Exploring the Dark Side of the Web: Collection and Analysis of U.S. Extremist Online Forums

Yilu Zhou, Jialun Qin, Guanpi Lai, Edna Reid, and Hsinchun Chen

Department of Management Information Systems, The University of Arizona
Tucson, AZ 85721, USA
{yilu, qin, guanpi}@email.arizona.edu,
{ednareid, hchen}@eller.arizona.edu

Abstract. Contents in extremist online forums are invaluable data sources for extremism reseach. In this study, we propose a systematic Web mining approach to collecting and monitoring extremist forums. Our proposed approach identifies extremist forums from various resources, addresses practical issues faced by researchers and experts in the extremist forum collection process. Such collection provides a foundation for quantitative forum analysis. Using the proposed approach, we created a collection of 110 U.S. domestic extremist forums containing more than 640,000 documents. The collection building results demonstrate the effectiveness and feasibility of our approach. Furthermore, the extremist forum collection we created could serve as an invaluable data source to enable a better understanding of the extremism movements.

1 Introduction

Previous studies provided an abundance of illustrations of how computer-mediated communication (CMC) tools were used by extremist organizations to support their activities [3, 5]. Contents generated by extremist organizations' use of online CMC tools, especially online forums, provide snapshots of their activities, communications patterns, and ongoing developments. They could serve as invaluable data sources for researchers to better study extremist movements. However, due to problems such as information overload and the covert nature of extremism, no systematic methodologies have been developed for collection and analysis of extremists' Internet usage.

To address these research gaps, in this research, we propose a systematic Web mining approach to collecting, monitoring, and analyzing extremist online forums. We discuss in detail the primary issues in extremist forum collection and how to address them. We also report the preliminary results of a case study where we built a U.S. domestic extremist forum collection using the proposed approach to demonstrate its effectiveness and feasibility.

The remainder of the paper is structured as follows. In section 2, we briefly review relevant research in extremists' exploitation of the Internet. In section 3, we describe the proposed extremist forum collection approach. In section 4, we present the preliminary results of a case study with U.S. domestic extremist forum collection. In the last section, we conclude this paper and provide future recommendations.

S. Mehrotra et al. (Eds.): ISI 2006, LNCS 3975, pp. 621 – 626, 2006.

2 Literature Review

2.1 U.S. Domestic Extremists and the Internet

U.S. domestic extremist groups have continuously exploited technology to enhance their operations. Stormfront.org, a neo-Nazi's Web site set up in 1995, is considered the first major domestic "hate site" on the Internet [5]. Nowadays, extremist groups have established a significant presence on the Internet with several hundred multimedia Web sites, online chat rooms, online forums [4].

The dynamic contents in extremist forums and chat rooms could serve as invaluable data sources for extremism research. However, chat rooms are difficult to monitor, because chatting history is often not preserved after chatting sessions are over. Forums retain all communication messages for users to view and reply at a later time. Thus, forums are of special interest to us because of their rich information and accessibility.

2.2 Existing Studies on Extremists' Use of the Internet

Research on extremist groups' use of the Internet is in its early stages [2, 4]. Table 1 identifies several studies that used various methodologies to explore a range of research questions about domestic extremist groups' exploitation of Internet technology.

Table 1. Summary of Research on Domestic Extremist Groups' Use of the Internet

Methodology	Finding
Observation	Whine [5] traced of the early usages of the Internet by extremists and identified patterns of usage of USENET, bulletin boards, and Web sites.
Content Analysis	Bunt [1] analyzed Islamic Web sites and found that those sites formed part of a religious conceptual framework to inspire and motivate followers.
Content & Link Analysis	Burris et al. [2] found hyperlinks provided a accurate representation of inter-organizational structure among related organizations.
Content & Link Analysis	Zhou et al. [6] used a semi-automated approach in mining U.S. extremist Web sites. They performed link structure analysis and content analysis to facilitate understanding of extremists' virtual communities.

Except for our previous research on U.S. extremist Web sites, most of the studies identified in Table 1 involved manual processes for monitoring and collecting extremist Web site and forum data. Due to the inefficiency of manual approaches and complexity of forums monitoring, the scope of these studies was limited. None of these studies resulted in a comprehensive extremist forum testbed for large-scale in-depth analysis.

3 Proposed Approach

To effectively and efficiently collect and analyze extremist forum contents, we propose a systematic Web mining approach that combines expert knowledge and

automatic Web mining techniques to monitor and collect extremist forums. As shown in Figure 1, our proposed approach contains three major steps: forum identification, forum collection and parsing, and forum analysis. In the following sub-sections, we discuss the propose approach in detail.

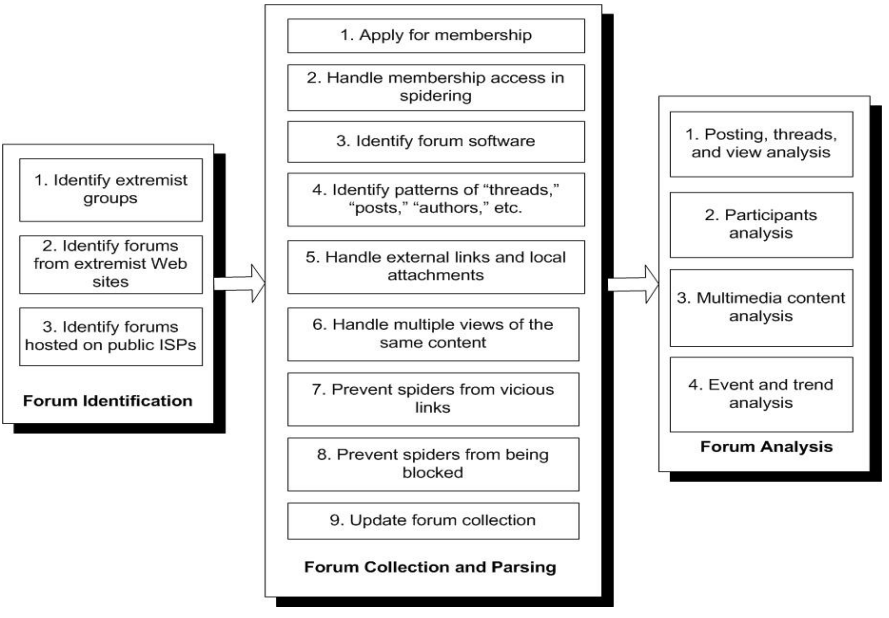

Fig. 1. A Web Mining Approach to Monitoring and Collecting Extremist Forums

3.1 Forum Identification

The identification of extremist forums is complicated by the fact that the Dark Web is covert and hidden away from general public. To ensure the quality of the collection, we identify extremist forums in the following three steps:

1) Identify extremist groups. We start the forum identification process by identifying the groups that are considered by authoritative sources as extremist groups. The sources include government agency reports (e.g., U.S. State Department reports, etc.), authoritative organization reports (e.g., UN Security Council reports, etc.), and studies published by terrorism research centers. Information such as group names, leader names, and jargons are identified to create an extremist keyword lexicon for use in the next steps.

2) Identify forums from extremist Web sites. In order to identify extremist forums, we first manually identify an initial set of terrorist group Web sites (called seed Web sites) from two sources: first, from authoritative sources used in the first step; second, from querying major online search engines with the extremist keyword lexicon. We expand the initial set of Web sites by extracting their out-links and back-links. We then identify and record forum entries from the expanded set of extremist Web sites.

3) Identify forums hosted on public ISPs. Besides forums on extremists' own Web sites, many extremist forums are hosted on public ISP servers such as Yahoo! Groups and Google Groups. We identify a list of these popular public ISPs and search for relevant forums using the extremist domain lexicon. By browsing through the messages of the forums returned from our search, relevant extremists' forums can be extracted.

The forums identified from both extremist Web sites and public ISP servers are filtered by domain experts to make sure that irrelevant or bogus forums do not make way into our final collection.

3.2 Forum Collection and Parsing

We proposed to use an automatic Web spidering approach to address the low efficiency problems of traditional manual forum collection approaches. Traditional Web crawling techniques cannot be directly applied in our case due to the defensive and diverse nature of extremist forums. To address these problems, we collect extremist forum documents in the following nine steps.

1) Apply for forum membership and handle access in spidering. Most extremist forums require membership to access. We manually send the application request to extremist forum masters as a curious neophyte. Once the application is approved, we use the user name and password to manually access the forums for the first time. Our access information is then stored in Internet cookies on our local computers. Then, we direct our spider program to use the cookies to access the forum contents.

2) Handle different forum software. Extremist forums are implemented using different forum software packages. The spider program needs to use different set of parameters to access those packages. We have created parameter templates for 12 most popular forums packages (e.g., vBulletin, ezboard, etc.) such that our spider program can generate parameters required by the corresponding forum packages. We also created parsers for each of the 12 popular forum such that our spider program can correctly extract thread information such as title, author, and post date, from the messages in different types of forums.

3) Handle external links, local attachments, and multiple views. Attachments posted by forum participants in their messages are very important. We set up our spider program to not only download textual messages, but also download multimedia documents, archive documents (e.g., ZIP files, etc.), and other non-standard files (files with extension names not recognizable by Windows operating system). Forums sometimes support views of the same message. These redundant documents need to be filtered based on the URL patterns to keep the collection concise.

4) Prevent Spiders from Vicious Links. Some forums may contain hyperlinks that trap a spider program in a loop (e.g. calendars, forum internal search engines, etc.). If the spidering process does not finish in a reasonable time, we need to examine the spidering log and exclude the vicious links from future spidering process.

5) Prevent Spiders from Being Blocked. We need to set a random time delay between hits and make the spider program mimic human browsing behaviors such that it will not be blocked by the forum servers. Some forums only allow specific

types of Web browsers to access their contents. We need to set up the spider program such that it mimics a certain Web browser to access the target forums.

6) Update Forum Collection. Forum contents are constantly being updated. The spider program needs to revisit the target forums periodically to download new threads and posts.

Once the terrorist/extremist forum collection is built, automatic Web mining and text mining techniques can be applied to the collection to identify the characteristics of active participants and listeners, analyze the number and types of multimedia files posted, and analyze the correlation between forum activities and major world events. Results of such analysis can facilitate researchers' in-depth analysis on those forums.

4 Case Study: A U.S. Domestic Extremist Forum Collection

In order to test the proposed approach, we conducted a case study in which we collected and analyzed contents from major U.S. domestic extremist forums. We believe that Web-based research on domestic extremist groups should prove valuable for supplementing and improving studies on domestic extremist movements.

Following the proposed approach, we started our forum collection process by identifying U.S. extremist groups from authoritative sources. We referred to the authoritative sources and identified 224 U.S. domestic extremist groups. Using the information of these groups as queries, we searched major search engines and public ISP servers for additional extremist forums. After the expansion and filter steps, we identified a total of 110 extremist forums of which 18 are hosted on extremist Web sites, 31 are hosted on Google Groups, 47 are hosted Yahoo! Groups, nine are hosted on MSN Groups, and five are hosted on AOL groups.

After obtaining membership for the password-protected forums, we spidered documents from the identified extremist forums. Table 2 is a summary of the number and volume of different types of documents we downloaded from the extremist forums.

Table 2. Summary of Document Types in the Forum Collection

	Stand Alone Forums		Public ISP Forums	
	# of Files	Volume (Bytes)	# of Files	Volume (Bytes)
Total	116,419	7.7G	524,652	20G
Textual Files	93,655	6.5G	350,046	10.7G
Multimedia Files	21,518	1.1G	6,511	1.3G
Non-Standard Files	1,246	45M	168,095	9G

As we can see from Table 2, our collection contains not only textual files, but also rich multimedia files and non-standard files. The non-standard files on extremist forums could be encrypted materials that were deliberately made inaccessible for normal software. Such rich contents could be used for various analysis purposes, such as hot topic analysis, time-series analysis, authorship analysis, and social network analysis.

5 Conclusions and Future Directions

In this study, we proposed a systematic Web mining approach to monitoring and collecting information from extremist forums. Using the proposed approach, we created a U.S. domestic extremist forums collection containing more than 600,000 multimedia documents. The comprehensiveness and quality of this collection demonstrated the effectiveness and feasibility of the proposed approach. Furthermore, this collection could serve as an invaluable data source for extremism research.

We have several future directions to pursue. First, we plan to get feedback from more domain experts to further improve the proposed approach. Second, we plan to apply our approach in case studies of larger scale. Third, we plan to explore more advanced machine learning and natural language processing techniques in the context of extremist forum analysis.

Acknowledgements

This research has been supported in part by the following grants:

- NSF, "COPLINK Center: Information & Knowledge Management for Law Enforcement," July 2000-September 2005.
- DHS/CNRI, "BorderSafe Initiative," October 2003-March 2005.

We also thank all anonymous domain experts who have contributed to the projects.

References

1. Bunt, G. R.: Islam In The Digital Age: E-Jihad, Online Fatwas and Cyber Islamic Environments. Pluto Press, London (2003)
2. Burris, V., Smith, E. Strahm, A.: White Supremacist Networks on the Internet, Sociological Focus, 33(2) (2003) 215-235
3. Dennings, D. E: Information Operations and Terrorism. (2004)
4. Gustavson, A. T., Sherkat, D.E.: Elucidating the Web of Hate: the Ideological Structuring of Network Ties Among Right Wing Hate Groups on the Internet, Annual Meetings of the American Sociological Association (2004)
5. Whine, M.: Far Right on the Internet, Governance of Cyberspace. B. Loader, ed., London Routledge (1997)
6. Zhou, Y., Reid, E., Qin, J., Chen, H., Lai, G.: US Domestic Extremist Groups on the Web:Link and Content Analysis, IEEE Intelligent Systems 20(5) (2005) 44-51