites for investigation.

V. PROPOSED FRAMEWORK

The simple sequential crawler in [27] makes a very incompetent use of these resources since at any time two of them are idle and the crawler appears in the third. The most direct way to speed up a crawler is through concurrent threads or procedures. Multiprocessing may be to some extent easier to use than multi threading. It is based on the platform and programming language, but it may also experience a higher overhead base on the participation of the functional system in the management of child procedures. A concurrent crawler tracks a standard parallel computing model as illustrated in Figure 1.

Principally each process or thread acts as an independent crawler, but in accessing to the shared data structures must be synchronized.

Dealing with a concurrent crawler is a bit more complicated for an empty frontier than for a sequential crawler [27]. An empty frontier not used for a long time till the crawler has stretched to a dead-end. So that other processes may be fetching pages and adding new URLs in the future. The thread or process manager may manage such a situation by sending a temporary sleep signal for processing that to report an empty frontier. The process manager needs to follow the number of sleeping processes; when all the processes are asleep, the crawler must stand still.

In this study, a concurrent crawler fetches some pages which are related to the crimes. Earlier, crawler fetched the pages at a time, which was incompetent because the resource was wasted. The suggested model aims to support efficiency by taking advantage of multiple threads, processes and asynchronous access to resources. The concurrent crawler can easily speed up a crawler by a factor of 5 or 10.

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.

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.

The following steps laid out the procedure for the second part of the proposed model for parsing the contents of rank 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 community 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.

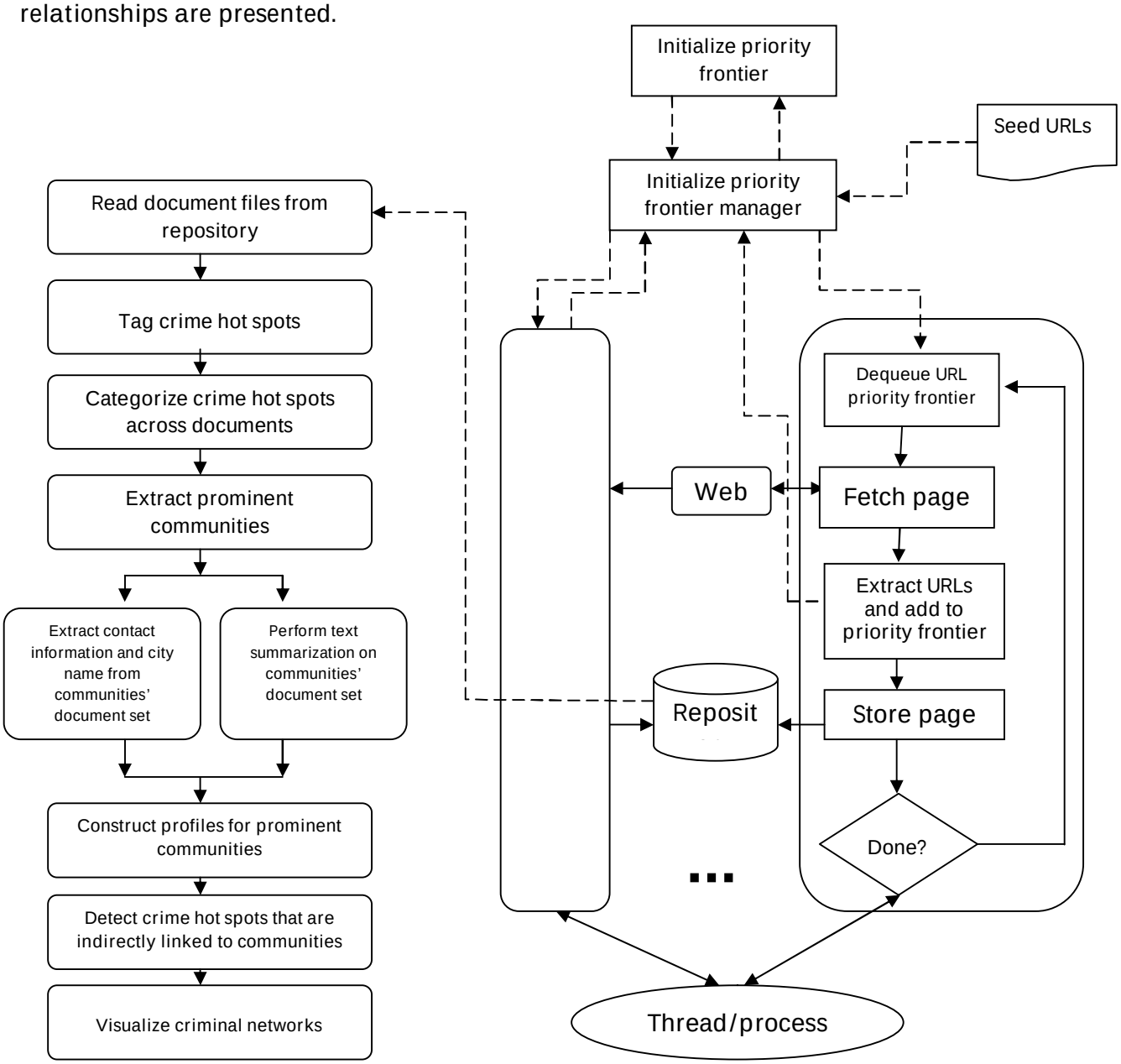

FIGURE I: COMBINED TEXTUAL DOCUMENT FRAMEWORK (CWTDF) AND WEBSITES FOR CRIMINAL NETWORK MINING

VI. CONCLUSION AND FUTURE WORK

The aim of this study is to recommend an integrated framework for finding the doubtful data evaluation by applying simultaneous crawler. The previous studies have done on the importance of criminal network assessments specially by concentrating on evaluating links between criminals in plain text documents or in structured data. Therefore, this paper has proposed the framework in two parts that in the first part extracts the related pages to crimes by using concurrent crawler, and in the second part mines and analysis the contents of the pages. Therefore, we suggest the presentation of some actual ontology-based crime web miner algorithms for different mechanisms of the framework which are designated in this paper as well further steps to obtain more scalable crawlers.

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.

[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.

[26] Spector, A. Z. 1989. Achieving application requirements. In *Distributed Systems*, S. Mullender, Ed. ACM Press Frontier Series. ACM, New York, NY, 19-33.

[27] Hosseinkhani, J, Chuprat. S, and Taherdoost. H. (2012). Criminal Network Mining by Web Structure and Content Mining, Advances in Remote Sensing, Finite Differences and Information Security, Prague, Czech Republic, September 24-26, 210-215.