Proposed Architecture for Terrorist Web Miner

R.D. Gaharwar
Assistant. Professor
G. H. Patel Department of computer Science and Technology,
Sardar Patel University,
Vallabh Vidyanagar, India

D.B. Shah
Professor
G. H. Patel Department of computer Science and Technology,
Sardar Patel University,
Vallabh Vidyanagar, India

G.K.S. Gaharwar
Assistant. Professor
School of Business and Law,
Navrachana University,
Vadodara, India

ABSTRACT

Terrorist Web Mining generates the novel information which can be used by SNA tools for Terrorist Network Mining. We assume that information available on open source media such as media reports, press reports etc can be used to draw a sociogram of these terrorist organizations. These sociograms can be used to study the relationships and roles of these terrorist organizations. In this article we propose architecture for a Terrorist Web Miner (TWM) which will search different search engines for terrorist attack information and the found web pages will be parsed for useful links. These links of the web pages which have relevant information then these web pages are stored and rank will be allocated to them. This links can farther be used for TNM.

Keywords

Web mining, Terrorist Networks, Terrorist Network Miner (TWM), SNA, TNM

1. INTRODUCTION

Search Engines and Web Search

Nowadays web based search engines are sprouting like mushrooms after a rainfall. Some of the popular search engines are Google, Bing, Yahoo! Search, Ask, Wow, Dogpile, AOI Search, WebCrawler, MyWebSearch etc. The people all round the world uses search engines for getting the useful information. Number of people using search engines in day to day life is growing exponentially.These search engines respond to query immediately for which each of these search engines uses their own indexes to make the search faster. For producing effective search results search engines like Google, indexes the relevant Web sites, images, Web pages, Usenet news groups, content-based directories, and news sources. This indexing of different web sources is done using proprietary algorithms, which work based on the assumption that if a page is useful , other web pages covering the other related topics are likely to provide a link to it. [1]These engines provide the useful information as per the user requirements but if we want the information from multiple search engines then we have to do web mining.

Web Mining

Extracting the useful data or information from different web pages available on internet is called web mining, Web Mining is used to extract knowledge from heterogeneous, semi or unstructured data moreover in web mining data collection involves crawling through large number of targeted web pages. For the mining of such heterogeneous data a number of algorithms have been proposed over the past decade. Primary data can be classified into 3 types: usage data, page content and web page structure. Hence the web mining can be categorized into 3 types: Web structure mining, Web content and Web usage mining. Hence there in considerable difference in web mining and conventional data mining.This makes the task of web mining more difficult. Moreover as the web mining becomes a manual task the effectiveness of it starts depending on the expertise of the investigators. The result of web mining will change from person to person. Hence the consistency of the search will be affected.As we all know information is distributed on the Web among billions of pages that are stored on millions of servers all around the world. The users browse the internet and for getting the useful information; they surf from one page to the next on World Wide Web. If a crawler is meant just for gathering information, it will visit many sites for evaluation and mined in a central location whereas for the long term usage the crawler should fetch the web pages, for crawling and save them in a source whereas the Web is a dynamic entity changing at exponential rates. Hence Web crawlers are facing a need to keep web requests stay dynamic and updated whenever links and pages are updated, deleted, added or modified. The most significant use of crawlers is in the provision of search engines. Hence the main consumers of Internet bandwidth are these web crawlers that generate their indexes by getting web pages from different search engines.

2. CHALLENGES IN THE INFORMATION PROCESSING OF TERRORIST NETWORKS

Sparrow [4] has studied the application of social network analysis for criminal activity analysis. Sparrow has penned down three problems of criminal network analysis

1. Incompleteness – This is the most common property of terrorist network. These networks are covert hence the interaction amongst its members is minimum which leads to the possibility that the missing nodes and links will remain uncovered by the investigators. This minimum interaction helps to maintain their secrecy.
2. Fuzzy boundaries – The boundaries of terrorist networks cannot be firmly decided as there is a confusion about whom to include in network and whom to not.
3. Dynamic – the terrorist networks are constantly changing hence they are dynamic and active. Sparrow in his study focused on the waxing and fading potency of a link depending upon the time and the task at hand instead of focusing at the presence or absence of a link between two social actors.

Apart from above mentioned problems terrorist networks also posses few more problems:

1. Inconsistency: Different terrorist attacks may have similar/different terrorist organizations involved. Hence every time new information has to be mined and new linkage maps has to be created. This may challenge the consistency of the data collected.
2. Incorrectness: Terrorists may forge their identities to hide themselves from the law enforcement agencies. This may result into fake/ incorrect data entries in databases.
3. Mining tools: Terrorist data is spread on World Wide Web (WWW). Hence for mining this data from different web pages may require more complex and sophisticated web mining tools. The process of designing such tools may sometimes turn into time consuming and cumbersome task.
4. Unreliability: Terrorist networks are often incomplete and incorrect hence the information collected may prove unreliable sometimes. Such unreliable can no longer be used for terrorist network mining.

3. RELATED RESEARCH

A Long- ago the terrorist networks mining used to be done manually due to the unavailability of prominent techniques. Hence terrorist networks mining started becoming both time and effort consuming task. However nowadays Social Network Analysis (SNA) has emerged as an efficient method for understanding and destabilizing these terrorist networks. In past few years the SNA techniques and methodology has revolutionalised the planning and strategies for counter terrorism. [5]Hosseinkhan J. et al [6] has done work for developing a architecture for criminal Mining. This architecture divides the task of criminal Mining into two phases: URL Fetching and Content Mining. In first phase, the pages which are concerned with the targeted offence are fetched and in the second phase the content of pages is parsed & mined.

Xiang Y. et al [7] has penned down the experience in applying two different visualization techniques namely hyperbolic tree view and a hierarchical list, for a crime analysis application. They have developed a prototype user interface for Criminal Relationship Visualization called COPLINK. This tool is used for interactively fetching & visualizing the criminal relationship information. They have successfully created a hyperbolic tree and a hierarchical view to visualise criminal relationships but in their research, the authors have concentrated more on visualisation of data rather than collecting data.

Valdis E. Krebs [8] in one of his article named “Mapping Networks of Terrorist Cells” conducted an exclusive analysis of 9/11 terrorist attacks. He collected the information from press reports of Sydney Morning Herald, The Washington Post, etc. These press reports showed that there were total 19 hijackers. The media reports also revealed the other details of the hijackers like who called whom, the addresses shared by them, their associations with each other and information that they used the same frequency flier number, etc. Valdis used graph theory concepts like centrality principles to derive a linkage map of the associations of these hijackers from the above information showed in media reports. From the linkage map of hijackers Valdis successfully proved that highest links led to Mohammad Atta (the group’s leader) than to any other hijackers.Zhou D. et al [9] worked for discovering semantic community discovery in Social Networks (SNs) for which they proposed two generative Bayesian models. They used probabilistic modeling in combination with community detection in SNs. The authors assumed that most communications in SNs usually occurs by exchanging communication documents, such as posts on blogs, emails, messages or posts on message boards. All these carrier of communication documents are called communication document. They carried a community structure study of an SN by modeling the various communication documents among its social actors and the format of communication documents is taken as email because emails have valuable information regarding shared knowledge and the SN infrastructure.

Gaharwar R. et al [10] says that to carry out effective Terrorist Networking Mining and extract useful information for counter- terrorism and national security, all kinds of information about these terrorist organizations should be collected, classified, and analyzed. However these terrorist networks operate in a covert manner and their secrecy proves to be their biggest strength. Hence this information is not available in open source. However the press reports, emails, media reports etc can be used to collect the information about terrorist organizations to carry out SNA on these Terrorist Networks.By studying the literature review, we found that there is a need of Terrorist Web Miner (TWM) and we propose the architecture for the same. The following figure shows the proposed architecture of TWM

4. PROPOSED ARCHITECTURE

TWM will mine the information related to terrorist attacks from popular search engines like Yahoo! Search, Ask, Bing and Dogpile. Terrorist Web Mining will one by one fetch the search pages according to the investigator’s requirement. Investigator may enter the keyword/ search word which will be searched from the web pages available on internet. Terrorist Web Mining will use search results from these web pages.

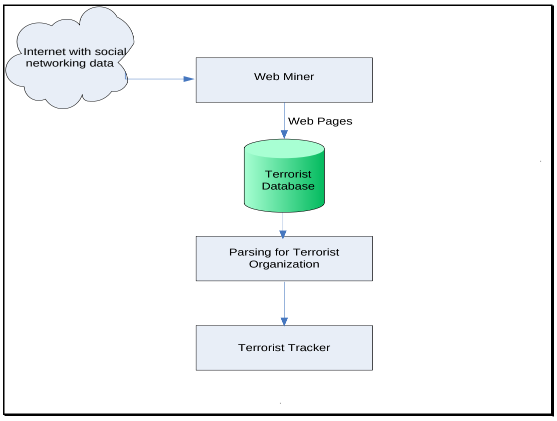
Internet with social networking data
Web Miner
Web Pages
Terrorist Database
Parsing for Terrorist Organization
Terrorist Tracker

Fig 1. The proposed architecture of TWM

Orthodox Information Retrieval (IR) system uses two types of documents: Conventional/Traditional text document and web pages. Web pages contain anchors tags and Hyperlinks along with simple text. These hyperlinks present in web pages are extremely important as they play a significant part in search ranking algorithms. On the other hand anchor tags may often give a more accurate explanation of the current web page. Therefore both of these; hyperlinks and anchor tags are important from information retrieval point of view. Another difference between a Conventional/Traditional text document and a web page is that a simple text document is rather structured whereas a web page is unstructured or semi structured. A Web page is not simply a collection few text paragraphs like in a conventional text document. A Web page has different fields, e.g., title, header, metadata, body, etc. Sometimes it is possible that the information contained in one field (e.g., the body field) can be more important than the information in other field (e.g., the title field). Furthermore, the content in a web page is typically arranged and presented in certain structured blocks. Some blocks are important and others may not be important (e.g., advertisements, privacy policy, copyright notices, etc.). Hence Web search engine has to detect the main content block(s) of a Web page effectively because fields appearing in such blocks are more important hence mining a web page is very different from mining any conventional textual document. [11]Hence this difficult task extracting the URL from the complete web page is done by the Terrorist Web Miner. This Miner will parse the complete fields and extract the URL from the web pages return by different search engines.

One of the major problems with the effectiveness of the search engines is Web Spamming. Every search engine indexes the web pages which are also called ranks of the web pages. If the web page has very low rank then the probability of that web page to appear as a search result decreases even if it is relevant to the user query. Search engines may illegitimately increase the rank of selected web pages. This is called web spamming. Hence to decrease web spamming the web miner will select only common links .The links that appeared in the search result of all the search engines will be filtered to increase the reliability of search and only the most common links returned by the search engines will be kept in the database to refer in future.

Once these common links are entered into database, the web pages that are relevant to these links are fetched. These web pages are then parsed to search the names of Terrorist organizations. Hence the links between Terrorist organizations can be found if the web pages have Terrorist organizations names in them.These links between Terrorist organizations can be farther given to Terrorist Tracker to show the relationships of these Terrorist organizations through the use of SNA and Graph Theory concepts.

5. ADVANTAGES OF THE SYSTEM

The architecture developed during this study will work towards automating the task of terrorist web mining. This system will automatically search different search engines for extracting useful links/URLs. This architecture will fetch the web pages related to these links and will automate the task of parsing the web pages for other related useful links also. In manual terrorist web mining extracting the useful links from heterogeneous web pages becomes a time consuming task. This system will automatically extract all the related links which can have the information related to terrorist attacks. Hence this system will eliminate all the problems present manual terrorist web mining.Moreover the web pages returned by most search engines will have maximum rank and only those web pages having highest ranks will be used for extracting the links of the terrorist organizations. Hence, this system will reduce the web spamming to minimum level by considering only those web pages which have maximum rank for future processing.

6. CONCLUSION

This study aims at creating a TWM which extracts the links of terrorist organizations. This TWM will work in two phases. During first phase the TWM will extract the web pages containing the information about terrorist attacks from different search engines and during second phase it will extract the useful URLs from these web pages. Hence this system will successfully extract useful links from the complex unstructured text data. This system will automate the complex task of collecting the information about terrorist networks.

This TWM can be integrated with Terrorist Web Visualization tools and SNA tools. SNA tool which uses graph theory concepts can use the information collected from the TWM to study terrorist networks and derive the linkage maps of the terrorist organizations. These maps can farther be used to study the relationship among terrorist network and destabilize these terrorist networks.

7. REFERENCES

[1] Sergey Brin and Lawrence Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine," in *7th World Wide Conference(WWW7)*, Brisbane, 1998.

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

[3] J Hosseinkhani, S Chuprat, and H Taherdoost, "Criminal Network Mining by Web Structure and Content Mining," in *Advances in Remote Sensing, Finite Differences and Information Security*, Prague, 2012, pp. 210-215.

[4] M K Sparrow, "The application of network analysis to criminal intelligence: An assessment of the prospects," *Social Networks 13*, pp. 251-274, 1991.

[5] U K Will, N Menon, and P Karampelas, "Detecting new trends in terrorism networks," *International Conference on Advances in Social Networks Analysis and Mining*, 2010.

[6] Javad Hosseinkhani, Suriayati Chaprut, Hamed Taherdoost, Morteza Harati, and Sadegh Emami Korani, "Criminal Communities Mining on the Web," *International Journal of Advanced Computer Science and Information Technology (IJACSIT)*, vol. 2, no. 2, pp. 13-22, April 2013.

[7] Yang Xiang, Michael Chau, Homa Atabakhsh, and Hsinchun Chen, "Visualizing criminal relationships: comparison of a hyperbolic tree and a hierarchical list," *Decision Support Systems*, pp. 69–83, 2005

[8] V E Krebs, "Mapping networks of terrorist cells," *Connections 24*, pp. 43–52, March 2001.

[9] D Zhou, R Manavoglu, J Li, C L Giles, and H Zha, "Probabilistic models for discovering e-communities," *15th international conference on world wide web (WWW)*, pp. 173–182, 2006.

[10] R D Gaharwar, D B Shah, and G. K. S. Gaharwar, "TERRORIST NETWORK MINING: ISSUES AND CHALLENGES," *International Journal of Advance Research In Science And Engineering*, vol. 4, no. 1, pp. 33-37, March 2015.

[11] Javad Hosseinkhani, Suriayati Chaprut, Hamed Taherdoost, and Amin Shahraki Moghaddam, "Propose a Framework for Criminal Mining by Web Structure and Content Mining," *International Journal of Advanced Computer Science and Information Technology*, vol. 1, no. 1, pp. 1-13, October 2012.