Terror Tracking Using Advanced Web Mining Perspective

T.Anand
Professor & Head,
Dept of Computer Science & Engg
AVCCollege of Engg
Mayiladuthurai,
India.
anand_thi@hotmail.com

S.Padmapriya
Assistant Professor,
Dept of Information Technology
AVCCollege of Engg
Mayiladuthurai,
India.
padmanand02@yahoo.co.in

E.Kirubakaran
Senior D.GeneralManager,
Bharat Heavy Electrical Ltd.
Trichy.
India.
e_kiru@ yahoo.co.in

Abstract— Web mining is a rapidly growing research area. It consists of Web usage mining, Web structure mining, and Web content mining. Web usage mining refers to the discovery of user access patterns from Web usage logs. Web structure mining tries to discover useful knowledge from the structure of hyperlinks. Web content mining aims to extract/mine useful information or knowledge from web page contents. Web mining techniques can be used for detecting and avoiding terror threats caused by terrorists all over the world.

Keywords- usage mining; structure mining; content mining and patterns

I INTRODUCTION

Web mining - is the application of data mining techniques to discover patterns from the 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. There are about 5,000 terrorist web sites as of 2006 and 50,000 sites of extremist and terrorist content as of 2007, including: web sites, forums, blogs, social networking sites, video sites, and virtual world sites (e.g., Second Life) in more than 15 languages.

II EXPLANATION

A. Web content mining

Web content mining is related but different from data mining and text mining. It is related to data mining because many data mining techniques can be applied in Web content mining. It is related to text mining because much of the web contents are texts. However, it is also quite different from data mining because Web data are mainly semi-structured and/or unstructured, while data mining deals primarily with structured data. Web content mining is also different from text mining because of the semi-structure nature of the Web, while text mining focuses on unstructured texts. Web content mining thus requires creative applications of data mining and/or text mining techniques and also its own unique approaches. In the past few years, there was a rapid expansion of activities in the Web content mining area. This is not surprising because of the phenomenal growth of the Web contents and significant economic benefit of such mining. However, due to the heterogeneity and the lack of structure of Web data, automated discovery of targeted or unexpected knowledge information still present many challenging research problems

Web information integration and schema matching- Although the Web contains a huge amount of data, each web site (or even page) represents similar information differently. How to identify or match semantically similar data is a very important problem with many practical applications. Some existing techniques and problems are examined.

Data/information extraction-Our focus will be on extraction of structured data from Web pages, such as products and search results. Extracting such data allows one to provide services. Two main types of techniques, machine learning and automatic extraction are covered.

Opinion extraction from online sources- There are many online opinion sources, e.g., customer reviews of products, forums, blogs and chat rooms. Mining opinions (especially consumer opinions) is of great importance for marketing intelligence and product benchmarking.

Knowledge synthesis- Concept hierarchies or ontology are useful in many applications. However, generating them manually is very time consuming. A few existing methods that explores the information redundancy of the Web will be presented. The main application is to synthesize and organize the pieces of information on the Web to give the user a coherent picture of the topic domain.

Segmenting Web pages and detecting noise- In many Web applications, one only wants the main content of the Web page without advertisements, navigation links, copyright notices. Automatically segmenting Web page to extract the main content of the pages is interesting problem. A number of interesting techniques have been proposed in the past few years.

978-1-4244-4711-4/09/$25.00 ©2009 IEEE

B. Web Structure Mining

World Wide Web can reveal more information than just the information contained in documents. For example, links pointing to a document indicate the popularity of the document, while links coming out of a document indicate the richness or perhaps the variety of topics covered in the document. This can be compared to bibliographical citations. When a paper is cited often, it ought to be important. By means of counters, higher levels cumulate the number of artifacts subsumed by the concepts they hold. Counters of hyperlinks, in and out documents, retrace the structure of the web artifacts summarized.

C. Web Usage Mining

Web servers record and accumulate data about user interactions whenever requests for resources are received. Analyzing the web access logs of different web sites can help understand the user behavior and the web structure, thereby improving the design of this colossal collection of resources. There are two main tendencies in Web Usage Mining driven by the applications of the discoveries: General Access Pattern Tracking and Customized Usage Tracking. The general access pattern tracking analyzes the web logs to understand access patterns and trends. These analyses can shed light on better structure and grouping of resource providers. Many web analysis tools exists but they are limited and usually unsatisfactory. Applying data mining techniques on access logs unveils interesting access patterns that can be used to restructure sites in a more efficient grouping, pinpoint effective advertising locations, and target specific users for specific selling ads. Customized usage tracking analyzes individual trends. Its purpose is to customize web sites to users. The information displayed, the depth of the site structure and the format of the resources can all be dynamically customized for each user over time based on their access patterns. While it is encouraging and exciting to see the various potential applications of web log file analysis, it is important to know that the success of such applications depends on what and how much valid and reliable knowledge one can discover from the large raw log data.

Current web servers store limited information about the accesses. Some scripts custom tailored for some sites may store additional information. However, for an effective web usage mining, an important cleaning and data transformation step before analysis may be needed.

D. Terrorism and control through web mining

The terrorists operate in web through:

Web sites -As many as 50,000 terrorists web sites were in existence across the web

Forums- There are about 300 terrorist forums which includes more than 30,000 members with close to 1,000,000 messages posted.

Blogs, social networking sites, and virtual worlds- Blogs and social networking sites, mostly hosted by terrorist sympathizers.

Videos and multimedia content- Terrorist sites are extremely rich in content with heavy usage of multimedia formats containing about 1,000,000 images and 15,000 videos from many terrorist sites and specialty multimedia file-hosting third-party servers.

E. Methodology for countering terrorism

The computational tools are grouped in two categories:

i. *Collection*

ii. Analysis and Visualization.

iii. Collection-

Web site spidering- Spiders/crawlers are based on previous digital library research. Spiders can access password-protected sites and perform randomized (human-like) fetching. Spiders are trained to fetch all html, pdf, and word files, links, PHP, CGI, and ASP files, images, audios, and videos in a web site. To ensure freshness, selected web sites must be searched for every 2 to 3 months.

Forum spidering- The forum spidering tool recognizes forum hosting software and their formats. by collecting the complete forum including: authors, headings, postings, threads, time-tags, etc., which allow us to re-construct participant interactions and by processing forum contents in Arabic, English, Spanish, French, and Chinese using selected computational linguistics techniques.

Multimedia (image, audio, & video) spidering-

Specialized techniques are in existence for spidering and collecting multimedia files and attachments from web sites and forums using stenography to identify encrypted images .

ii. Analysis and Visualization:

Social network analysis (SNA)- Various SNA techniques can be used to examine web site and forum posting relationships using topological metrics (betweeness, degree, etc.) and properties (preferential attachment, growth, etc.) to model terrorist and terrorist site interactions. We have developed several clustering (e.g., Blockmodeling) and projection (e.g., Multi-Dimensional Scaling, Spring Embedder) techniques to visualize their relationships by understanding “Dark Networks” (unlike traditional “bright” scholarship, email, or computer networks) and their unique properties (e.g., hiding, justice intervention, rival competition, etc.).

Content analysis- Several detailed (terrorism-specific) coding schemes can be developed to analyze the contents of terrorist and extremist web sites. Content categories include: recruiting, training, sharing ideology, communication, propaganda, etc and computer programs to help automatically identify selected content categories (e.g., web master information, forum availability, etc.).

Web metrics analysis- Web metrics analysis examines the technical sophistication, media richness, and web interactivity of extremist and terrorist web sites by examining technical features and capabilities (e.g., their ability to use forms, tables, CGI programs, multimedia files, etc.) of such sites to determine their level of “web-savvy-ness.” Web metrics provides a measure for terrorists/extremists’ capability and resources. All terrorist site web metrics are extracted and computed using computer programs.

Sentiment and affect analysis- Not all sites are equally radical or violent. Sentiment (polarity: positive/negative) and affect (emotion: violence, racism, anger, etc.) analysis allows one to identify radical and violent sites that warrant further study. We also examine how radical ideas become “infectious” based on their contents, and senders and their interactions by recent advances in Opinion Mining – analyzing opinions in short web-based texts.

Authorship analysis and write print technique-Grounded in authorship analysis research, for identifying anonymous senders based on the signatures associated with their forum messages by expanding the lexical and syntactic features of traditional authorship analysis to include system (e.g., font size, color, web links) and semantic (e.g., violence. racism) features of relevance to online texts of extremists and terrorists. We have also developed advanced Inkblob and Writeprint visualizations. The Writeprint technique to help visually identify web signatures has been developed for Arabic, English, and Chinese languages. The Arabic Writeprint consists of more than 400 features, all automatically extracted from online messages using computer programs.

Video analysis-A significant portion of our videos are IED related. Based on previous terrorism ontology research, we have developed a unique coding scheme to analyze terrorist-generated videos based on the contents, production characteristics, and meta data associated with the videos. We have also developed a semi-automated tool to allow human analysts to quickly and accurately analyze and code these videos.

Dark Web analysis-A smaller number of sites are responsible for distributing a large percentage of IED related web pages, forum postings, training materials, explosive videos, etc and unique signatures can be used for those sites based on their contents, linkages, and multimedia file characteristics. Much of the content needs to be analyzed by military analysts. Training materials also need to be developed for troops before their deployment.

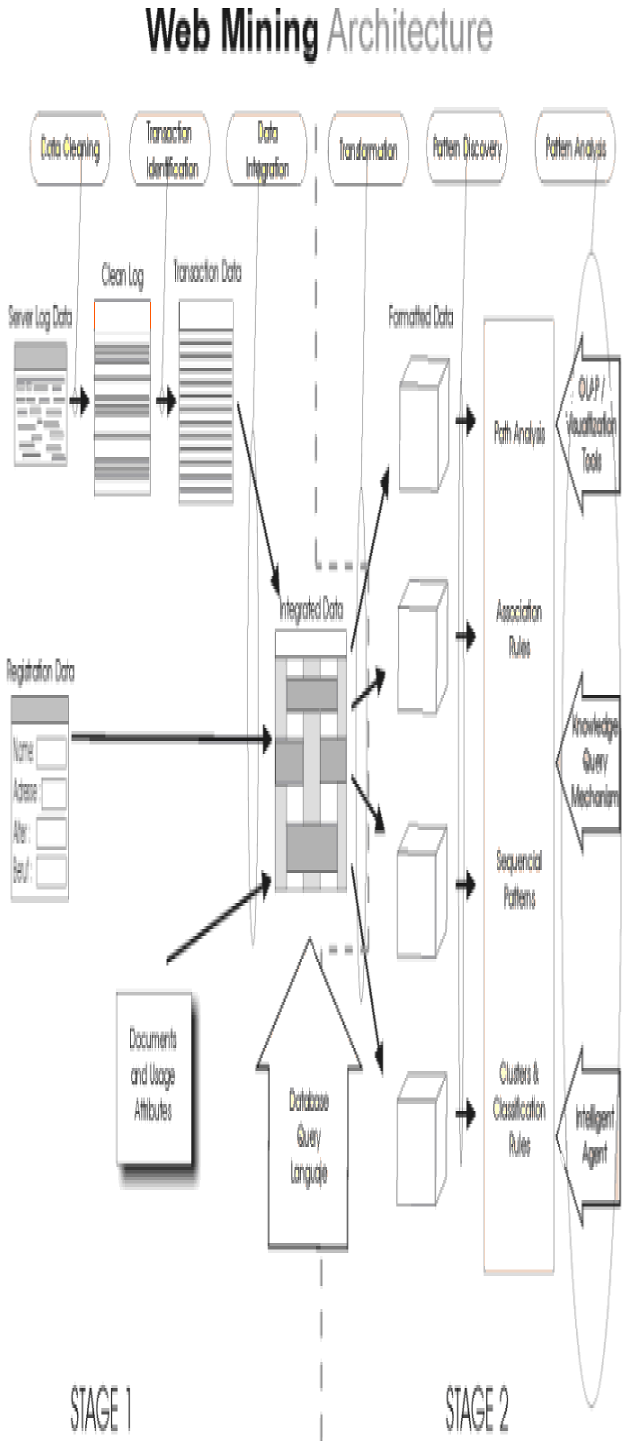

Figure 1 - Web Mining Architecture

III CONCLUSION

Thus, Web mining techniques can be used for detecting and avoiding terror threats caused by terrorists all over the world using the computational tools which is grouped under the two categories viz., Collection and Analysis.

REFERENCES

[1] H. Chen and C. Yang (Eds.), “Intelligence and Security Informatics,” Springer, forthcoming, 2008.

[2] H. Chen, E. Reid, J. Sinai, A. Silke, and B. Ganor (Eds.), “Terrorism Informatics: Knowledge Management and Data Mining for Homeland Security,” Springer, forthcoming, 2008.

[3] H. Chen, T. S. Raghu, R. Ramesh, A. Vinze, and D. Zeng (Eds.), “Handbooks in Information Systems -- National Security,” Elsevier Scientific, 2007.

[4] C. Yang, D. Zeng, M. Chau, K. Chang, Q. Yang, X. Cheng, J. Wang, F. Wang, and H. Chen. (Eds.), Intelligence and Security Informatics, Proceedings the Pacific-Asia Workshop, PAISI 2007, Lecture Notes in Computer Science (LNCS 4430), Springer-Verlag, 2007.

[5] S. Mehrotra, D. Zeng, H. Chen, B. Thursaisingham, and F. Wang (Eds.), Intelligence and Security Informatics, Proceedings of the IEEE International Conference on Intelligence and Security Informatics, ISI 2006, Lecture Notes in Computer Science (LNCS 3975), Springer-Verlag, 2006.

[6] H. Chen, F. Wang, C. Yang, D. Zeng, M. Chau, and K. Chang (Eds.), Intelligence and Security Informatics, Proceedings of the Workshop on Intelligence and Security Informatics, WISI 2006, Lecture Notes in Computer Science (LNCS 3917), Springer-Verlag, 2006.

[7] H. Chen, “Intelligence and Security Informatics for International Security: Information Sharing and Data Mining,” Springer, 2006.

[8] P. Kantor, G. Muresan, F. Roberts, D. Zeng, F. Wang, H. Chen, and R. Merkle (Eds.), Intelligence and Security Informatics, Proceedings of the IEEE International Conference on Intelligence and Security Informatics, ISI 2005, Lecture Notes in Computer Science (LNCS 3495), Springer-Verlag, 2005.

[9] H. Chen, R. Moore, D. Zeng, and J. Leavitt (Eds.), Intelligence and Security Informatics, Proceedings of the Second Symposium on Intelligence and Security Informatics, ISI 2004, Lecture Notes in Computer Science (LNCS 3073), Springer-Verlag, 2004.

[10] H. Chen, R. Miranda, D. Zeng, T. Madhusudan, C. Demchak, and J. Schroeder (Eds.), Intelligence and Security Informatics, Proceedings of the First NSF/NIJ Symposium on Intelligence and Security Informatics, ISI 2003, Lecture Notes in Computer Science (LNCS 2665), Springer-Verlag, 2003.

[11] Chen, H.. "Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums." ACM Transactions on Information Systems, forthcoming, 2008.

[12] Reid, E. and H. Chen, “Contemporary Terrorism Researchers’ Patterns of Collaboration and Influence,” *Journal of the American Society for Information Science and Technology*, forthcoming, 2008.

[13] Measuring Radicalization on the Internet, " in Proceedings of the IEEE International Intelligence and Security Informatics Conference (Taipei, Taiwan, July 17-20, 2008). Springer Lecture Notes in Computer Science.

[14] H. Chen, S. Thoms, T. Fu. "Cyber Extremism in Web 2.0: An Exploratory Study of International Jihadist Groups," in Proceedings of the 2008 IEEE Intelligence and Security Informatics Conference, Taiwan, June 17-20, 2008.