Communications on Applied Electronics (CAE) – ISSN : 2394-4714
Foundation of Computer Science FCS, New York, USA
Volume 4– No.6, March 2016 – www.caeaccess.org

A Survey of the Factors of Optimal Noise Reduction Algorithm for Terrorist Web Mining

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
Over the past few decades there have been frequent terrorist attacks around the world including India. This article describes Terrorist Network Mining and problems faced during studying such networks. A major challenge faced by the law enforcement agencies is the large crime 'raw' data volumes and the lack of sophisticated network analysis tools and techniques to utilize the data effectively and efficiently. This article states different data collection techniques used for terrorist networks. The major challenge is the development of the optimal noise reduction algorithm which will help in creating accurate linkage map of terrorist network without the loss of any key player node. This article successfully lists the factors that can be taken under consideration during generation of optimal noise reduction algorithm for Terrorist Web Mining. Once the accurate linkage map is generated the identification and removal of the key player for the destabilization of terrorist networks will become lot easier.

Keywords
Social Network Analysis, Terrorist Networks, Terrorist Network Mining

1. INTRODUCTION
The expansion of World Wide Web has lead to the accumulation of huge amount of information on internet. The number of people accessing internet for sending/receiving information is also increasing exponentially. People feel connected to each other through the use of internet but having huge amount of information makes it almost impossible to dig correct information without the help of effective algorithms. Web mining techniques can be used by internet users who want specific information.

1.1 Web Mining
Web Mining is a technique used for identifying the patterns from the web data which initially seemed unrelated. According to Bhatia [1] web mining is meant to extract information/ knowledge from web pages. During this extraction process the useful patterns are discovered from the information. The author described web mining process having steps like resource search, data selection, generalization and analysis.

Bhatia categorized the web mining process into 3 categories like Web Content Mining, Web Structure Mining and Web Usage Mining.

Web Content Mining: Web Content Mining focuses on the extraction of useful information in from the actual text of web documents. Web pages can be unstructured or semi-structured. Web Content mining process does the difficult task of extracting important information from web document.

Web Structure Mining: Web Structure Mining concentrates on the structural composition of the web documents. Web page is different from other normal page as it has different types of tags in it. [2] The work of the Web Structure Mining is to find the linkage structure of the hyperlinks in the web documents.

Web Usage Mining: Web Usage Mining is the process of identifying the navigational patterns from the user's browsing behavior. Web server logs, browser history, navigational information etc can be studied for Web Usage Mining.

1.2 Terrorist Network Mining
Using web mining techniques and tools to dig the linkage information about terrorists/terrorist organizations can be referred as Terrorist Network Mining.

Gaharwar et al. [7] described Terrorist Network Mining as a novel research field which can be used to investigate organized crime. Terrorists/Terrorist organizations communicate using internet to accomplish their inhuman activities. Hence internet becomes repository of terrorist information too. The terrorist organizations can be considered as nodes and their communication relationship can be considered as link joining these nodes. The linkage map can be created showing the relationship behavior of the terrorist organizations. Such information about terrorist networks can be used for the detailed study of these networks and destabilization of such dark networks.

1.2.1 Shortcoming of Terrorist Network Mining
Terrorist Network Mining is a novel research area that can be used for digging the information about terrorist networks but there are some factors that can make the Terrorist Network Mining ineffective. These factors are as follows:

1) Lack of related data
2) Lack of proper data mining tools
3) Lack of effective noise reduction algorithms

Lack of related data: Terrorist Network Mining intends to use the data from internet. Internet is an ocean of information. Extraction of the relevant and useful information about

terrorists/terrorist organizations from internet would take lot of efforts. The information collection of terrorists/terrorist organizations from web is done with the help of Terrorist Network Mining tools. These tools can be used to eavesdrop on all traffic of web sites, showing any association with any terrorist/ terrorist organizations to get the IP address.[3] But again the problem stands as these IP addresses are not static.

Lack of proper data mining tools: Nowadays different types of quantitative and analogical models can be used for terrorist network mining like IDS (Intrusion Detection System), Vector Space method (VSM), Clustering techniques etc. These techniques are used for the detection of terrorist-related activities on the network. However there are some factors that can decrease the efficiency of these tools. These factors are as follows:

1) Large-quantity of data to be in analysis to gain small piece of useful information
2) Incompleteness of the terrorist related information
3) Dynamic behavior of the terrorist networks makes it very difficult to get static information of such covert networks. [4]

Due to all the above stated factors building effective data mining tool becomes highly difficult.

Lack of effective noise reduction algorithms: From the huge data available on internet only a small portion of the information is used for further processing. This data can be filtered later on to make it more useful. Noise reduction techniques can help in the process of data mining to give more accurate results. Noise reduction algorithms used nowadays are less effective for Terrorist Web Mining. Hence there is a need to develop new noise reduction algorithm to generate precise information.

2. RELATED RESEARCH

Mishra et al. [5] studied web traffic for Topic Sensitive Link Analysis. The authors analyzed various orthodox ranking methods like PageRank, Weighted PageRank and Topic Sensitive PageRank.

The authors proposed an architecture using graph theory concepts. Graph is created considering every node as a category and the links between each node as a subcategory which also assigned some weight based on the fact that every web document is labeled with a specific category. When any page belongs to any specific category its relevance/importance will decide its rank. For any search engine the final rank to the page will be assigned after combining search engine's ranking and the system's ranking techniques. In the next step Weighted PageRank vectors according to the ranks already assigned to the web pages from Open Directory Project (ODP).

When user fires any query, only highlighted words in the query are collected because a single web page may have different topics in it. For each highlighted word an importance score is computed which is later on combined with the importance score of the other highlighted words to generate composite PageRank score. Different scoring schemes are used to generate final ranks with respect to user query. Hence the authors proposed a scalable algorithm for link analysis.

Lal et al. [6] proposed a novel algorithm for semantic web based on data cloud. The authors explained different types of cloud computing modes like SOMF (Service-Oriented Modeling Framework) which encompasses the entity models like IaaS (Infrastructure-as-a-Service), PaaS (Platform-as-a-Service) and SaaS (Software-as-a-Service).

The authors proposed data mining scheme which uses the sector/sphere framework in combination with associative rules. The proposed model is divided into following 3 steps:

1) Creating large optional word sets
2) Filtration of the large optional word set
3) Discovering the association rules

During first step, database is scanned and user query is kept in the word sets. As authors keep on adding the query words to word set, it keeps on growing in size. Once a large optional word set is generated filtration is carried which leads to second step. Filtration process is done by calculating the support of each word in database, depending on support the confidence of that word will be calculated from the database by assigning the probabilities. If the probability is too high, word can be filtered out from the original database. This process reduces the size of the original word set.

During the last step, the appropriate association rules are applied. Due to the applications of associative rules this algorithm tends to avoid the redundant rules and thus improves the efficiency of the algorithm.

Gaharwar et al. [7] has described graph theory concepts that can be used for SNA (Social Network Analysis). The authors considered social network of internet user as sociogram where each node is one of the user and communication connection between users is considered as links between the nodes. Once sociogram of the social actors is created successfully, different graph theory concepts such as degree, betweeness, prestige, closeness etc can be applied to the sociograms. Each of these centrality principles has different implications on graph for example betweeness means that the number of paths that join two nodes, passes through any given node. Hence if the betweeness index of the node is high, it can be considered as a broker node. If any of the nodes has high degree of index, it means that the node is very active in network. Hence these centrality principles can be useful in understanding the working of any social network including terrorist networks. The role of each actor in the social network can be identified easily by correct application of these centrality principles.

Valdis E. Krebs studied 9/11 attacks very closely. He wrote an article named "Mapping Networks of Terrorist Cells" based on his study. Valdis stated that despite of the fact that the actors on the social network may belong to same network; there are many different types of relationships amongst individual actors. The relationship between individuals can be dependency relationship which is built only if two individuals in the network are associated with each other very closely. The other type of relationship can be of trust. This relationship is built between individuals in the network if they have already worked with each other before. The mission relationship develops between individual if they are not very much in touch with each other. Valdis created linkage map of the individuals involved in the 9/11 terrorist attacks. This linkage map was analyzed against different strengths of relationships. In the end Valdis successfully concluded that

most of the links lead to the mastermind of 9/11 terrorist attacks. [8]

Elovici et al. [9] used vector space model for the purpose of Information retrieval from World Wide Web. The process of fetching only the useful information/knowledge from internet is known as Information retrieval process. The authors created n-dimensional vectors of the web page content. The similarities between any two web pages are computed using Cosine/Euclidean distance. The Euclidean distance finds the similarity between terrorist topics from web page and accessed web page. Hence the terrorist topics of interest are showed as n-dimensional vectors.

3. TECHNICAL CHALLENGES

Internet users are increasing exponentially every year and so is the web traffic. To investigate the terror traces on web traffic enormous amount of information is to be analyzed. Even if such vast web traffic is analyzed the interpretation and representation of the analyzed summary is a difficult task. Hence some filtration is necessary to give concise output. The main difficulty with the filtration technique is that not much of the research is been carried out about this issue. Moreover the recent technique used by most researchers in the field of Terrorist Network Mining is the simple filtering technique.

3.1 Problems with Simple Filtration technique

Terrorist networks can be studied using the different Social Network Analysis centrality principles. Different nodes in terrorist networks have different type and different level of involvement. These centrality principles not only play key role in the analysis of terrorist networks but they are also very sensitive to every small change in network. [10] Krebs has successfully discovered the shortcoming of the covert/dark networks. Krebs stated sometimes even the stronger ties may have low frequency of activation and because of this stronger ties appear weak. Simple Filtration technique will cut off all the nodes which have low frequency/weak ties. Hence sometimes it will discard stronger ties in this process.

Moreover the key node/leader node does not communicate more with other nodes and hence it appears as less active node. Such less active node is difficult to find with Simple Filtration technique. Hence there is a need of new optimized noise reduction techniques which may consider various critical parameters of the terrorist networks.

4. PROPOSED METHODOLOGY

A lot of study has been carried out in the field of data mining of terrorist networks. Different kinds of algorithms like vector space, clustering algorithms etc are developed and customized for this purpose. The information so collected can be used further for understanding the terrorist networks deeply only after proper filtrations. The noise reduction techniques that are used in most of the existing research are the simple filtering technique.

The list of the factors that can be used as critical parameters in the noise reduction algorithm are as follows:

1) Number of causalities
2) Freshness of attacks
3) Links with other key nodes in the networks
4) Headquarter epicenter

Number of causalities: If the node is involved in attacks which had less causality then such nodes can be considered as weaker. The aim of the nodes in covert network is to plan an activity that can cause high number of causalities. If the number of causalities is high the node was trusted with important mission hence the strength of the node in the network also increases.

Freshness of attacks: The biggest dilemma the researchers face is deciding the timeline of the data collection. The researcher has to restrict his/her research boundaries. If the data used in the research is barbaric, it may prove to be irrelevant for the research in long run. Terrorist network are decided based on their freshness. The more they are fresh, the more they will be involved in current attacks. If any terrorist organization has been inactive for a long time it may be considered as a weaker node.

Links with other key nodes in the networks: Generally if any node in the network has less frequency it is considered as a weaker node. Although it might be possible that a node having less frequency can be a key node and it is using other nodes to complete its mission. In such a situation the leader node would appear as a weaker node. During noise reduction if such node is eliminated, then one of the key players would be lost. Hence rather than just eliminating any node based on its frequency, its linkage with other key nodes should also be considered. This will help in successfully indentifying the role of key players in the terrorist networks.

Headquarter epicenter: Here authors assume that terrorists are trained before assigning them a mission. This training program is generally held at headquarter of the terrorist organization. Different terrorist organizations are related through motto, goals etc. The related terrorist organization tends to share headquarters, exchange the individuals etc. If any weaker node is related to stronger node in network then there are chances that it was also involved in the mission or it may be held responsible for any future mission. Hence if the terrorist organizations are related to others via headquarter they both can be considered as stronger nodes.

These are factors that may decide the strength and weakness of the nodes in the terrorist networks. Hence these factors should be taken into consideration while designing optimal noise reduction algorithm. Here a primitive design of optimal noise reduction algorithm is proposed by assigning '1' as a probability to each factor and computing the strength.

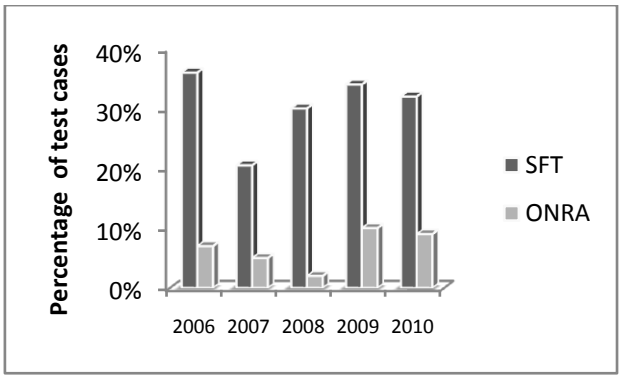

Fig 1: Comparision of percentages of the test cases using Simple Filtration Technique and Optimal Noise Reduction Algorithm.

The proposed Optimal Noise Reduction algorithm (ONRA) takes the above critical factors under consideration for efficient filtration of information related to terrorist networks. Let the importance of any node be represented as Critical Relational Factor (CRF).

CRF = $\sum p_i x_i$, where i=1,…,n and x_i be any of the above critical parameters and p_i represents the weight assigned to respective critical parameter.

Numbers of test cases for different years have been filtered using both Simple Filtration Technique (SFT) and proposed Optimal Noise Reduction Algorithm (ONRA) which shows that proposed Optimal Noise Reduction Algorithm gives better filtration results than the Simple Filtration Technique. The following graph shows the percentage of the test cases where important nodes are filtered by the use of Simple Filtration Technique (SFT) and proposed Optimal Noise Reduction Algorithm (ONRA).

5. CONCLUSION

The current paper not only describes what terrorist network mining is? But also tries to explore the key problems faced during terrorist network mining. Terrorist networks are covert/dark networks having secrecy as their key strength. Due to this fact collecting the data for studying such networks becomes difficult task. However this paper explains Topic Sensitive Link Analysis which can be used for the data collection process of Terrorist Networks. The relevant data can also be collected through different types of cloud computing models. The collected data is processed and filtered to create linkage map of the Terrorist Networks using Social Network Analysis tools. These linkage maps can easily be studied by applying graph theory concepts like centrality principles, key players of the network can be identified and destabilization of such networks becomes easier. This research presents a list of factors which can be used in the optimal noise reduction algorithm. These factors affect the strength of the node in the terrorist networks. This research work justifies the effectiveness of each factor on the strength/weakness of the node by designing a primitive optimal noise reduction algorithm which takes under consideration the factors and testing it against previous algorithm. The results clearly show that this new optimal noise reduction algorithm performs better than previous work. This primitive optimal noise reduction algorithm can be fined tuned in future by assigning more appropriate probabilities to each of the factors.

6. REFERENCES

[1] T. Bhatia, "Link Analysis Algorithm For Web Mining", *International Journal of Computer Science And Technology*, Vol. 2, Issue 2, June 2011, pp. 243-246.

[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, March 2013, pp. 683-687.

[3] N. Chaurasia, A. Tiwari, "Novel Algorithm for Terrorist Network Mining", *International Journal of Computer Science And Communication Technologies*, Vol. 6, no. 1, July 2013, pp. 898-903.

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

[5] S. N. Mishra, A. Jaiwal, A. Ambhaikar, "An Effective Algorithm for Web Mining Based on Topic Sensitive Link Analysis", *International Journal of Advance Research in Computer Science and Software Engineering*, Vol. 2, Issue 4, April 2012, pp. 278-282.

[6] K. Lal, N. C. Mahanti, "A Novel Data Mining Algorithm for Semantic Web Based Data Cloud", *International Journal of Computer Science and Security*, Vol. 4, Issue 2, 2010, pp. 160-175.

[7] 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, 2015, pp. 33-37.

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

[9] Y. Elovici, A. Kandel, M. Last, B. Shapira, O. Zaafrany, "Using Data Mining Techniques for Detecting Terror-Related Activities on the Web", *Proc. Second Int'l Conf. on Computer and Electrical Engineering (ICCEE 2009)*, 2009, pp. 152-157

[10] V. E. Krebs, (2002), "Uncloacking Terrorist Networks", First Monday, Vol. 7, 2002, pp. 4 - 1.