Research Journal of Applied Sciences, Engineering and Technology 6(19): 3617-3624, 2013
DOI:10.19026/rjaset.6.3568
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: January 02, 2013 Accepted: February 18, 2013 Published: October 20, 2013

**Research Article**

# Proposing a Framework for Exploration of Crime Data Using Web Structure and Content Mining

[1]Amin Shahraki Moghaddam, [2]Javad Hosseinkhani, [2]Suriayati Chuprat,
[3]Hamed Taherdoost and [4]Hadi Barani Baravati
[1]Department of Computer, Zahedan Branch, Islamic Azad University, Zahedan, Iran
[2]Advanced Information School (AIS), Universiti Teknologi Malaysia (UTM),
Kuala Lumpur, Malaysia
[3]Deparment of Computer Engineering, Islamic Azad University, Semnan Branch, Semnan, Iran
[4]Department of Computer, Iranshahr Branch, Islamic Azad University, Iranshahr, Iran

**Abstract:** The purpose of this study is to propose a framework and implement High-level architecture of a scalable universal crawler to maintenance the reliability gap and present the evaluation process of forensic data analysis criminal suspects. In Law enforcement agencies, criminal web data provide appropriate and anonymous information. Pieces of information implemented the digital data in the forensic analysis to accused social networks but the assessment of these information pieces is so difficult. In fact, the operator manually should pull out the suitable information from the text in the website and find the links and classify them into a database structure. In consequent, the set is ready to implement a various criminal network evaluation tools for testing. As a result, this procedure is not efficient because it has many errors and the quality of obtaining the analyzed data is based on the expertise and experience of the investigator subsequently the reliability of the tests is not constant. Therefore, the better result just comes from the knowledgeable operator. The objectives of this study is to show the process of investigating the criminal suspects of forensic data analysis to maintenance the reliability gap by proposing a structure and applying High-level architecture of a scalable universal crawler.

**Keywords:** Crime web mining, criminal network, forensics analysis, framework, social network, terrorist network, universal crawler

## INTRODUCTION

Anonymous and suitable information always are provided by criminal web data for Law enforcement agencies. 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 (Fayyad and Uthurusamy, 2002; Hosseinkhani *et al*., 2012b).

In many suspect situations, suspicions have measured the computers for instance desktops smart phones notebooks. Computers have an important knowledge and information about social networks of the suspect, they also are the main target of criminal (Chang *et al*., 2003).

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 (Al-Zaidy *et al*., 2012). September 11$^{th}$ 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 (Xu and Chen, 2005).

The majority of the collected digital evidence is regularly in the textual context such as e-mails, chat logs, blogs and web pages. The utilized data are being usually uncategorized and need the investigator to apply new techniques to pull out information from them. The data entry also done manually that is very difficult. Based on the collector's proficiencies the totality of information may be so broad and criminals can hide whatever they want (Al-Zaidy *et al*., 2012).

For crawling of the Web, many applications exist. One is surfing on the Internet and visiting web sites, it can help a user to notify when new information

**Corresponding Author:** Amin Shahraki Moghaddam, Department of Computer, Zahedan Branch, Islamic Azad University, Zahedan, Iran

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 suggested framework uses a special crawler for crime web mining. Special crawlers are one that go and bring the web pages based on the ranking (Tao, 2007; Hosseinkhani *et al*., 2012a; Peng and Ji-Hua, 2010).

If the borders are applied as a priority queue sooner than a FIFO (First in First Out) queue, a various crawling strategy is achieved. Based on the assessment of the linked page's value, preferential crawlers usually give each unvisited link a main concern. The assessment might be based on content properties, topological properties or other mixture of measurable types. For instance, the objective of a topical crawler is to track edges which are lead to portions of the Web graph that are appropriate for a user-selected topic. In this case, the seeds chosen are more important than breadth-first crawlers (Hosseinkhani *et al*., 2012b).

The proposed framework of this study for crime web mining consists of two sections. The first part is High-level architecture of a scalable universal crawler, the crawl the web that constructed on ranked pages which content mining rank the downloaded pages to discover key URLs early section all over the crawl. The second part is criminal networks mining. In order to reduce the running time of the procedure in prioritizing URLs for crawling the desired pages, the priority algorithm are used.

## WEB MINING

Data mining and information detection are obtained by World Wide Web that are also shows an challenging possibilities 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 the internet can places anyone with a computer and can prepare variety 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 (Xu and Chen, 2005).

The aim of web mining is to extract an appropriate information from the page content, Web hyperlink structure and usage data. Data mining are different from web mining in case that over the past decade, online data are heterogeneous and semi or unstructured for the mining of which a number of proposed algorithms. Web usage mining Web mining tasks and Web content mining are characterized into three classes based on their used types of primary data in their mining. In addition, in Web mining, the gathered data are comprised of crawling a huge amount of target Web pages (Tao, 2007).

**Information Retrieval (IR) and web search:** No introduction is required for Web search. Base on the Web searching opportuneness and the productivity of the information on the web, it is progressively the main information seeking method. Therefore, there is a few people go to the libraries and more searches have been done on the Web. Actually, the writing of this section has been much harder without rich Web contents and effective search engines (Duda *et al*., 1995; Baeza-Yates and Ribeiro-Neto, 1999).

Web search has its root in information retrieval (or IR for short), a field of study that helps the user find needed information from a large collection of text documents. Traditional IR assumes that the basic information unit is a document and a large collection of documents is available to form the text database. On the Web, the documents are Web pages (Brin and Page, 1998).

Retrieving information simply means finding a set of documents that is relevant to the user query. A ranking of the set of documents is usually also performed according to their relevance scores to the query. The most commonly used query format is a list of keywords, which are also called terms. IR is different from data retrieval in databases using SQL queries because the data in databases are highly structured and stored in relational tables, while the information in the text is unstructured. There is no structured query language like SQL for text retrieval (Ding and Marchionini, 1997).

It is safe to say that Web search is the single most important application of IR. To a great extent, Web search also helped IR. Indeed, the tremendous success of search engines has pushed IR to the center stage. Search is, however, not simply a straightforward application of traditional IR models. It uses some IR results, but it also has its unique techniques and presents many new problems for IR research.

At the first, efficiency is a dominant issue in the Web search, but due to the fact that document collections in most IR systems are not very big, it is only secondary in traditional IR systems. On the other hand, the number of pages on the Web is huge for instance through the writing of this chapter, Google indexed more than 8 billion pages. The Web users also

are willing to have very fast responses and replies. If the retrieval is not user friendly and cannot do efficiently, it is not important how effective an algorithm is so just a few people can use it (Debnath *et al*., 2005).

There is a difference between conventional text documents and Web pages applied in traditional IR systems. The first one is the Web pages that have anchor texts and hyperlinks; this Web page does not be present in traditional documents. 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 (Forman, 2003).

As a final point, the most important issue in the Web is spamming, in the case that the rank position of a page is returned by a search engine. If a page is relevant to a query but is ranked very low (e.g., below top 30), then the user is unlikely to look at the page. If the page sells a product, then this is bad for the business. In order to improve the ranking of any target pages, "illegitimate" means, called spamming, are often used to boost their rank positions. Detecting and fighting Web spam is a critical issue as it can push low quality (even irrelevant) pages to the top of the search rankings, which harms the quality of the search results and the user's search experience.

**Criminal web mining:** From the Sept. 11, 2001, the fear of characteristics confirmation got new heights. Investigating identity deception is attracting more interest these days with national security issues. Identity deception is an intentional falsification of identity in order to deter investigations. Conventional investigation methods run into difficulty when dealing with criminals who use deceptive or fraudulent identities, as the FBI discovered when trying to determine the true identities of the 19 hijackers involved in the attacks. Besides its use in the post-event investigation, the ability to validate identity can also be used as a tool to prevent future tragedies (Krebs, 2001).

A sender is defined as a Interpersonal deception that significantly proposed communicating messages to raise a false conclusion or belief by the receiver (Burgoon *et al*., 1996). Methods have been developed to detect deception using physiological measures (for example, polygraph), nonverbal cues and verbal cues. Nonverbal cues are indications conveyed through communication channels such as micro-expression (for example, facial expression), eye movement and body language. Verbal cues are linguistic patterns exhibited in messages that may include deception. The empirical techniques can measure the reliability of verbal cues, The empirical techniques can be Criteria-Based Content Analysis and Statement Validity Assessment (Vrij, 2000). Police officers are trained to detect lies by observing nonverbal behaviors, analyzing verbal cues and/or examining physiological variations. Some are also trained as polygraph examiners. Because of the complexity of deception, there is no universal method to detect all types of deception. Some methods, such as physiological monitoring and behavioral cues examination, can only be conducted while the deception is occurring. Also, there is little research on detecting deception in data where few linguistic patterns exist (for example, profiles containing only names, addresses and so on). As a result, current trick detection techniques are technologically advanced for applications in physiology and communication that are not suitable for determining deception in identity profiles.

It is a common practice for criminals to lie about the particulars of their identity, such as name, date of birth, address and Social Security number, in order to deceive a police investigator. For a criminal using a falsified identity, even if it is one quite similar to the real identity recorded in a law enforcement computer system, an exact-match query can do very little to bring up that record. In fact, criminals find it is easy and effective to escape justice by using a false identity.

A criminal might provide a falsely characteristic that apply for an innocent person's identity. Law enforcement officers can determine two false ways characteristics. 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. Therefore, many deceptive records are available in law enforcement data (Tavel, 2007; Sannella, 1994).

The second one is crime of huge amounts of manual information processing analysts who can identify some deceptive identities by crime analysis techniques that make a connection analysis is applied to build a criminal networks from textual documents or

database records. In addition, by focusing on criminal identity information, the connection analysis studies the suggestions among organizations, criminals and vehicles. On the other hand, crime analysis is a time consuming analytical activity that is comprised of huge amounts of manual information processing in real life.

## LITERATURE REVIEW

Nowadays criminal network analysis has attracted more attention of the researchers. Based on the previous studies (Chen *et al*., 2004), by applying of data mining techniques, the criminal relations has been shown in a large volume of event reviews by police departments. In order to control relationships between pairs of criminals, they apply co-occurrence frequencies (Yang and Ng, 2007) that demonstrate a method to extract criminal networks from websites. 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*. (2012) have done a work that is different in three aspects. First, they just emphases on unstructured textual data that are achieved from a suspect's hard drive. 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 criminal networks track social network paradigm, for that reason, the method which is proposed for social network analysis also can be utilized for 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*. (2007) proposed a framework to extract social networks from text documents available on the web. A method has been stated by Hope *et al*. (2006) 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.

Through the literatures, the other social network works concentrate on other kind of text documents, for example e-mails. Another approach proposed by Zhou *et al*. (2006) that finds communities in email messages and pulls out the association information utilizing semantics to label the associations. 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*. (2009) introduced a concept association graph-based approach to search for the best evidence trail across a set of documents that connects two given topics. 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 (Skillicorn and Vats, 2007).

A framework has been developed for crime web mining consists of two parts (Hosseinkhani, *et al*., 2012b). 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.

The objective of the study was to suggest a framework by using concurrent crawler to show the process of exploring the criminal accused of legal data evaluation which insures the reliability gap.

The open finding approach is applied to search for keywords that the users need and bring the documents that are consist of related topics. Furthermore, clustering techniques are used to assess the findings and offer the operator clusters of new information by the open finding approach. This novel information is correlated to each other in the initial request terms. Consequently, these open discovery approaches are searching for new connections between concepts to develop the results of web queries. On the other hand, the aim of this study is to concentrate on extracting web published textual information and documents from criminal network sites by using High-level architecture of a scalable universal crawler for investigation.

## METHODOLOGY

The simple chronological crawler in research by Hosseinkhani *et al*. (2012b) creates very useless resources that two of them are idle and in any time the crawler performs at 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 research by Hosseinkhani *et al*. (2013). Principally each process or thread acts as an independent crawler, but in accessing to the shared data structures must be synchronized.

Moreover, concurrent crawler for an empty frontier is a bit complex than for a sequential crawler. 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.

A concurrent crawler in creates some pages that have some connection to the crimes. Previously, crawler fetched the pages immediately and they were incompetent because of the waste of resource. 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. On the other hand, the simultaneous architecture does not upgrade to the required performance of an effective search engine. Here the further steps are present to achieve more scalable crawlers.

In this study, we use a High-level architecture of a scalable universal crawler. The entire procedure begins with a list of unvisited URLs that are called the frontier. Actually, an important queue that is used in ranking pages based on its sensitivity is frontier. Users prepared the list of URALS that comes from the seed URLs. Preparing the URLS make some opportunities for each main loop that URL be picked from the frontier by crawler. At that time, the page correlated 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 complete in various forms, for example when the intended number of pages is crawled, the crawling ends once. Alongside, based on the frontiers' getting empty, the process can be pushed to be ended. On the other hand, this situation may not be happening due to the high average number of links.

For the second part of the offered model for parsing the contents of rank pages, the following steps are presented: first of all in order to pull out the crime hot spot, text documents should be explored. 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. The last one is the preparation of a total scheme that is a visual representation of related information, the prominent communities and the indirect relationships.

The main purpose of each search engine is using the Web crawlers to keep their amortizing the cost of crawling, indices and indexing over the millions of queries that are established between successive index updates. There is a different between the concurrent breadth-first crawlers and large-scale universal crawlers in two major dimensions.

**Performance:** They require upgrading fetching and processing hundreds of thousands of pages per second that is suitable for various architectural improvements.

**Policy:** They attempt to cover most of the possible important pages on the Web. And meanwhile they continue their index more fresh and update. These purposes are somehow incompatible with the purpose of the designed crawlers to obtain good tradeoffs between their objectives. Following the discussion of the main issues in meeting need these requirements.

Figure 1 shows the architecture of a large-scale crawler. One of the most significant of the concurrent model is the utilizing of procedures with synchronous sockets or asynchronous sockets in place of threads (Hosseinkhani *et al*., 2012a). Asynchronous sockets are non blocking, as a result a thread or single process can retain hundreds of network connections that are open simultaneously and help to use of network bandwidth efficiently. Based on the managing processes or threads, Asynchronous sockets remove the overhead, it also make access locking to shared data structures unnecessary. As a replacement, the sockets are asked to monitor their state. It is treated for the relationship between indexing and extraction, when the entire page has been fetched into memory. This "pull" model removes the need of locks and the contention of resources.

The efficiency of the crawler of frontier manager are improved by keeping various parallel queues that the URLs in each queue denote to a single server. Besides distribution, the load through various servers within any short time interval. Therefore, this approach have to create a connection with servers through many page requests, hence it minimizes the overhead of TCP and also closing and opening handshakes.

The crawler required to determine their host names in URLs to IP addresses. One of the most important bottlenecks of a naïve crawler is the Domain Name System (DNS) that crawler makes connection with that opens a new TCP link to the DNS server for each URL. The crawler should take several steps to discus on this bottleneck such as, first, it can utilize UDP as a replacement for TCP and as the transport protocol for DNS requests. However, UDP does not assure the delivery of a request and pockets that can be dropped irregularly. In contrast, UDP no connection acquires above an important speedup over TCP. Then, the DNS server utilizes insistent, large and fast caches. As a final

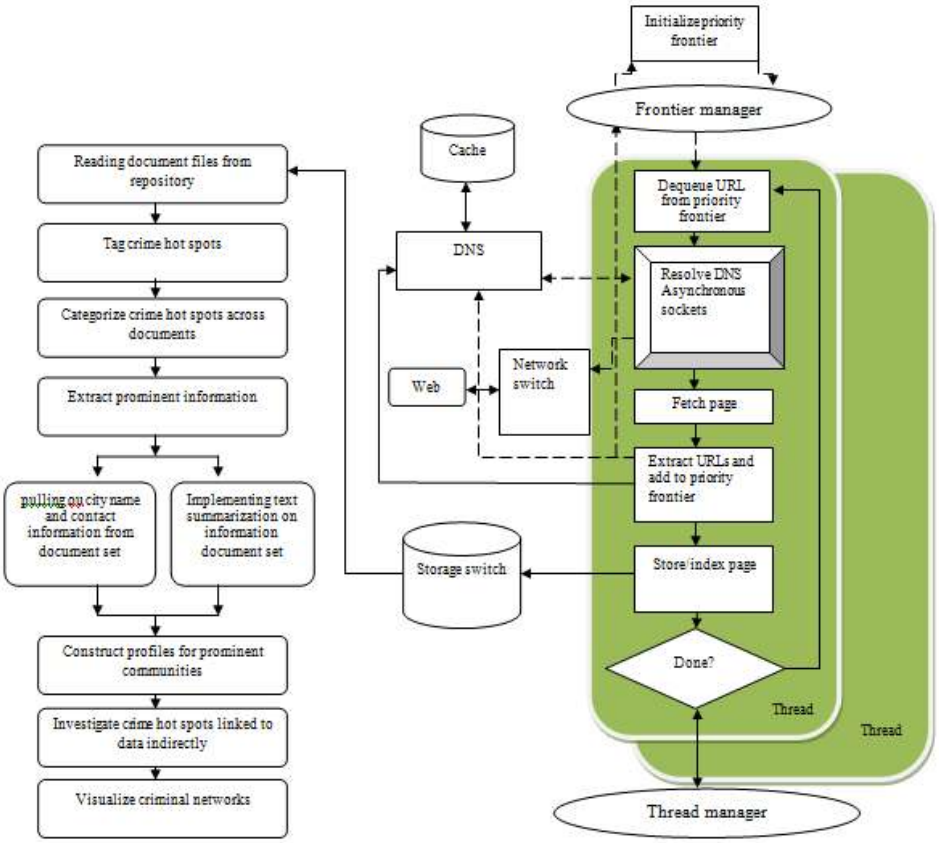


Fig. 1: The Combination of a Textual Document Framework (CWTDF) and websites by using high-level architecture of a scalable universal crawler

point, the pre-fetching of DNS requirements performs when the connections extract from a page. The URLs can be scanned for the sent host names to the DNS server alongside of being added to the frontier. The host IP address is likely to be found in the DNS cache when a URL is later ready to be fetched, avoiding the need to propagate the request through the DNS tree.

## ADVANTAGE OF THE FRAMEWORK

In this study, the framework which is proposed for crime web mining consists of two sections. The first part is crawling the web due to the ranked pages which is High-level architecture of a scalable universal crawler. High-level architecture of a scalable universal crawler are ranking of the downloaded pages to discover main URLs by content mining early section all over the crawl and the second part is for criminal networks mining. The aim of applying the priority algorithm is to reduce the running time of the procedure in arranging URLs.

The purpose is to cover a lot of pages that is in encounter with the need to keep an index fresh. Through the start of added pages and due to the highly dynamic nature of the Web, modified and deleted all the time, for a crawler to reconsider pages that are already in the index is necessary to keep the index up-to-date.

Besides to creating more efficient utilize of network bandwidth over and done with asynchronous sockets, the multiple network connections can increase network bandwidth which switches to multiple routers. Therefore, they are applying the networks of multiple Internet service providers. In the same way, disk I/O throughput can be improved via a storage area network connected to a storage pool through a fiber channel switch.

## CONCLUSION

The evaluation of the retrieved information and supporting the study procedure can be done by reviewing of files that is involve searching content for knowledge and information it means that reviewing propose and address other information sources which is based on the ways that the investigator used for searching evidence. In this study, the main objective is to bridge the gap between unstructured text data and criminal network mining. It means that, the challenging is that the mining criminal communities from a set of text files have been collected from a suspect's data. In the other hand, in a "Text files" such as blogs, chat

logs, web pages, e-mails, or any textual data, investigators effort to implement some other search tools to extract and classify appropriate information from the text due to its unstructured nature and after that for further analysis, enter the suitable pieces into a well-structured database manually. Thus this manual process is error prone, time consuming and boring and also the quality of an analysis and the comprehensiveness of a search pretty much depends on the investigators expertise and experience.

The aim of this study is to combine frameworks to discover suspicious to be evaluated by using simultaneous crawler. Due to the fact that the former studies are focuses on the criminal network evaluations significance by evaluating links between criminals in structured data, this study have been just focused on the framework in two parts such as a discussion on the related pages extraction to crimes by using a scalable universal crawler and investigating the contents of the pages. Accordingly, the actual ontology-based crime web miner algorithms are offered for different mechanisms of the framework as for further steps to achieve more scalable crawlers.

For future studies, we recommend some effective algorithms for various components of the framework that are indicated in this study, for example, An Enhanced Ontology-based Crime Web Miner Algorithm for crawling the web contents and importing different types of crime ontologies to find out the suspicious or malicious crimes. In that case, combining an effective and existing crime web miner algorithm with semantic Web and specifically OWL (as a Web Ontology Language) might be taken into account. As a result, a prototype uses Java programming to support the applicability of the proposed framework and the evaluation of the result of the proposed approach with other current approaches to determine the effectiveness of the approach will be conducted.

## REFERENCES


Al-Zaidy, R.F., C.M. Benjamin, A.M. Youssef and F. Fortin, 2012. Mining criminal networks from unstructured text documents. Digit. Invest., 8(3-4): 147-160.

Baeza-Yates, R. and B. Ribeiro-Neto, 1999. Modern Information Retrieval. ACM Press, New York, pp: 463.

Brin, S. and L. Page, 1998. The anatomy of a large-scale hypertextual web search engine. Comput. Networ. ISDN Syst., 30(1-7): 107-117.

Burgoon, J.K., D.B. Buller, L.K. Guerrero, W.A. Afifi and C.M. Feldman, 1996. Interpersonal deception: XII information management dimensions underlying deceptive and truthful messages. Commun. Monogr., 63(1): 50-69.

Chang, W., W. Chung, H. Chen and S. Chou, 2003. An International Perspective on Fighting Cybercrime. Proceeding of 1st NSF/NIJ Symposium on Intelligence and Security Informatics (ISI 2003). LNCS 2665, Springer-Verlag, pp: 379-384.

Chen, H., W. Chung, J.J. Xu, G. Wang, Y. Qin and M. Chau, 2004. Crime data mining: A general framework and some examples. IEEE Comput. Soc., 37(4): 50-56.

Debnath, S., P. Mitra and C.L. Giles, 2005. Automatic extraction of informative blocks from webpages. Proceedings of the 2005 ACM Symposium on Applied Computing, pp: 1722-1726.

Ding, W. and G. Marchionini, 1997. A study on video browsing strategies. Technical Report, University of Maryland at College Park.

Duda, R.O., P.E. Hart and D.G. Stork, 1995. Pattern Classification and Scene Analysis. 2nd Edn., John Wiley & Sons Inc., USA.

Fayyad, U.M. and R. Uthurusamy, 2002. Evolving Data Mining into Solutions for Insights. Comm. ACM, 58: 28-31.

Forman, G., 2003. An extensive empirical study of feature selection metrics for text classification. J. Mach. Learn. Res., 3: 1289-1305.

Hope, T., T. Nishimura and H. Takeda, 2006. An integrated method for social network extraction. Proceeding of the 15th International Conference on World Wide Web (WWW), pp: 845-846.

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

Hosseinkhani, J., S. Chaprut and H. Taherdoost, 2012b. Propose a framework for criminal mining by web structure and content mining. Int. J. Adv. Comput. Sci. Inf. Technol., 1(1): 1-13.

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

Jin, W., R.K. Srihari and H.H. Ho, 2007. A text mining model for hypothesis generation. Proceeding of the 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI), pp: 156-162.

Jin, Y., Y. Matsuo and M. Ishizuka, 2009. Ranking Companies on the Web using Social Network Mining. In: Ting, I.H. and H.J. Wu (Eds.), Web Mining Applications in E-Commerce and E-Services. Studies in Computational Intelligence, 172. Springer, Berlin/Heidelberg, pp: 137-152.

Krebs, V.E., 2001. Mapping networks of terrorist cells. Connections, 24(3): 43-52.

Peng, J. and S. Ji-Hua, 2010. A method of text classifier for focused crawler. J. Chin. Inform. Proces., 26: 92-96.

Sannella, M.J., 1994. Constraint satisfaction and debugging for interactive user interfaces. Ph.D. Thesis, UMI Order Number: UMI Order No. GAX95-09398., University of Washington.

Skillicorn, D.B. and N. Vats, 2007. Novel information discovery for intelligence and counterterrorism. Decis. Support Syst., 43(4): 1375-1382.

Tao, P., 2007. Research on topical crawling technique for topic- specific search engine. Ph.D. Thesis, Jilin University.

Tavel, P., 2007. Modeling and Simulation Design. AK Peters Ltd., Natick, MA.

Vrij, A., 2000. Detecting Lies and Deceit*:* The Psychology of Lying and the Implications for Professional Practice. Wiley, Chichester.

Xu, J.J. and H. Chen, 2005. CrimeNet Explorer: A framework for criminal network knowledge discovery. ACM Trans. Inform. Syst., 23(2): 201-226.

Yang, C.C. and T.D. Ng, 2007. Terrorism and crime related weblog social network: Link, content analysis and information visualization. Proceeding of IEEE International Conference on Intelligence and Security Informatics (ISI), pp: 55-58.

Zhou, D., R. Manavoglu, J. Li, C.L. Giles and H. Zha, 2006. Probabilistic models for discovering e-communities. Proceeding of the 15th International Conference on World Wide Web (WWW), pp: 173-182.