

Collecting and Analyzing the Presence of Terrorists on the Web: A Case Study of Jihad Websites

Edna Reid¹, Jialun Qin¹, Yilu Zhou¹, Guanpi Lai², Marc Sageman³, Gabriel Weimann⁴, and Hsinchun Chen¹

¹ Department of Management Information Systems, The University of Arizona,
Tucson, AZ 85721, USA
{ednareid, qin, yiluz, hchen}@bpa.arizona.edu

² Department of Systems and Industry Engineering, The University of Arizona,
Tucson, AZ 85721, USA
guanpi@email.arizona.edu

³ The Solomon Asch Center For Study of Ethnopolitical Conflict,
University of Pennsylvania, St. Leonard's Court, Suite 305, 3819-33 Chestnut Street,
Philadelphia, PA 19104, USA
sageman@sas.upenn.edu

⁴ Department of Communication, Haifa University, Haifa 31905, Israel
weimann@soc.haifa.ac.il

Abstract. The Internet which has enabled global businesses to flourish has become the very same channel for mushrooming ‘terrorist news networks.’ Terrorist organizations and their sympathizers have found a cost-effective resource to advance their courses by posting high-impact Websites with short shelf-lives. Because of their evanescent nature, terrorism research communities require unrestrained access to digitally archived Websites to mine their contents and pursue various types of analyses. However, organizations that specialize in capturing, archiving, and analyzing Jihad terrorist Websites employ different, manual-based analyses techniques that are inefficient and not scalable. This study proposes the development of automated or semi-automated procedures and systematic methodologies for capturing Jihad terrorist Website data and its subsequent analyses. By analyzing the content of hyperlinked terrorist Websites and constructing visual social network maps, our study is able to generate an integrated approach to the study of Jihad terrorism, their network structure, component clusters, and cluster affinity.

1 Introduction

Nowadays, the Internet has allowed terrorist groups to easily acquire sensitive intelligence information and control their operations [19]. Some research showed that terrorists use the Internet to develop a world-wide command, control, communication and intelligence system (C3I). For example, Jenkins posited that terrorists have used the Internet as a broadcast platform for the “terrorist news network” [12] which is an effective tactic because they can reach a broad audience with relatively little chance of detection.

Although this alternate side of the Internet, referred to as the Dark Web has recently received extensive government and media attention, our systematic understanding of how terrorists use the Internet for their campaign of terror is limited. According to studies by the Institute for Security and Technology Studies (ISTS) at Dartmouth College [11] and Anderson [2], there is a need to address this under-researched issue. In this study, we explore an integrated approach for harvesting Jihad terrorist Websites to construct a high quality collection of Website data that can be used to validate a methodology for analyzing and visualizing how Jihad terrorists use the Internet, especially the World Wide Web, in their terror campaigns. Jihad terrorist Websites are Websites produced or maintained by Islamic extremist groups or their sympathizers.

In this study, we answer the following research questions: What are the most appropriate techniques for collecting high-quality Jihad terrorism Webpages? What are systematic approaches for analyzing and visualizing Jihad terrorist information on the Web so as to identify usage and relationships among groups? How do you conduct a content analysis of the Jihad terrorists' collection?

2 Previous Research

In this section, we briefly review related studies on collection and analyzing terrorist Websites.

2.1 Terrorist Websites

The Web has been intensively used by terrorist organizations for their advantages. Arquilla and Ronfeldt [3] described this trend as netwar, an emerging model of conflict in which the protagonists use network forms of organization and exploit information technology. Many studies have been conducted on analyzing the terrorists' use of the Web. Examples include Elison [9], Tsfati and Weimann [21], ISTS [11], and Weimann [22]. All of them used terrorists' and their sympathizers' Websites as their primary data sources and provided brief descriptions of their methodologies.

To ensure the researchers and experts have access to terrorist Websites for research and intelligence analysis, several organizations are collecting, archiving, and analyzing Jihad terrorist Websites. These organizations include: the Internet Archive, the Project for Research of Islamist Movements, (PRISM) at the Interdisciplinary Center Herzliya, the Jihad and Terrorism Studies Project at the Middle East Research Institute (MEMRI), the Search for International Terrorist Entities (SITE Institute), and Professor Gabriel Weimann's collection at the University of Haifa, Israel [17]. Although all of them manually capture and analyze terrorist Websites to publish research reports, none publish their specific collection building and analytical approaches. Except for using search engines to identify terrorist Websites, none of the organizations seem to use any other automated methodologies for capturing and analyzing terrorist Websites.

2.2 Automated Web Harvesting

Previous research from the digital library community suggested automatic approaches to harvesting WebPages in particular domains. Web harvesting is the process of gathering and organizing unstructured information from pages and data on the Web [13]. Albertsen [1] uses an interesting approach in the “Paradigma” project. The goal of Paradigma is to archive Norwegian legal deposit documents on the Web. It employs a Web crawler that discovers neighboring Websites by following Web links found in the HTML pages of a starting set of WebPages. Metadata is then extracted and used to rank the Websites in terms of relevance.

In this study, we use a web spider to discover new Jihad terrorist Websites and use them as seeds to perform backlink searches (i.e., Google’s backlink search tool). However, we do not use metadata but rely instead on judgment calls by human experts because there are so many fake Jihad terrorism Websites. The “Political Communications Web Archiving” group also employs a semiautomatic approach to harvesting Websites [16]. Domain experts provide seed URLs as well as typologies for constructing metadata that can be used in the spidering process. Their project’s goal is to develop a methodology for constructing an archive of broad-spectrum political communications over the Web. In contrast, for the September 11 and Election 2002 Web Archives projects, the Library of Congress’ approach was to manually collect seed URLs for a given theme [18]. The seeds and their immediate neighbors (distance 1) are then crawled.

2.3 Web Link and Content Analysis

Web link analysis has been previously used to discover hidden relationships among commercial companies [10]. Borgman [4] defines two classes of web link analysis studies: relational and evaluative. Relational analysis gives insight into the strength of relations between web entities, in particular Websites, while evaluative analysis reveals the popularity or quality level of a Web entity. In this study, we are more concerned with relational analysis as it may bring us insight into the nature of relations between Websites and, possibly, terrorist organizations. Gibson [10] describes a methodology for discerning Web communities on the WWW. His work is based on Hyperlink-Induced Topic Search (HITS), a tool that searches for authoritative hypermedia on a given broad topic. In contrast, we construct a Website topology from a high quality Jihad Terrorism collection.

To reach an understanding of the various facets of Jihad terrorism Web usage, a systematic analysis of the Websites’ contents is required. Researchers in the terrorism domain have traditionally ignored existing methodologies for conducting a systematic content analysis of Website data [21,11]. In Bunt’s [5] overview of Jihadi movements’ presence on the Web, he described the reaction of the global Muslim community to the content of terrorist Websites. Tsfati and Weimann’s [21] study of terrorism on the Internet acknowledges the value of conducting a systematic and objective investigation of the characteristics of terrorist groups’ communications. All the studies mentioned above are qualitative studies. We believe Jihad terrorism content on the

Web falls under the category of communicative contents and a quantitative analysis is critical for a study to be objective.

Demchak and Friis' [8] work focused on measuring "openness" of government Websites using a Website Attribute System tool that is basically composed of a set of high level attributes such as transparency and interactivity. Each high level attribute is associated with a second layer of attributes at a more refined level of granularity. This system is an example of a well-structured and systematic content analysis exercise and provides guidance for the present study.

3 Proposed Approach and Preliminary Results

We propose an integrated approach to the study of Jihad Terrorism Web infrastructure. We combined a sound methodology for constructing a high-quality Jihad terrorism collection, a hyperlink analysis for the study of Jihad terrorism group relationships, and a systematic content analysis to study the details on the terrorists' Web usage. In the following sub-sections, we will describe the details of our approach and report the preliminary results from our analysis.

3.1 Jihad Collection Building

A systematic and sound methodology for collecting the Jihad terrorism Websites guarantees that our collection, which is the cornerstone of the study, is comprehensive and representative. We take a three step systematic approach to construct the collection:

1) *Identify seed URLs of Jihad terrorism groups and perform backlink expansion:* We first identified a set of Jihad terrorist groups from the US Department of State's list of foreign terrorist organizations. Then, we manually search major search engines (google.com, yahoo.com ...etc) using information such as the group names as queries to find Websites of these groups. Three Jihad terrorist Websites were identified: www.qudsway.com of the Palestinian Islamic Jihad, www.hizbollah.com of Hizbollah, and www.ezzedine.net which is a Website of the Izzedine-Al-Qassam, the military wing of Hamas. Then, we used Google back-link search service to find all the Websites that link to the three terrorist Websites mentioned above and obtained a total of 88 Websites.

2) *Filtering the collection:* Because bogus or unrelated terrorist sites can make their way into our collection, we have developed a robust filtering process based on evidence and clues from the Websites. We constructed a short lexicon of Jihad terrorism with the help of Arabic language speakers. Examples of highly relevant keywords included in the lexicon are: "حرب صليبية" ("Crusader's War"), "المجاهدين" ("Moujahedin"), "الكفار" ("Infidels"), etc. The 88 Websites were checked against the lexicon. Only those Websites which explicitly identify themselves as the official sites of a terrorist organization and the Websites that contain praise of or adopts ideologies

espoused by a terrorist group are included in our collection. After the filtering, 26 out of the 88 Websites remained in our collection.

3) *Extend the search manually*: To ensure the comprehensiveness of our collection we augment the collection by means of manually search large search engines using the lexicon constructed in the previous step. The Websites that are found are then filtered using the same rules used for filtering the backlink search results. As a result, 16 more Websites were identified and our final Jihad collection contains 39 terrorist Websites.

After identifying the Jihad terrorist Websites, we download all the Web pages within the identified sites. Our final collection contains more than 300,000 high-quality Web pages created by Jihad terrorists.

3.2 Link Analysis

Our goal here is to shed light on the infrastructure of Jihad Websites and to provide the necessary tools for further analysis of Jihad terrorist group relationships. We believe the exploration of hidden Jihad communities over the Web can give insight into the nature of relationships and communication channels between the Jihad terrorist groups.

Uncovering hidden Web communities involves calculating a similarity measure between all pairs of Websites. We define similarity to be a real-valued multivariable function of the number of hyperlinks between Website “A” and Website “B.” In addition, a hyperlink is weighted proportionally to how deep it appears in the Website hierarchy. For instance, a hyperlink appearing at the homepage of a website is given a higher weight than hyperlinks appearing at a deeper level. We calculated the similarity between each pair of Websites to form a similarity matrix. Then, this matrix was fed to a multidimensional scaling (MDS) algorithm which generated a two dimensional graph of the Website link structure. The proximity of nodes (Websites) in the graph reflects the similarity level. Figure 1. shows the visualization of the Jihad Website link structure.

Interestingly, domain experts recognized the existence of six clusters representing hyperlinked communities in the network. On the left side of the network resides the Hizbollah cluster. Hizbollah is a Lebanese militant organization. Established in 1982 during the Israeli invasion of Lebanon, the group routinely attacked Israeli military personnel until their pullout from south Lebanon in 2000. A cluster of Websites of Palestinian organizations inhabits the bottom left corner of the network: Hamas, Al-Aqsa Martyr’s Brigades, and the Palestinian Islamic Jihad. Hizbollah community and the Palestinian militant groups’ community were connected through hyperlink. Hizbollah has traditionally sympathized and supported the Palestinian cause. Hence, it is not surprising at all to see a link between the two virtual communities.

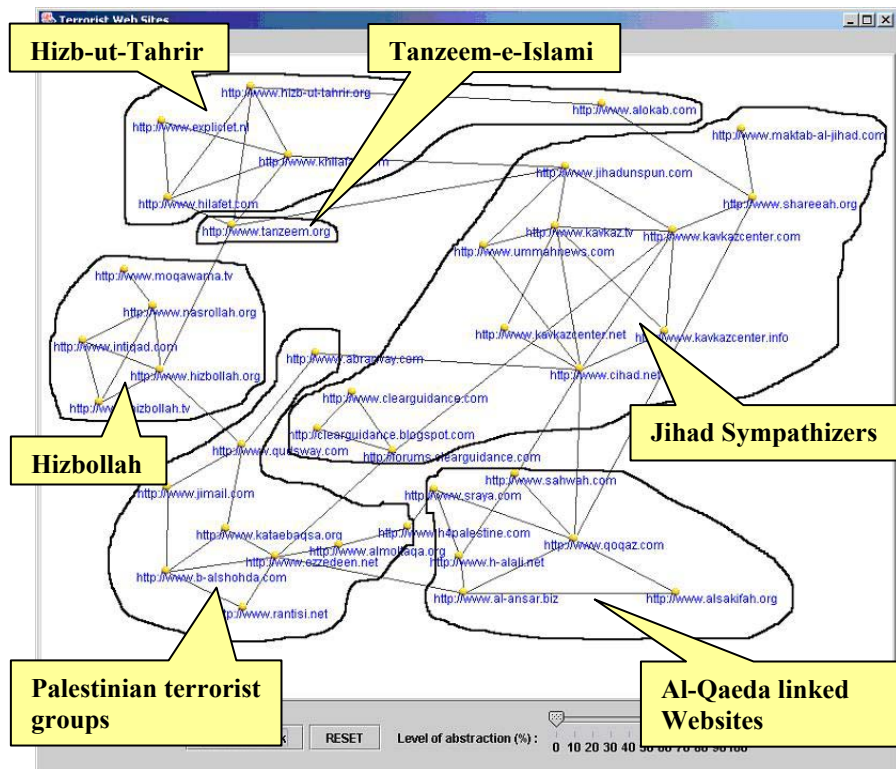


Fig. 1. The Jihad Terrorism Network with Automatically Generated Hyperlinked Communities

On the top left corner sits the Hizb-ut-Tahrir cluster which is a political party with branches in many countries over the Middle-East and in Europe. Although groups in this cluster are not officially recognized as terrorist groups, they do have links pointing to the Hizbollah cluster.

Looking at the bottom right corner one can see a cluster of Al-Qaeda affiliated sites. This cluster has links to two Websites of the radical Palestinian group Hamas. Al-Qaeda sympathizers with Palestinian groups. As well, some Palestinian Islamist groups like Hamas and Islamic Jihad share the same Salafi ideology with Al-Qaeda. In the top right hand corner, the Jihad Sympathizers Web community gathers Websites maintained by sympathizers of the Global Salafi movement. This community of Salafi sympathizers and supporters has links to three other major Sunni Web communities: the Al-Qaeda community, Palestinian extremists, and Hizb-ut-Tahrir communities. As expected the sympathizers community does not have any links to Hezbollah's community as they follow radically different ideologies.

Visualizing hyperlinked communities can lead to a better understanding of the underlying Jihad terrorism Web infrastructure. In addition, the visualization serves as a tool for showing the relationships between various hyperlinked communities. Furthermore, it helps foretell likely relationships between terrorist groups in the real world.

3.3 Content Analysis

To complete our analysis of Jihad terrorism on the Web we propose a framework for content analysis. The framework consists of high level attributes, each of which is composed of multiple fine grained low level attributes. This approach is similar to what is presented in Demchak and Friis' study [8]. Table 1 shows the high level and associated low level attributes used in this study.

Table 1. Attributes used in the study

High Level Attribute	Low Level Attribute
Communications	Email
	Telephone
	Multimedia
	Online Feedback Form
	Documentation
Fundraising	External Aid Mentioned
	Fund Transfer
	Donation
	Charity
	Support Groups
Sharing Ideology	Mission
	Doctrine
	Justification of the use of violence
	Pin-pointing enemies
Propaganda (insiders)	Slogans
	Dates
	Martyrs
	Leaders
	Banners and Seals
	Narratives of operations and Events
Propaganda (outsiders)	References to Western media coverage
	News reporting
Virtual Community	Listserv
	Text chat room
	Message board
	E-conferencing
	Web ring

Currently we only consider the presence of an attribute in a Website. In other words, the attribute for a given Website is assigned a "0" if it does not appear in a Website and a "1" if it does appear. However, this binary scheme does not capture the true contribution of the attributes. Hence, we assigned weights to each attribute such that

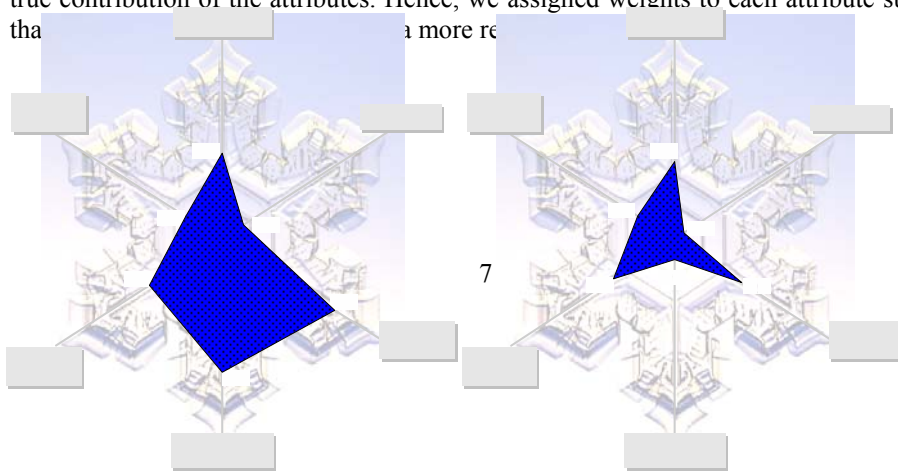


Fig. 2. Snowflake Diagrams for Palestinian terrorist groups and Jihad Supporters Web Communities

We asked our domain expert to go through each Website in our collection and record the presence of low-level attributes. The manual coding of the attributes in a Websites takes around 45 minutes of work. After completing the coding scheme for 32 Websites in the collection, we then compared the content of the clusters or hyper-linked communities in the network shown in Figure 2. We aggregated data from all Websites belonging to a cluster and displayed the result in snowflake diagrams. Figure 3 shows two such diagrams.

An interesting observation in these snowflake diagrams is the discrepancy in the “propaganda towards insiders” attribute. Militant groups, in this case Palestinian groups, tend to use the Web for disseminating their ideas in their own communities. They utilize propaganda as an effective tool for influencing youth and possibly recruiting new members. Conversely, the sympathizers try to explain their views to outsiders (Westerners) and try to justify terrorist actions.

4 Discussion and Future Work

We have developed an integrated approach to the study of the Jihad terrorism Web Infrastructure. Hyperlinked communities’ analysis brings an overall view of the terror web infrastructure. Visualizing hyperlinked communities facilitates the analysis of Web infrastructures and paves the way for more refined microscopic content analysis of the Websites. We then conducted a systematic content analysis of the websites and compared the content of various clusters. As part of our future work, we envisage implementing feature extraction algorithms for automatically detecting attributes in Web pages. We believe that our methodology can be an effective tool for analyzing Jihad terrorism on the Web. Moreover, it can be easily extended to analyze other Web contents.

5 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.
- NSF/ITR, “COPLINK Center for Intelligence and Security Informatics Research – A Crime Data Mining Approach to Developing Border Safe Research,” September 2003-August 2005.
- DHS/CNRI, “BorderSafe Initiative,” October 2003-March 2005.

We would like to thank Dr. Joshua Sinai from the Department of Homeland Security, Dr. Rex A. Hudson from the Library of Congress, and Dr. Chip Ellis from the MIPT organization for their insightful comments and suggestions on our project. We

would also like to thank all members of the Artificial Intelligence Lab at the University of Arizona who have contributed to the project, in particular Homa Atabakhsh, Cathy Larson, Chun-Ju Tseng, and Shing Ka Wu.

References

1. Albertsen, K.: The Paradigma Web Harvesting Environment. 3rd ECDL Workshop on Web Archives, Trondheim, Norway (2003)
2. Anderson, A.: Risk, Terrorism, and the Internet. *Knowledge, Technology & Policy* 16(2) (2003) 24-33
3. Arquilla, J., Ronfeldt, D.F.: Advent of Netwar. Rand Report (1996) <http://www.rand.org/>
4. Borgman, C. L., Furner, J.: Scholarly Communication and Bibliometrics. *Annual Review of Information Science and Technology*, ed. B. Cronin. Information Today, Inc (2002)
5. Bunt, G. R.: *Islam In The Digital Age: E-Jihad, Online Fatwas and Cyber Islamic Environments*. Pluto Press, London (2003)
6. Carley, K. M., Reminga, J., Kamneva, N.: Destabilizing Terrorist Networks. NAACSOS Conference Proceedings, Pittsburgh, PA (2003)
7. Carmon, Y.: Assessing Islamist Web Site Reports Of Imminent Terror Attacks In The U.S. MEMRI Inquiry & Analysis Series #156 (2003)
8. Demchak, C. C., Friis, C., La Porte, T. M.: Webbing Governance: National Differences in Constructing the Face of Public Organizations. *Handbook of Public Information Systems*, G. David Garson, ed., New York: Marcel Dekker Publishers (2000)
9. Elison, W.: Netwar: Studying Rebels on the Internet. *The Social Studies* 91 (2000) 127-131
10. Gibson, D., Kleinberg, J., Raghavan, P.: Inferring Web Communities from Link Topology: Proceedings of the 9th ACM Conference on Hypertext and Hypermedia (1998)
11. Institute for Security Technology Studies: Examining the Cyber Capabilities of Islamic Terrorist Groups. Report, ISTS (2004) <http://www.ists.dartmouth.edu/>
12. Jenkins, B. M.: World Becomes the Hostage of Media-Savvy Terrorists: Commentary. *USA Today* (2004) <http://www.rand.org/>
13. Kay, R.: Web Harvesting. *Computerworld* (2004) <http://www.computerworld.com>.
14. Kenney, A R., McGovern, N.Y., Botticelli, P., Entlich, R., Lagoze, C., Payette, S.: Preservation Risk Management for Web Resources: Virtual Remote Control in Cornell's Project Prism. *D-Lib Magazine* 8(1) (2002)
15. Reid, E. O. F.: Identifying a Company's Non-Customer Online Communities: a Prototyping. Proceedings of the 36th Hawaii International Conference on System Sciences, (2003)
16. Reilly, B., Tuchel, G., Simon, J., Palaima, C., Norsworthy, K., Myrick, L.: Political Communications Web Archiving: Addressing Typology and Timing for Selection, Preservation and Access. 3rd ECDL Workshop on Web Archives, Trondheim, Norway (2003)
17. Research Community PRISM: The Project for the Research of Islamist Movements. <http://www.e-prism.org>. MEMRI: Jihad and Terrorism Studies Project (2003)
18. SITE Institut : Report (2003) <http://www.siteinstitute.org/mission.html>.
19. Schneider, S. M., Foot, K., Kimpton, M., Jones, G.: Building thematic web collections: challenges and experiences from the September 11 Web Archive and the Election 2002 Web Archive. 3rd ECDL Workshop on Web Archives, Trondheim, Norway (2003)
20. Tekwani, S.: Cyberterrorism: Threat and Response. Institute of Defence and Strategic Studies, Workshop on the New Dimensions of Terrorism, Singapore (2002)
21. The 9/11 commission report (2004) <http://www.gpoaccess.gov/911/>

22. Tsfati, Y., Weimann, G.: www.terrorism.com: Terror on the Internet. *Studies in Conflict & Terrorism* 25 (2002) 317-332
23. Weimann, G.: www.terrorism.net: How Modern Terrorism Uses the Internet. Special Report 116, U.S. Institute of Peace (2004) <http://usip.org/pubs/>