THE IMPLEMENTATION OF TERRORIST WEB MINER FOR THE TERRORIST NETWORKS

R. D. Gaharwar[1], D. B. Shah[2], G.K.S. Gaharwar[3]

[1,2]*G. H. Patel Department of computer Science and Technology, Sardar Patel University, Vallabh Vidyanagar, India*
[3]*School of Business and Law, Navrachana University, Vadodara, India*

Abstract

Terrorist web information is very valuable for Law enforcement agencies. This type of information is available on World Wide Web but is in fragments. These fragments of information are to be collected and analyzed for the terrorist network mining. This process when done manually may result into inefficient and error prone mining of data. Hence during this study the attempt is made to automate this difficult task of collecting, classifying and analysis of the terrorist web data. Developed Terrorist Web Miner will search terrorist information available on World Wide Web, it will fetch all the web pages related to terrorist web data and will extract the useful links. These links are stored into databases for future reference. This system can replace the manual system for collecting terrorist web data. The web pages related to terrorist attacks can be used in future to extract the links between terrorist organizations and hence a communication pattern, linkage patterns, role and relationship analysis can be derived for terrorist networks.

Keywords: Terrorist Networks, Terrorist Network Miner, SNA, TNM.

--***---

1. INTRODUCTION

In modern era, Social Network Analysis (SNA) has emerged as a novel tool for Terrorist Network Mining. SNA can be used effective and efficiently for Terrorist Network Mining only if the investigators are well equipped with adequate information about Terrorist Network. This process of information collection about terrorist networks is a difficult task as these networks work in a covert manner. For this we have to rely on media reports, press reports etc to collect such data from open source. For this task search engines can be used.

2. WEB MINING

The World Wide Web is a gigantic collection of different types of homogenous as well as heterogeneous documents. The web pages are themselves a collection of different types of tags like anchor tags, header tags etc and they may be human generated or machine generated. These documents can be in different languages i.e. Human or programming languages. The heterogeneous web documents can be a collection of text, html, images, sound clips, pdf files etc. The unique thing about web documents is that they also have meta information associated with them .i.e. information about information contained in these web documents. [1]

From the World Wide Web, investigator tries to focus only on those web documents which are related to particular topic and then carefully extract useful links from these web pages. This task can be performed by a web miner. Moreover a web miner detects the unusual patterns that exist amongst the web pages which initially seems unrelated.

Web mining is one of the data mining technique to dig out useful information from web which includes web documents, hyperlinks between documents, web site visited logs, etc. Internet has became an indispensable part of our lives now a days so the techniques which are helpful in extracting data present on the web is an interesting area of research. These techniques help to extract knowledge from Web document, in which at least one of structure or usage (Web log) data is used in the mining process (with or without other types of Web). According to analysis targets, web mining can be divided into three different types, which are Web usage mining, Web content mining and Web structure mining. [2]

3. TERRORIST WEB MINING

Terrorist Web mining is a challenge for detecting patterns of terrorists/ terrorist organizations from the vast collection of heterogeneous web pages available on internet. In a way terrorist web mining encompasses web mining as one has to mine all the data on the web as well as mine the structure and usage patterns. Information on the web includes structured text as well as unstructured text data. Therefore the tools developed for web mining apply for terrorist web mining also. In addition, we need tools to extract the structure of the terrorist network as well as the communication patterns.

Information about terrorists/ terrorist organizations is rarely available. This is because of that these terrorists/ terrorist organizations work in a secret manner. The complete terrorist network working depends on their secrecy. The terrorist networks are sometimes also referred as dark/covert networks because of their pattern of working in a secret manner.

4. RELATED RESEARCH

Valdis E. Krebs in "Mapping Networks of Terrorist Cells" described a connection between relationship of individual in Terrorist Networks and Data Sources for example Trust/Dependence relationship can only be confirmed amongst terrorists if terrorists have family contacts, has been in neighborhood, school, military or organization. This type of information can be gathered from public and court records and hence this information may be available in individual's native country only. Task/mission relationship can only be drawn from phone calls, electronic mail, chat messages, web site visits or board messages records. Travel records are also useful in collecting such information. Valdis built a matrix showing links between terrorists. He decided to draw linkage map of three strengths. Valdis also concluded that the binding strength amongst the terrorists will depend largely on the amount of time spend either at school, learning institutes, clubs, work place, or at training program. [3]

Xaing Y. et al have developed a prototype for criminal relationship visualize tool called COPLINKS. Authors have successfully developed an efficient, user friendly information analysis and sharing system. Authors have used data and information from various sources including law enforcement agencies data. Hence COPLINK database contains 2 million criminal case reports from the criminal departments. This visualizing tool creates hierarchical lists and hyperbolic tree from this database. After creating hierarchical lists and hyperbolic tree of criminal data authors have done empirical study to compare effectiveness and efficiency of both these views. During this empirical study they found out that both views can be useful in criminal relationship visualization. [4]

Javal. et. al. has created a high-level architecture for criminal mining, they have used a scalable universal crawler for this architecture. This crawler begins its procedure by listing all the unvisited URLs. Thereafter the pages correlated to these URLs are fetched and the URLs from the fetched web pages are added to the list. Moreover to upgrade the performance of this crawler Javal et. al. have used multithreaded architecture. Hence the request to web pages is loaded in different threads and all these threads will run simultaneously and the system performance is enhanced. [5]

Gaharwar R. et al [6]has mentioned that all the terrorist organizations interact to generate covert networks and they operate in a secret manner which is also their biggest strength. This secrecy help these terrorist organizations to work efficiently without being at the surface, Hence collection of information about such covert networks can be a challenge. However if the information from open source can be taken as authentic, applying effective data mining techniques can generate the useful information. Such information can be farther used for counter terrorist activities.

Ding Z. et al described the importance of semantic information on social networks compared to previous work in SNA where prime focus was on social network (SN) topology. Authors have suggested two different semantic community discovery models in social networks. These are Bayesian models which combine probabilistic modeling with community detection in social networks. The authors mainly focus on documents as the much of the communication in social networks (SN) is done through exchanging documents. These documents may be emails, blogs, instant massages, chat history, boards, posts etc. The authors have assumed that the users having related communication interests will be associated with their communication of a semantic society in SN. The communication is done through document called communication document which is generally taken as email. Hence emails are considered as carrier document of communication because they have valuable information regarding SN infrastructure and semantic shared knowledge. Here the authors study the semantics of SN. [7]

The following figure is the architecture of Terrorist Web Miner. The Terrorist Web Miner shown here will search on internet for social networking data of terrorist attacks in India. Once the web pages related to the terrorist attacks in India are fetched, these web pages will be parsed for extracting the links of the related web pages. These web pages related to Terrorist Networks will be stored in Terrorist Database. [8]
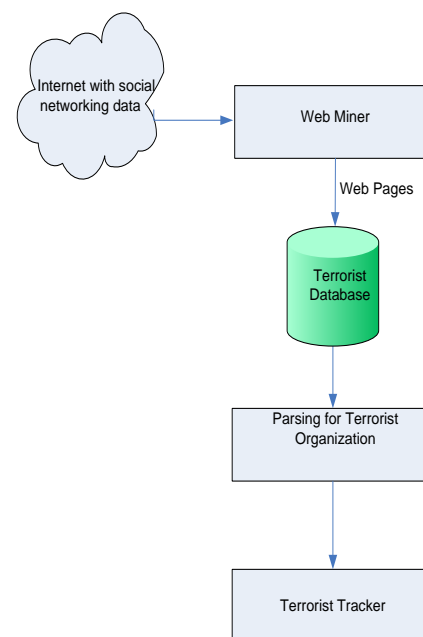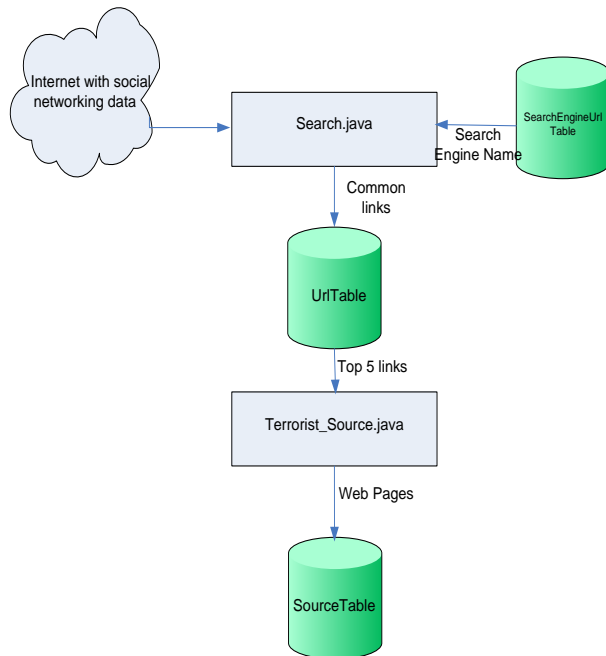

Figure 1: the architecture of Terrorist Web Miner

Terrorist Database contains all the web pages which farther will be used to extract terrorist links among terrorist organizations. These web pages will be parsed to extract the terrorist links. The terrorist links can be used to create a linkage map of any terrorist attack. Linkage map can also be used to study the relationships to each organization/terrorist in the network as well as its can be used to study the role of to each organization/terrorist in the network.

5. IMPLEMENTATION OF TERRORIST WEB MINER

The following figure shows the implementation of Terrorist Web Miner. This web miner is developed using Java and MySql database.

Figure 2: the architecture of Terrorist Web Miner

There are 3 tables in Terrorist Tracker Database namely SearchEngineUrl, UrlTable and SourceTable. SearchEngineUrl Table contains the URLs of the search engines from which investigator wants to search results and may contain URLs of popular search engines like Yahoo! Search, Ask, Bing and Dogpile etc. Each of these search engine uses its own internal indexing system to rank the page for making searching effective.

The task of mining terrorist networks is divided into two steps:

Step 1: Extraction
There is a java program named Search.java that will run the search on all the search engines stored in SearchEngineUrl Table. Search.java will have a search keyword for example "Terrorist attacks in India". This search keyword will be used to the useful web pages.

As a output of this keyword, these search engines will return number of links contain in a web page. Extracting these links from the web pages is a difficult task because a web page is a collection of header tags, anchor tags, links etc. Hence, parsing a web page is far more difficult than parsing a text document. Search.java will parse the web page returned by these search engines as a result of search. The links from the web page will be extracted and these links will be stored in UrlTable table along with their ranks. Ranks are allocated by the Terrorist Web Miner. As the frequency of the link URL increases its rank also increases

i.e. In UrlTable table, ranks of the web pages changes dynamically. Hence the URL of the link that appears in almost all search engines will have highest rank.

Step 2: Filtration
In search, investigator may face the problem of Web Spamming. Web Spamming means increasing the rank of any page illegitimately. If the page have low rank it will not be returned as a search result irrespective of whether it has appropriate content or not. Hence to reduce this web spamming only 5 highest rank URLs are selected for farther study by Terrorist _Source.java program. This program will fetch the contents of the top 5 ranked URLs/ links stored in UrlTable table. If the content of these links/URLs contains the name Terrorist Organization then this web page is stored in SourceTable table. Hence SourceTable table will contain only those web pages which have Terrorist Organization information in them. Later on these web pages can be used to study the relationship between different Terrorist Organizations.

6. EVALUATION

The following table shows the content of UrlTable table in MySql Database. UrlTable table have fields like SiteUrl and Rank. As described in implementation, Search.java will look for Terrorist attack information on popular search engines like Yahoo! Search, Ask, Bing and Dogpile etc. This program will parse the contents of the web pages and will look for the useful links. When the links is found, it is added in this table and if this link already exists in this table then simply rank is increased. This will reduce the level of the web spamming. When UrlTable table is ready Terrorist_Source.java will select top five links from this table. These top 5 links will be mined to get all the others links related to this topic because in web mining we assume that if current web pages are useful then other pages covering the related topic are likely to provide a like to it. Hence the implementation of Terrorist Web Miner is successfully done.

Table: UrlTable

Sr No	SiteUrl	Rank
1	www.satp.org/satporgtp/countries/india/database/index.html	5
2	www.ndtv.com/topic/india-terror-attacks	5
3	www.theguardian.com/world/mumbai-terror-attacks	4
4	en.wikipedia.org/wiki/Terrorism_in_India	3
5	www.india.com	3
6	en.wikipedia.org/wiki/List_of_terrorist_incidents_in_India	2
7	www.ibtimes.com/major-terrorist-attacks-india-over-last-20-years-timeline-1752731	2
8	www.chacha.com	2
10	en.wikipedia.org/Terrorist_incidents_in_India	2

CONCLUSION

The counter terrorism war can be fought and won only by superior knowledge and effective SNA tools. Terrorist Networks can be destabilized only when we have enough knowledge about terrorist networks and its topologies. Terrorist Web Mining aids in searching, extracting and collecting this knowledge. During this study we successfully created a terrorist web miner which crawls through different search engines and returns the related web pages. These web pages are then parsed for extracting the terrorist links from the web pages which may contain information about terrorist organizations. This links can be farther stored to create linkage map of terrorist networks.

REFERENCES

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

[2]. Monika Yadav and Pradeep Mittal, "Web Mining: An Introduction," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, no. 3, pp. 683-687, March 2013.

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

[4]. Yang Xiang, Michael Chau, Homa Atabakhsh, and Hsinchun Chen, "Visualizing criminal relationships: comparison of a hyperbolic tree and a hierarchical list," Decision Support Systems 41, pp. 69–83, 2005.

[5]. Javad Hosseinkhani, Suriayati Chaprut, Hamed Taherdoost, Morteza Harati Cool, 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.

[6]. 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, 2015.

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

[8]. R. D. Gaharwar, D. B. Shah, and G.K.S. Gaharwar, "Proposed Architecture for Terrorist Web Miner," International Journal of Computer Applications, vol. 128, no. 9, pp. 18-20, October 2015.