Uncovering the Dark Web: A Case Study of Jihad on the Web

Hsinchun Chen
Artificial Intelligence Lab, Department of Management Information Systems, The University of Arizona, Tucson, AZ 85721, USA. E-mail: hchen@eller.arizona.edu

Wingyan Chung
Department of Operations and Management Information Systems, Leavey School of Business, Santa Clara University, Santa Clara, CA 95053, USA. E-mail: wchung@scu.edu

Jialun Qin
Management Department, College of Management, University of Massachusetts Lowell, Lowell, MA 01854, USA. E-mail: jialun_qin@uml.edu

Edna Reid
Department of Library Science, Clarion University, Clarion, PA 16214, USA. E-mail: ereid@clarion.edu

Marc Sageman
The Solomon Asch Center for Study of Ethnopolitical Conflict, University of Pennsylvania, Philadelphia, PA 19104, USA. E-mail: sageman@sas.upenn.edu

Gabriel Weimann
Department of Communication, University of Haifa, Haifa 31905, Israel. E-mail: weimann@soc.haifa.ac.il

While the Web has become a worldwide platform for communication, terrorists share their ideology and communicate with members on the "Dark Web"—the reverse side of the Web used by terrorists. Currently, the problems of information overload and difficulty to obtain a comprehensive picture of terrorist activities hinder effective and efficient analysis of terrorist information on the Web. To improve understanding of terrorist activities, we have developed a novel methodology for collecting and analyzing Dark Web information. The methodology incorporates information collection, analysis, and visualization techniques, and exploits various Web information sources. We applied it to collecting and analyzing information of 39 Jihad Web sites and developed visualization of their site contents, relationships, and activity levels. An expert evaluation showed that the methodology is very useful and promising, having a high potential to assist in investigation and understanding of terrorist activities by producing results that could potentially help guide both policymaking and intelligence research.

Received September 20, 2006; revised June 29, 2007; accepted January 4, 2008

© 2008 ASIS&T • Published online 7 April 2008 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20838

1. Introduction

The Internet has evolved to become a global platform through which anyone can conveniently disseminate, share, and communicate ideas. Despite many advantages, misuse of the Internet has become ever more serious, however. Terrorist organizations, extremist groups, hate groups, and racial supremacy groups are using the Web to promote their ideology, to facilitate internal communications, to attack their enemies, and to conduct criminal activities. Warnings have been made that terrorists may launch attacks on such critical infrastructure as major e-commerce sites and governmental networks (Gellman, 2002). Insurgents in Iraq have posted Web messages asking for munitions, financial support, and volunteers (Blakemore, 2004). It therefore has become important to obtain from the Web intelligence that permits better understanding and analysis of terrorist and extremist groups. We define this reverse side of the Web as a "Dark Web," the portion of the World Wide Web used to help achieve the sinister objectives of terrorists and extremists.

Currently, intelligence from the Dark Web is scattered in diverse information repositories through which investigators need to browse manually to be aware of their content. Much of the information stored in search engine databases could be properly collected and analyzed for transformation into intelligence and knowledge that would enhance understanding of terrorists' activities. However, search engines often overwhelm users by producing laundry lists of irrelevant results and creating information overload problems. Related but unfocused information makes it difficult to obtain a comprehensive description of a terrorist group or a terrorism topic. Many Web resources contain information *about* terrorism, but a relatively small proportion comes from terrorist groups themselves and data on the Web often are not persistent and may be misleading. Many terrorist Web sites do not use English, so investigators who do not speak that language may be unable to understand a site's content.

In this article, we have addressed the aforementioned problems by proposing and implementing a semiautomated methodology for collecting and analyzing Dark Web information. Leveraging human preciseness and machine efficiency, the methodology consists of various steps including collection, filtering, analysis, and visualization of Dark Web information. We used this comprehensive methodology to collect and analyze data from 39 Arabic terrorist Web sites and conducted an evaluation of the results. This research aimed to study to what extent the methodology can assist terrorism analysts in collecting and analyzing Dark Web information. From a broader perspective, this research contributes to the development of the new science of "Intelligence and Security Informatics (ISI)," the study of the use and development of advanced information technologies, systems, algorithms, and databases for national security related applications through an integrated technological, organizational, and policy based approach (Chen, 2005; Strickland & Hunt, 2005). We believe that many existing computer and information systems techniques need to be reexamined and adapted for this unique domain to create new insights and innovations.

The rest of this paper is structured as follows. The second section presents a review of terrorists' use of information technologies to facilitate terrorism, information services for studying terrorism, and advanced techniques for collecting and analyzing terrorism information. The third section describes a methodology for collecting and analyzing Dark Web information. The fourth section illustrates the use of the methodology in a case study of Jihad on the Web (where "Jihad" is an Islamic term referring to a holy war waged against enemies) and discusses the evaluation results. The last section concludes the study and discusses future directions.

2. Literature Review

2.1. *Terrorists' Use of the Web*

Recent studies have shown how terrorists use the Web to facilitate their activities. Tsfati and Weimann used the names of terrorist organizations to search six search engines and found 16 relevant sites in 1998 and 29 such sites in 2002 (Tsfati & Weimann, 2002). Their analysis of site content revealed heavy use of the Web by terrorist organizations to share ideology, to provide news, and to justify use of violence. Relying on open source information (e.g., court testimony, reports, Web sites), researchers at the Institute for Security Technology Studies identified five categories of terrorist use of the Web (Technical Analysis Group, 2004): propaganda (to disseminate radical messages); recruitment and training (to encourage people to join the Jihad and get online training); fundraising (to transfer funds, conduct credit card fraud and other money laundering activities); communications (to provide instruction, resources, and support via email, digital photographs, and chat session); and targeting (to conduct online surveillance and identify vulnerabilities of potential targets such as airports). Among these, using the Web as a propaganda tool has been widely observed.

Identified by the U.S. Government as a terrorist site, Alneda.com called itself the "Center for Islamic Studies and Research," a bogus name, and provided information for Al Qaeda (Thomas, 2003). To group members (insiders), terrorists use the Web to share motivational stories and descriptions of operations. To mass media and non-members (outsiders), they provide analysis and commentaries of recent events on their Web sites. For example, Azzam.com urged Muslims to travel to Pakistan and Afghanistan to fight the "Jewish-backed American Crusaders." Qassam.net appealed for donations to purchase AK-47 rifles (Kelley, 2002). Al Qaeda and some humanitarian relief agencies used the same bank accounts via www.explizit-islam.de (Thomas, 2003).

Terrorists also share ideologies on the Web that provide religious commentaries to legitimize their actions. Based on a study of 172 members participating in the global Salafi Jihad, Sageman concluded that the Internet has created a concrete bond between individuals and a virtual religious community (Sageman, 2004). His study reveals that the Web appeals to isolated individuals by easing loneliness through connections to people sharing some commonality. Such virtual community offers a number of advantages to terrorists. It no longer ties to any nation, fostering a priority of fighting against the far enemy (e.g., the United States) rather than the near enemy. Internet chat rooms tend to encourage extreme, abstract, but simplistic solutions, thus attracting most potential Jihad recruits who are not Islamic scholars. The anonymity of Internet cafés also protects the identity of terrorists. However, Sageman does not consider the Internet to be a direct contact with Jihad, because devotion to Jihad must be fostered by an intense period of face-to-face interaction. In addition, existing studies about terrorists' use of the Web mostly use a manual approach to analyze voluminous data. Such an approach does not scale up to rapid growth of the Web and frequent change of terrorists' identities on the Web.

2.2. *Information Services for Studying Terrorism*

Despite the public nature of the Web, terrorists often try to prevent authorities from tracing their Web addresses and activities, which has prompted several information services to monitor the Web sites of militant Islamic groups and to provide access to translated versions of information posted there. *The Jihad and Terrorism Project* was developed by the Middle East Media Research Institute to bridge the language gap between the West and the Middle East by providing timely translations of Arabic, Farsi, and Hebrew documents (Middle East Media Research Institute, 2004). The *Project for the Research of Islamist Movements* (www.e-prism.org) studies radical Islam and Islamist movements, focusing primarily on Arabic sources. These projects provide access to an array of information such as translated news stories, transcripts, video clips, and training documents produced by terrorists but fall short of supporting analysis and visualization of terrorist data from the Dark Web (Project for the Research of Islamist Movements, 2004).

2.3. Advanced Information Technologies for Combating Terrorism

Since the 9/11 attacks, there has been increased interest in using information technologies to counter terrorism. A study conducted by the U.S. Defense Advanced Research Projects Agency shows that their collaboration, modeling, and analysis tools speeded analysis (Popp, Armour, Senator, & Numrych, 2004), but these tools were not tailored to collecting and analyzing Web information. Although new approaches to terrorist network analysis have been called for (Carley, Lee, & Krackhardt, 2001), existing efforts have remained mostly small scale; they have used manual analysis of a specific terrorist organization and did not include resources generated by terrorists in their native languages. For instance, Krebs manually collected data from English news releases after the 9/11 attacks and studied the network surrounding the 19 hijackers (Krebs, 2001). Although automated social network analysis techniques have been proposed to analyze and portray criminal networks, it is not clear whether the techniques are applicable to the mostly unstructured data in terrorist Web sites that contain textual and multimedia data (Xu & Chen, 2005). Their use of structured data in a police department database also does not help understand terrorist Web sites. Other advanced information technologies having potential to help analyze terrorist data on the Web include information visualization and Web mining.

Information visualization technologies have been used in many domains (Zhu & Chen, 2005) such as criminal analysis (Chung, Chen, Chaboya, O'Toole, & Atabakhsh, 2005) and business stakeholder analysis (Chung, 2007). For example, multidimensional scaling (MDS) algorithms consist of a family of techniques that portray a data structure in a spatial fashion, where the coordinates of data points are calculated by a dimensionality reduction procedure (Young, 1987). MDS has been many different applications. Chung and his colleagues developed a new browsing method based on MDS to depict the competitive landscape of businesses on the Web (Chung, Chen, & Nunamaker, 2005). He and Hui applied MDS to displaying author cluster maps in their author co-citation analysis (He & Hui, 2002). Eom and Farris applied MDS to author co-citation in decision support systems (DSS) literature over 1971 through 1990 in order to find contributing fields to DSS (Eom & Farris, 1996). Kealy applied MDS to studying changes in knowledge maps of groups over time to determine the influence of a computer-based collaborative learning environment on conceptual understanding (Kealy, 2001). Although much has been done in different domains to visualize relationships of objects using MDS, no attempts to apply it to discovering terrorists' use of the Web have been found.

Web mining is the use of data mining techniques to automatically discover and extract information from Web documents and services (Chen & Chau, 2004; Etzioni, 1996). Chen and his colleagues (Chen, Fan, Chau, & Zeng, 2001) showed that the approach of integrating meta-searching with textual clustering tools achieved high precision in searching the Web. Web page classification, a process of automatically assigning Web pages into predefined categories, can be used to assign pages into meaningful classes (Mladenic, 1998). Web page clustering, a process of identifying naturally occurring subgroups among a set of Web pages, can be used to discover trends and patterns within a large number of pages (Chen, Schuffels, & Orwig, 1996). Although a number of Web mining technologies exist (e.g., Chen & Chau, 2004; Last, Markov, & Kandel, 2006), there has not yet been a comprehensive methodology to address problems of collecting and analyzing terrorist data on the Web. Unfortunately, existing frameworks using data and text mining techniques (e.g., Nasukawa & Nagano, 2001; Trybula, 1999) do not address issues specific to the Dark Web.

To our knowledge, few studies have used advanced Web and data mining technologies to collect and analyze terrorist information on the Web, though these technologies have been widely applied in such other domains as business and scientific research (e.g., Chung et al., 2004; Marshall, McDonald, Chen, & Chung, 2004). New approaches to collecting and analyzing terrorist information on the Web are needed.

3. A Methodology for Collecting and Analyzing Dark Web Information

3.1. The Methodology

To address threats from the wide range of information sources that terrorists and extremists use to spread their ideas and to conduct destructive activities, we have proposed a semiautomated methodology integrating various information collection and analysis techniques and human domain knowledge. Figure 1 shows the methodology aiming to effectively assist human investigators to obtain Dark Web intelligence using information sources, collection methods, filtering, and analysis.

- *Information sources* consist of a wide range of providers of terrorist or terrorism information on the Web. Some of these are readily accessible (e.g., search engines) while some, like terrorism incident databases and Web sites developed and

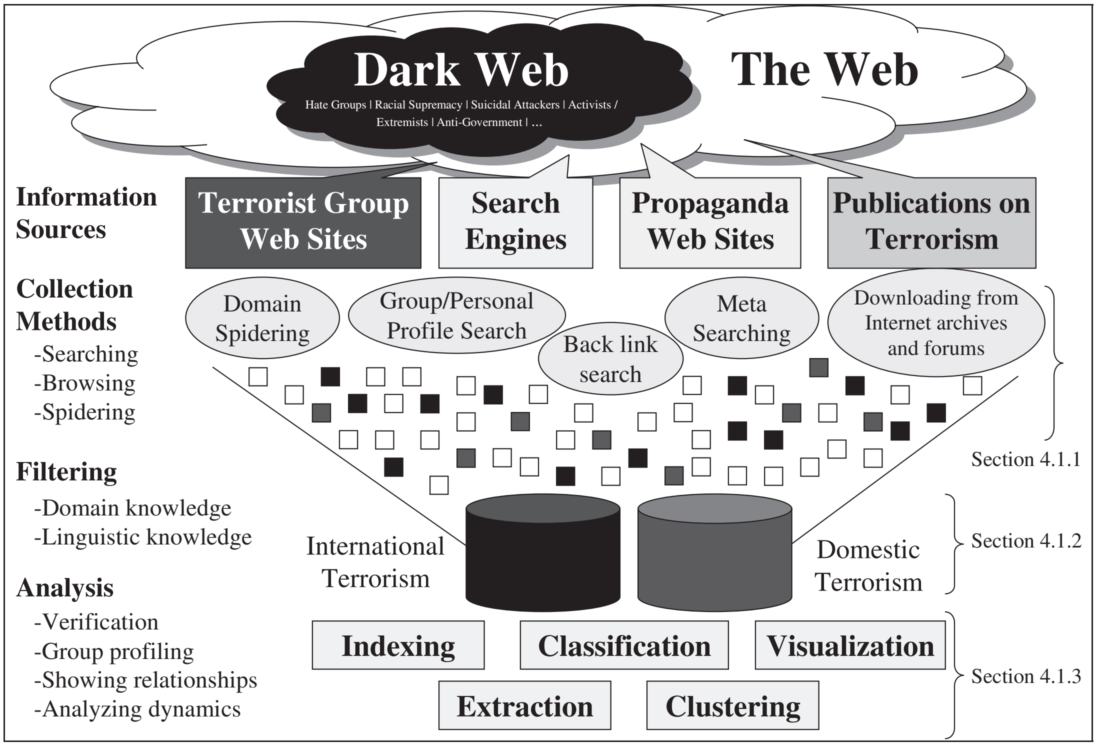

FIG. 1. A methodology for collecting and analyzing Dark Web information.

maintained by terrorists and their supporters, can only be reached with the help of domain experts.

- *Collection methods* make possible automatic searching, browsing, and harvesting of information from identified sources. *Domain spidering* starts with a set of relevant seed URLs and relies on an automatic Web page collection program, often called a spider or crawler, to harvest Web pages linked to the seed URLs. *Back-link search*, supported by some search engines such as Google (www.google.com) and AltaVista (www.altavista.com, acquired by Overture that was then acquired by Yahoo! in 2003), allows searching of Web pages that have hyperlinks pointing to a target Web domain or page. It helps investigators trace activities of terrorist supporters and sympathizers, whose Web pages often reference terrorist sites (e.g., glorify martyrs' actions, show a concurrence of terrorist attacks). *Group/personal profile search*, exemplified by major Web portals such as Yahoo! (members.yahoo.com) and MSN (groups.msn.com), reveals the profiles of groups or individuals who share the same interests. Terrorists and their supporters may perhaps put "hot links" in their profiles, which allow investigators to discover hidden linkages. *Meta-searching* uses related keywords as input to query multiple search engines from which investigators or automated programs can collate top-ranked results and filter out duplicates to obtain highly pertinent URLs of terrorist Web sites. With careful formulation of search terms and appropriate linguistic knowledge, they can obtain highly relevant results. For example, searching the Arabic name of "Usama Bin Laden" (اسامة بن لادن) in multiple search engines returns mixed results about terrorist news articles and terrorist Web sites, while augmenting "Usama Bin Laden" with the keyword "Sheikh" (the head of tribe or leader in Arabic), which is frequently used by Al Qaeda to refer to Bin Laden, can give more relevant terrorist and supporter Web sites. *Downloading from Internet archives and forums* exploits the temporal dimension of Web information. For instance, the Internet Archive (www.archive.org) offers access to historical snapshots of Web sites. Usenet discussion forums provide a wealth of textual communication that can be mined for hidden patterns over time.
- *Filtering* involves sifting through collected information and removing irrelevant results, but to perform this task requires domain knowledge and linguistic knowledge. Domain knowledge refers to knowledge about terrorist groups, their relationships with other terrorist and supporter groups, their presence on and usage of the Web, as well as their histories, activities, and missions. Linguistic knowledge deals with terms, slogans, and other textual and symbolic clues in the native languages of the terrorist groups. Filtering can be automatic or manual, depending on requirements for efficiency of process and precision of the results. Typically, manual filtering achieves high precision, but it is less efficient and relies on domain experts who have had years of experience in the field. Automatic filtering is very efficient as it often uses computers and machine learning to process large amounts of data but the results are less precise. Investigators can obtain high-quality data for analysis from filtered repositories.
- *Analysis* provides insights into data and helps investigators identify trends and verify conjectures. Several functions support these analytical tasks. *Indexing* relates textual terms to individual Web pages, thereby supporting precise searching of the pages. *Extraction* identifies meaningful entities such as terrorist names, frequently used slogans, and suspicious terms. *Classification* finds common properties among entities

and assigns them to predefined categories to help investigators predict trends of terrorist activities. *Clustering* organizes entities into naturally occurring groups and helps to identify similar terrorist groups and their supporters. *Visualization* presents voluminous data in a format perceivable by human eyes, so investigators can picture the relationship within a network organization of terrorist groups and can recognize their underlying structure.

3.2. Discussion of the Methodology

Although the Internet has been publicly available since the 1990s, the Dark Web emerged only in recent years. A lack of useful methodology designed for Dark Web data collection and analysis has limited the capability to fight against terrorism. As discussed above, the proposed methodology has incorporated various data and Web mining technologies while still allowing human domain knowledge to guide their application. Its semiautomated nature combines machine efficiency with the advantages of human precision, a useful complement to computers that usually fail to detect deception and ambiguity on the Dark Web. Its coverage of wide varieties of data sources and techniques ensures a comprehensive Dark Web data collection, a challenge often faced by terrorism and intelligence analysts. Therefore, the methodology and its integration and application of data and Web mining technologies to Dark Web analysis are novel contributions to the ISI research.

4. Jihad on the Web: A Case Study

To demonstrate the value and usability of our methodology, we have applied it to collecting and analyzing the use of the Web for *Jihad*, an Islamic term referring to a holy war waged against enemies as a religious duty. Believers contend that those who die in Jihad become martyrs and are guaranteed a place in paradise. In the recent decades, the concept of Jihad has been used as an ideological weapon to combat against Western influences and secular governments and to establish an ideal Islamic society (Encyclopedia Britannica Online, 2007). Jihad supporters are closely related to terrorist groups while maintaining anonymity using the Web. For example, prior to the 9/11 attacks, Al-Qaeda members sent each other thousands of messages in a password-protected section of an extreme Islamic Web site (Anti-Defamation League, 2002). Terrorist groups such as Hamas, Hizbollah, and Palestinian Islamic Jihad also use Web sites as propaganda tools. We describe the steps of applying the methodology as follows (see Figure 1). The data described below were collected in 2004.

4.1. Application of the Methodology

4.1.1. *Collection.* To collect data, we first identified four suspicious URLs through Web searching, referencing to published terrorism reports, and performing personal profile searches on Yahoo. (For example, we searched "hizbollah" in Google where we found its URL among the top-ranked results.) These URLs are Palestinian Islamic Jihad (PIJ; www.qudsway.com), Hizbollah (www.hizbollah.org), the military wing of Hamas (www.ezzedeen.net), and an Arabic Web site with a pro-Jihad forum (www.al-imam.net). A 2003 U.S. Department of State report confirmed that PIJ, Hizbollah, and Hamas to be terrorist or terrorist-affiliated groups (Department of State, 2003). Though Al-Imam.net is not classified as a terrorist organization, it contains pro-Jihad forums in which messages and links to terrorist Web sites are posted. We then used the back-link search function of Google to obtain several hundreds URLs that point to the four suspicious URLs. As Dark Web information can be scattered in many different sources and can be changed quickly over time, the several methods used to identify the four initial URLs enabled us to cover a broader scope and a more timely content than relying only on published reports (e.g., U.S. Department of State's annual report). While different initial URLs and different times of data collection could affect the content of the data collected, we believe that the choice of the four URLs are representative of the Dark Web. It would be an interesting future direction to study the extent to which data collection affect the quality of analysis results.

4.1.2. *Filtering.* We conducted two rounds of filtering. First, we manually filtered out unrelated sites, such as news or governmental Web sites that report or discuss only terrorist activities, religious Web sites with no reference to Jihad or violence, and political Web sites where there is no mention or approval of terrorist activities. We retained Web sites of terrorist organizations, those of terrorist leaders and those that praise terrorists or their actions. Forty-six sites remained after this round of filtering.

Second, with the help of a native Arabic speaker (who is not a terrorism expert), we manually added 14 terrorist and supporter sites identified by querying Google with the keywords (in Arabic) that we had found in the terrorist and supporter sites. Such keywords included the leaders' and organizations' names in Arabic ("mojahedin iran," "markaz dawa," "الشيخ المجاهد بن لادن," etc.). To limit the scope of analysis, we considered only the top 50 results returned from the search engine in each query search. In addition, we manually removed 21 sites from the set of all sites obtained based on their relevance to the domain. This round of filtering and refining resulted in 39 Arabic Web sites—24 terrorist sites and 15 supporter sites.

4.1.3. *Analysis.* We performed clustering, classification, and visualization on the 94,326 Web pages collected by crawling the 39 terrorist and supporter sites using an exhaustive breadth-first search spidering program (with a maximum depth of 10 levels). The first analysis task we performed was clustering in which we considered as input the 46 Web sites identified from the first round of filtering (see paragraph 1 of Section 4.1.2). The clustering involves calculating a similarity between each pair of Web sites in our collection to uncover hidden Web communities. We define similarity to be a real-valued multivariable function of the number of hyperlinks in one Web site ("A") pointing to another Web site ("B"), and the number of hyperlinks in the latter site ("B") pointing to the former site ("A"). In addition, a hyperlink is weighted proportionally to how deep it appears in the Web site hierarchy. For instance, a hyperlink appearing on the homepage of a Web site is given a higher weight than hyperlinks appearing at a deeper level. Specifically, the similarity between Web sites "A" and "B" is calculated as follows:

$$Similarity(A, B) = \sum_{\substack{\text{All links L} \\ \text{b/w A and B}}} \frac{1}{1 + lv(L)}$$

where $lv(L)$ is the level of link L in the Web site hierarchy, with homepage as level 0 and the level increased by 1 with each level down in the hierarchy. Using these heuristics, a computer program automatically extracted hyperlinks on Web pages and calculated their similarities.

In the second analysis task, we classified the sites by their affiliations with terrorist groups, ideologies, and religions, and by their Web site attributes. Our native Arabic speaker manually identified the affiliations of all the Web sites according to their site content. Even with the help of the Arabic speaker, the components of methodology are generic enough to be applicable to other domains. The choice of this Arabic speaker, (again, who is not a terrorism expert), also would not affect the results. Table 1 shows the details of the Web sites and their affiliations.

In addition to using affiliations, we classified the sites by indicating how terrorists and their supporters use the Web to facilitate their activities. From our literature review, we identified six types of terrorist use of the Web and 27 unique Web site attributes. Table 2 presents these attributes categorized under the six types. Following this coding scheme, the Arabic speaker manually read through all the subject Web pages to record terrorist uses of the Web. Similarly to that used in studying the openness of government Web sites (La Porte, Jong, & Demchak, 1999), our coding involved finding whether an attribute existed on the Web sites (i.e., binary scoring). Manual coding of each Web site required 45 minutes to 1 hour.

To reveal patterns of terrorist Web site existence and degree of a site's activities, we performed in the third analysis task two types of visualization: multidimensional scaling and snowflake visualizations.

Multidimensional scaling visualization provided a high-level picture of all the terrorist groups and their relationships. We used Multidimensional scaling (MDS) to transform a high-dimensional similarity matrix to a set of two-dimensional coordinates (Young, 1987). While other visualization techniques might have been applicable, we chose MDS because it suits the current data structure and provides a vivid picture summarizing terrorist groups' relationships. Figure 2 shows these relationships in which the sites appear as nodes and the lines connect pairs of sites that have at least one hyperlink pointing from one site to another. Using the similarity matrix as input, the MDS algorithm calculated coordinates of each site and placed the sites on a two-dimensional space where proximity reflects similarity. Upon closer examination of the figure, seven clusters of sites emerge. (The numbers in parentheses refer to the sites in Table 1. The URLs were filtered out in the second-round filtering but appeared in the collection after the first-round filtering.)

(1) *Hizballah Cluster* (# 7, 11, 12, hizbollah.org, and intiqad.org) contains the Web site of Hizballah group (www.hizbollah.org) and its affiliated sites such as Hizbollah E-magazine (www.intiqad.org), Hizbollah Support Association (#11), and the site of Sayyed Hassan Nasrollah (#12), a major leader of Hizbollah.
(2) *Palestinian Cluster* (# 4, 5, 6, 9, 13, 14, 15, 36, and h4palestine.com) includes militant groups fighting against Israel (e.g., Al-Aqsa Martyrs Brigade, Hamas). There are links between sites of the same group (e.g., # 4 and 14) and links between sites of different groups (e.g., # 9 and 6).
(3) *Al Qaeda Cluster* (# 26, 28, 31, 35, 37, and sahwah.com) includes Salafi groups' supporters' Web sites that often are linked to each other in their "Other friendly Web sites" section. They use their Web sites heavily to propagate their ideology. For example, Al-ansar.biz posted a video of the beheading of Nicholas Berg, one of the first civilians killed by terrorists (Newman, 2004). Alsakifah.org provides an online discussion forum.
(4) *Caucasian Cluster* (# 10, 34, kavkazcenter.com, kavkaz.tv, kavkazcenter.net, and kavkazcenter.info) consists of Web sites that link to Chechen rebels and provide news updates from Chechen areas. For example, Qoqaz.com has documented operations against Russian military.
(5) *Jihad Supporters* (# 29, 30, 32, 33, clearguidance.blogspot.com, and ummanews.com) consist of Web sites providing news and general information on the global Jihad movement. These sites rarely are linked to each other and often play a propaganda role that targets outsiders.
(6) *Hizb-Ut-Tahrir* (# 27, hizb-ut-tahrir.org, expliciet.nl, khilafah.com, and hilafet.com) contains a non-terrorist political group, Hizb-Ut-Tahrir, dedicated to the restoration of Islamic law and Khilafah (global leadership of Muslims). It has a presence in many Arab countries (e.g., Lebanon, Jordan) and some European countries. For instance, Expliciet.nl is a Dutch Web site based in the Netherlands.
(7) *Tanzeem-e-Islami Cluster* (tanzeem.org) consists of a single site representing the Pakistani "Tanzeem-e-Islami" party with no clear ties to terrorism.

Snowflake visualization supports analysis of different dimensions (or categories) of activities of a Web site cluster. It originates from a star plot that has been widely used to display multivariate data (Chambers, Cleveland, Kleiner, & Tukey, 1983). A snowflake shown in Figure 2 represents a terrorist site cluster. Figure 3 shows five snowflake diagrams, each representing the degree of activity of terrorist/supporter groups in the five terrorist clusters (Clusters 1–5) described above. (Clusters 6 and 7 are not included because they do not contain terrorist sites.) The six sides of a snowflake represent the six dimensions of terrorist use of the Web, as shown in Table 2 and explained above. Each of these six dimensions represents a normalized scale between 0 and 1 (activity index), showing the degree of activity on the dimensions.

TABLE 1. Analysis of Jihad terrorist groups and their supporters' sites.

No	Name	URL[a]	Description[b]	Terrorist group[c]	Religion
Terrorist Groups' Web Sites (total: 24)					
1	Special Force	www.specialforce.net	Provides computer game replicating the fighting scenes between Lebanese resistance and Israeli occupiers	Hizballah	Shi'a Muslim
2	Palestine Info in Urdu	palestine-info-urdu.com	Hamas news Web site in Urdu	Hamas	Sunni Muslim
3	Al-Manar	web.manartv.org	The Web site of Al-Manar, the TV channel of Lebanese Hizballah	Hizballah	Shi'a Muslim
4	Abrarway	www.abrarway.com	News Web site of Islamic Jihad of Palestine Guerrilla group	Palestinian Islamic Jihad	Sunni Muslim
5	Islamic Jihad Mail	www.jimail.com	News Web site of Islamic Jihad of Palestine Guerrilla group	Palestinian Islamic Jihad	Sunni Muslim
6	Ezz-al-dine Al-Qassam	www.ezzedeen.net	A general portal of Izz-Edeen Al-Qasam	Hamas	Sunni Muslim
7	Hizbollah	www.hizbollah.tv	The official Web site of Hizballah Organization	Hizballah	Shi'a Muslim
8	Info Palestina	www.infopalestina.com	Hamas information and news Web Site in Malay	Hamas	Sunni Muslim
9	Kataeb Al Aqsa	www.kataebalaqsa.com	The official Web Site of Al Aqsa Martyrs Brigades	Al-Aqsa Martyrs Brigade	Secular
10	Kavkaz	www.kavkaz.org.uk	The news Web Site of Chechen guerrilla fighters	Islamic International Brigade, Special Purpose Islamic Regiment, Riyadus-Salikhin Reconnaissance and Sabotage Battalion of Chechen Martyrs	Sunni Muslim
11	Moqawama	www.moqawama.tv	Web site of the Hizballah's support group	Hizballah	Shi'a Muslim
12	Nasrollah	www.nasrollah.org	Hizballah leader's site (Sheikh Hassan Nasrollah)	Hizballah	Shi'a Muslim
13	Alshohada	www.b-alshohda.com	Web site of Hamas and Islamic Jihad dedicated to martyrs	Hamas, Palestinian Islamic Jihad	Sunni Muslim
14	Quds Way	www.qudsway.com	Provides general news of Islamic Jihad of Palestine	Palestinian Islamic Jihad	Sunni Muslim
15	Rantisi	www.rantisi.net	Web site of Abdel Aziz Al Rantisi a Hamas leader	Hamas	Sunni Muslim
16	People's Mojahedin of Iran	www.iran.mojahedin.org	Web site posting statements by the People's Mojahedin Organization	Mujahedin-e Khalq Organization	Secular
17	National Council of Resistance of Iran	www.iranncrfac.org	Official Web site of the Foreign Affairs Committee of the National Council of Resistance of Iran	Mujahedin-e Khalq Organization	Secular
18	Iranian People's Fadaee Guerrillas	www.siahkal.com	The memorial Web Site of the Iranian People's Fadaee Guerrillas	Mujahedin-e Khalq Organization	Secular
19	The Organization of Iranian People's Fedaian	www.fadai.org	The Organization of Iranian People's Fedaian (Majority) official Web site	Mujahedin-e Khalq Organization	Secular
20	Organization of Iranian People's Fedayee Guerrillas	www.fadaian.org	Organization of Iranian People's Fedayee Guerrillas memorial Web site	Mujahedin-e Khalq Organization	Secular
21	The Union of People's Fedaian of Iran	www.etehadefedaian.org	News and information Web site of the Union of People's Fedaian of Iran	Mujahedin-e Khalq Organization	Secular

(*Continued*)

TABLE 1. (*Continued*)

No	Name	URL[a]	Description[b]	Terrorist group[c]	Religion
22	Revolutionary Peoples Liberation Front	www.dhkc.net	Revolutionary Peoples Liberation Front official Web site. Provides news and statements of the organization	Revolutionary People's Liberation Army/Front	Secular
23	DHKC International	www.dhkc.info	Web site of DHKC in Turkish	Revolutionary People's Liberation Army/Front	Secular
24	Crusade Begins	jorgevinhedo.sites.uol.com.br	The Brazil-based Web site links to Lashkar-e-Taiba – a terrorist organization based in Pakistan	Lashkar-e Tayyiba	Sunni Muslim
Supporters' Web sites (total: 15)					
25	Al Ansar	www.al-ansar.biz	Provides support to Al Qae'da organization, as well as articles about the Salafi Sunni Ideology	Al Qaeda	Sunni Muslim
26	Alokab	www.alokab.com	Provides articles about the Salafi Sunni Ideology and the Jihadist movement	Al Qaeda	Sunni Muslim
27	Alsakifah Forum	www.alsakifah.org	Provides educational services and a forum dedicated to the discussion of the Salafi Ideology	Al Qaeda	Sunni Muslim
28	Cihad	www.cihad.net	A general Jihad Web site providing information about all Jihad activities around the world	Al Qaeda	Sunni Muslim
29	Clear Guidance Forum	www.clearguidance.com	Forum of Jihad supporters	Al Qaeda	Sunni Muslim
30	Sheikh Hamid Bin Abdallah Al Ali	www.h-alali.net	Salafi Educational Web site with some Jihad ideas	Al Qaeda	Sunni Muslim
31	Jihadunspun	www.jihadunspun.com	Pro-Jihad news Web site	Al Qaeda	Sunni Muslim
32	Maktab-Al-Jihad	www.maktab-al-jihad.com	Pro-Jihad news Web site	Al Qaeda	Sunni Muslim
33	Qoqaz	www.qoqaz.com	Jihad news from the Caucasus	Islamic International Brigade, Special Purpose Islamic Regiment, Riyadus-Salikhin Reconnaissance and Sabotage Battalion of Chechen Martyrs	Sunni Muslim
34	Supporters of Shareeah	www.shareeah.org	A general portal dedicated to the Jihadist movement	Al Qaeda	Sunni Muslim
35	Moltaqa	www.almoltaqa.org	Hamas Forum	Hamas	Sunni Muslim
36	Saraya	www.saraya.com	Pro-Jihad Web site	Al Qaeda	Sunni Muslim
37	Osama Bin Laden	1osamabinladen.5u.com	A Web site dedicated to Osama Bin Laden	Al Qaeda	Sunni Muslim
38	Tawhed	www.tawhed.ws	Pro-Jihad Web site	Al Qaeda	Sunni Muslim
39	The Right Word	www.rightword.net	Pro-Al Qae'da Web Portal	Al Qaeda	Sunni Muslim

[a] Some of the URLs and sites may have been changed at the time of reading due to the rapid change of the Dark Web.

[b] The descriptions are obtained from the Web sites.

[c] Descriptions of these terrorist groups appear in the U.S. Department of State Report *Pattern of Global Terrorism, 2002*.

TABLE 2. Categories of terrorist use of the Web and Web site attributes.

Category	Attribute	Description
Communications	E-mail	Any listed email address or feedback form.
	Telephone (including Web phone)	Telephone numbers of organization officials.
	Multimedia tools	Video clips of bombings and other activities. Video, sound recording & game (e.g., leader's messages and instructions).
	Online feedback form	Allow the user to give feedback or ask questions to the Web site owners and maintainers.
	Documentation	Report, book, letter, memo and other resources provided (e.g., in pdf, Word, and Excel formats).
Fundraising	External aid mentioned	Other groups or governments supporting the organization.
	Fund transfer	Fund transfer methods.
	Donation	Donations under the form of direct bank deposits.
	Charity	Donations to religious welfare organizations associated with terrorist organization.
	Support groups	Suborganizational structures charged with the fund raising program.
	Others	Other attributes belonging to this category.
Sharing ideology	Mission	The major goals of the organization (e.g., destruction of an enemy state, liberation of occupied territories).
	Doctrine	The beliefs of the group (e.g., religious, communist, extreme right).
	Justification of the use of violence	Ideology condones the use of violence to accomplish goals (e.g., suicide bombing).
	Pinpointing enemies	Classifies others as either enemies or friends (e.g., U.S. is enemy, Taliban regime is friendly).
Propaganda (insiders)	Slogans	Short phrases with religious or ideological connotations.
	Dates	Mentions dates in the history of the terrorist group, such as the date of a major attack.
	Martyr's description	Lists the names of members who died in terrorism related operations or descriptions of the circumstances.
	Leader's name(s)	Terrorist groups leader(s) name as claimed by the Web site.
	Banner and seal	Banner depicting representative figures, graphical symbols, or seals of the organization.
	Narratives of operations and events	Provides narratives of the operations and attacks of the group.
	Others	Other attributes belonging to this category.
Propaganda (outsiders)	Reference to media coverage of events	For example, the Web site criticizes Western media coverage of events with explicit mention of outlets such as CNN, CBS.
	News reporting	Group's own interpretation of events.
Virtual community	Listserv	Automatic mailing list server that broadcasts to everyone on the list.
	Text chat room	Virtual room where a chat session takes place. Text messaging chat session such as ICQ.
	Message board	Allows members to post and read messages online.
	Web ring	A series of web sites linked together in a ring that by clicking through all of the sites in the ring the visitor will eventually come back to the originating site.

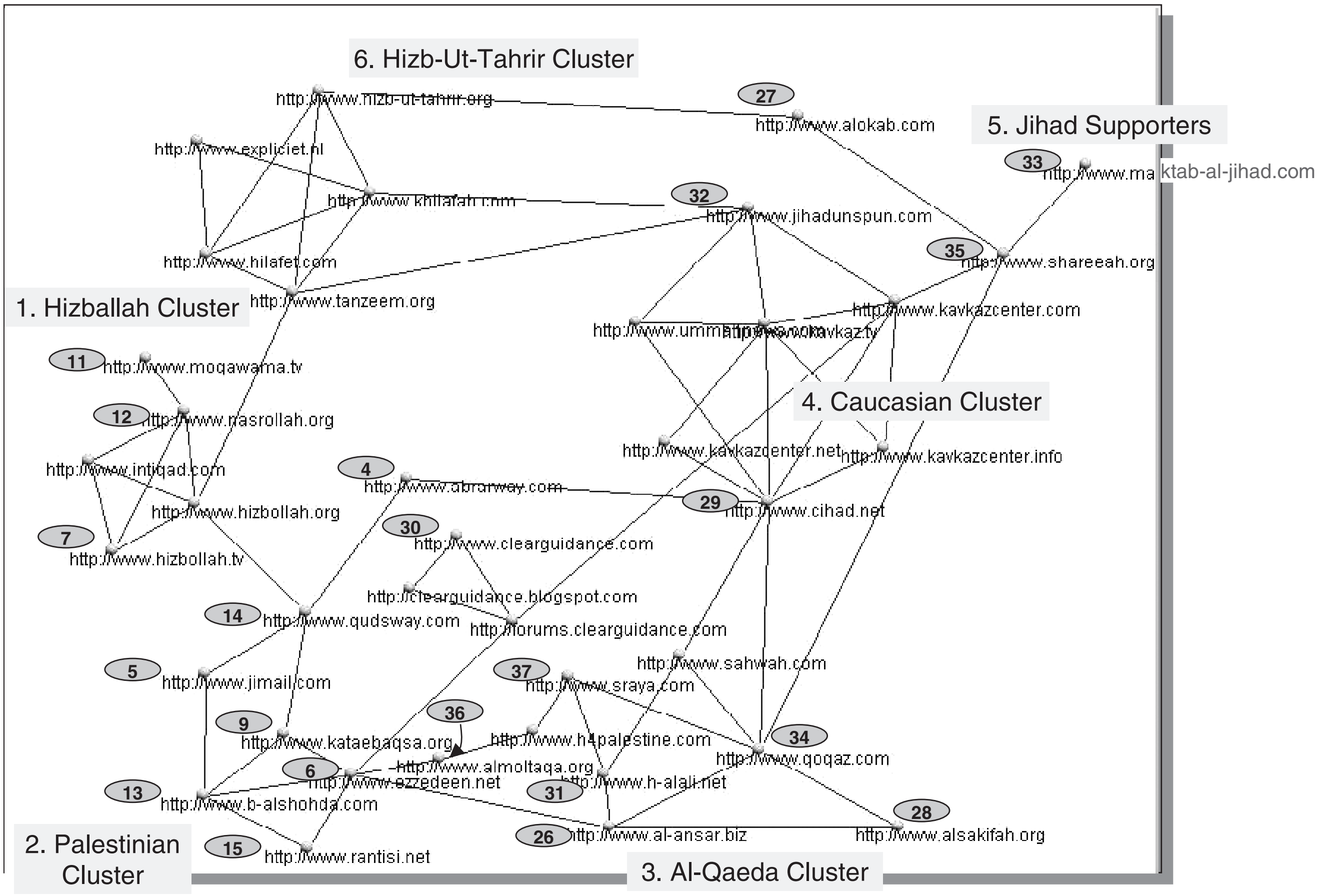

FIG. 2. Clustering and visualization of terrorist Web sites (The numbers refer to those appearing in Table 1)*.

The activity index of Cluster c on dimension d was calculated by the following formula:

$$\text{Activity Index}\,(c,d) = \frac{\sum_{i}^{n}\sum_{j}^{m} w_{i,j}}{m \times n}$$

$$\text{where } w_{i,j} = \begin{cases} 1 & \text{attribute i occurs in Web site j} \\ 0 & \text{otherwise} \end{cases}$$

n = total number of attributes in the specified dimension d; m = total number of Web sites belonging to the specified Cluster c.

The closer the activity index is to 1, the more active a cluster is on that dimension. This index reveals in what areas the terrorist groups are active and hence provides investigators and analysts with clues about how to devise strategies to combat a group.

4.2. *Results and Discussions*

Our preliminary observations show that the methodology yielded promising results. For example, it identified Web sites affiliated with 10 of the 26 groups classified as Jihad terrorist organizations in the U.S. State Department report on terrorism. Al-Ansar.biz (# 26), the site that posted the beheading video of Nicholas Berg, posted messages from Al Qaeda leaders such as Osama Bin Laden, Ayman Al-Zawahiri, and Al-Zarqawi, praising their attacks on enemies. Another site, Tawhed.com (site 39), posted a poem praising the 9/11 attacks. The rhetoric of the poem commonly appears in many Al Qaeda affiliated Web sites, referring to the Americans as crusaders (الصليبيين). Words like Sunna and Jama'h (السنة والجماعة) reflect the branch of Islam to which the Salafi groups belong.

From the snowflake diagrams (Figure 3), we found that terrorists and supporters use the Web heavily to share ideology and to propagate ideas, especially to their members. For example, the Palestinian cluster (Cluster 2) actively shares its ideology and heavily uses the Web as a propaganda tool for members. The Web sites in this cluster support liberation of Palestine, pinpoint and criticize their enemies, and describe details of operations and rationales supported by Quaran verses. In contrast, Jihad supporters (Custer 5) rarely use the Web for propaganda but share ideology and communicate there. The Hizbollah cluster (Cluster 1) resembles the Palestinian cluster in heavy use of the Web for sharing ideology and insider propaganda. For example, the sites in this cluster glorify martyrs and leaders and also were used moderately for outsider propaganda and communications. In all the five clusters, we found little evidence of using the

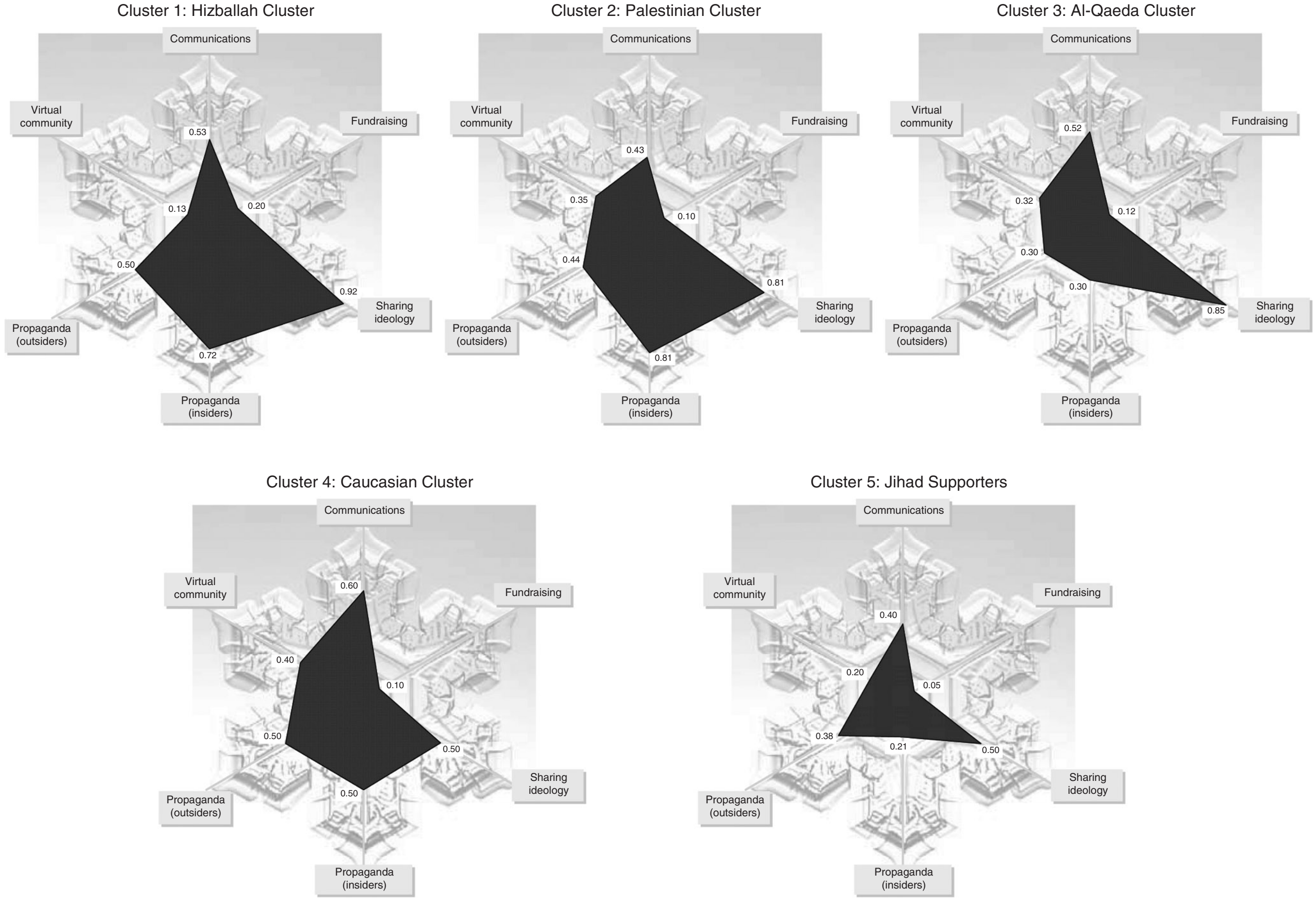

FIG. 3. Snowflake visualization of five terrorist site clusters.

Web for fundraising or building a virtual community. Probably such uses have gone underground or do not appear on the Web.

4.3. *Expert Evaluation and Results*

Based on the above results, we have invited a terrorism expert to conduct an evaluation of the methodology. A senior fellow of the U.S. Institute of Peace at Washington D.C., the expert is a professor of communication in a major research university in Israel. Having expertise in modern terrorism and the Internet, he has published more than 80 refereed journal articles and books and is a frequent speaker at international conferences on counter terrorism. This expert also leads a team of about 16 research assistants who regularly monitor 4,300 sites on the Dark Web for terrorist activities. The approach he and his team use to collect and analyze terrorists' use of the Web is largely manual, relying on laborious human browsing and monitoring of selected Web sites. His experience in manual analysis served to contrast with our methodology that automated part of the Dark Web data collection and analysis. We decided to use expert validation instead of other evaluation methods because of two reasons: (1) Lab experiment is not suitable because typical experimental subjects do not have much knowledge in the Dark Web, and (2) it is not feasible to invite terrorists to participate in an interview or empirical evaluation. The expert was not involved in writing this article.

The evaluation was conducted using an unbiased structured questionnaire and a formal procedure. We showed the results to our expert and asked him to provide detailed comments on the categorization of Web sites and attributes, the visualization and clustering of terrorist groups, and the usability of the snowflake visualization. In general, he deemed the results to be very promising and the methodology design to be excellent. He believed that this was the start of a very important research that will result in a useful database and a reliable methodology to update and maintain the database.

The expert was greatly impressed by the visualization and clustering capabilities of the methodology, and he provided valuable comments on our work. However, he said that the 39 Web sites shown in Table 1 do not represent the entire population of all terrorist Web sites, the number of which he estimated to be over four thousands. Because we focused only on Middle Eastern terrorist groups (rather than all terrorist groups in the world), we believe that our methodology has yielded representative results and has automated much of the manual work of identifying and analyzing terrorist Web sites. He suggested adding qualitative measures such as persuasive appeals, rhetoric, and attribution of guilt to the Web site attributes shown in Table 2. We believe that these important attributes are difficult to be incorporated into the automated processing of our methodology because of their qualitative nature. He considered the clustering and visualization shown in Figure 2 to be very important because of its usefulness to investigation of terrorist activities on the Web. He called the snowflake visualization very accurate and very useful to investigation of terrorist Web sites but criticized the way we created linkages among Web sites. He suggested considering textual citations and other references in addition to using only hyperlinks.

Overall, the expert agreed that the results were very promising because they offer useful investigation leads and would be very helpful to improve understanding of terrorist activities on the Web. Because of the high qualification and relevant experience of this expert, we believe that the evaluation results can accurately reflect the effectiveness of the methodology. These results also contributed to advancing the ISI discipline by showing the applicability of the methodology to Dark Web data collection and analysis.

5. Conclusions and Future Directions

Collecting and analyzing Dark Web information has challenged investigators and researchers because terrorists can easily hide their identities and remove traces of their activities on the Web. The abundance of Web information has made it difficult to obtain a comprehensive picture of terrorists' activities. In this article, we have proposed a methodology to address these problems. Using advanced Web mining, content analysis, visualization techniques, and human domain knowledge, the methodology exploited various information sources to identify and analyze 39 Jihad Web sites. Information visualization was used to help to identify terrorist clusters and to understand terrorist use of the Web. Our expert evaluation showed that the methodology yielded promising results that would be very useful to assist investigation of terrorism. The expert considered the visualization results very useful, having potential to guide policymaking and intelligence research. Therefore, this research has contributed to developing a useful methodology for collecting and analyzing Dark Web information, applying the methodology to studying and analyzing 39 Jihad Web sites, and providing formal evaluation results of the usability of the methodology.

We are pursuing a number of directions to further our research. As terrorists often change their Web sites to remove traces of their activities, we plan to archive the Dark Web content digitally and apply our methodology to tracing terrorist activities over time. We will develop scalable techniques to collect such volatile yet valuable content to visualize large volumes of Dark Web data and extract meaningful entities from terrorist Web sites. These efforts will help investigators trace and prevent terrorist attacks.

6. Acknowledgments

This research was partly supported by funding from the U.S. Government Department of Homeland Security and Corporation for National Research Initiatives and by Santa Clara University. We thank contributing members of the University of Arizona Artificial Intelligence Lab for their support and assistance.

References

Anti-Defamation League. (2002). Jihad Online: Islamic Terrorists and the Internet, retrieved March 26, 2008 from http://www.adl.org/internet/jihad_online.pdf.

Blakemore, B. (November 23, 2004). Web posting may provide insight into Iraq insurgency. ABC News, retrieved March 26, 2008 from http://abcnews.go.com/WNT/story?id=277421.

Carley, Kathleen M. Ju-Sung Lee and David Krackhardt, 2001, Destabilizing Networks, Connections, 24(3): 31–34.

Chambers, J., Cleveland, W., Kleiner, B., & Tukey, P. (1983). Graphical methods for data analysis. Wadsworth International Group (Belmont, CA) and Duxbury Press (Boston, MA).

Chen, H. (2005). Introduction to the special topic issue: Intelligence and security informatics. Journal of the American Society for Information Science and Technology, 56(3), 217–220.

Chen, H., & Chau, M. (2004). Web mining: Machine learning for Web applications. In M. E. Williams (Ed.), Annual review of information science and technology (Vol. 38, pp. 289–329). Medford, NJ: Information Today, Inc.

Chen, H., Fan, H., Chau, M., & Zeng, D. (2001). MetaSpider: Meta-searching and categorization on the Web. Journal of the American Society for Information Science and Technology, 52(13), 1134–1147.

Chen, H., Schuffels, C., & Orwig, R. (1996). Internet categorization and search: A self-organizing approach. Journal of Visual Communication and Image Representation, 7(1), 88–102.

Chung, W. (2008). Visualizing E-Business Stakeholders on the Web: A Methodology and Experimental Results. International Journal of Electronic Business, 6(1), 2008, 25–46.

Chung, W., Chen, H., Chaboya, L.G., O'Toole, C., & Atabakhsh, H. (2005). Evaluating event visualization: A usability study of COPLINK Spatio-Temporal Visualizer. International Journal of Human-Computer Studies, 62(1), 127–157.

Chung, W., Chen, H., & Nunamaker, J.F. (2005). A visual framework for knowledge discovery on the Web: An empirical study on business intelligence exploration. Journal of Management Information Systems, 21(4), 57–84.

Chung, W., Zhang, Y., Huang, Z., Wang, G., Ong, T.-H., & Chen, H. (2004). Internet searching and browsing in a multilingual world: An experiment on the Chinese business intelligence portal (CBizPort). Journal of the American Society for Information Science and Technology, 55(9), 818–831.

Department of State. (2003). Patterns of Global Terrorism 2002: The United States Government, retrieved March 26, 2008 from http://www.state.gov/s/ct/rls/crt/2002/.

Encyclopedia Britannica Online. (2007). Jihad. Retrieved March 26, 2008 from http://www.britannica.com/ebc/article-9368558, Britannica Concise Encyclopedia.

Eom, S.B., & Farris, R.S. (1996). The contributions of organizational science to the development of decision support systems research subspecialties. Journal of the American Society for Information Science, 47(12), 941–952.

Etzioni, O. (1996). The World Wide Web: Quagmire or gold mine? Communications of the ACM, 39(11), 65–68.

Gellman, B. (June 27, 2002). Cyber-attacks by Al Qaeda feared. Washington Post, page A01, retrieved March 26, 2008 from http://www.washingtonpost.com/ac2/wp-dyn/A50765-2002Jun26.

He, Y., & Hui, S.C. (2002). Mining a Web citation database for author co-citation analysis. Information Processing and Management, 38(4), 491–508.

Kealy, W.A. (2001). Knowledge maps and their use in computer-based collaborative learning. Journal of Educational Computing Research, 25(4), 325–349.

Kelley, J. (July 10, 2002). Militants Wire Web With Links to Jihad. USA Today, retrieved March 26, 2008 from http://www.usatoday.com/news/world/2002/07/10/web-terror-cover.htm.

Krebs, V.E. (2001). Mapping network of terrorist cells. Connections, 24(3), 43–52.

La Porte, T. M., Jong, M. d., & Demchak, C. C. (1999). Public Organizations on the World Wide Web: Empirical Correlates of Administrative Openness. Paper presented at the Proceedings of the 5th National Public Management Research conference, College Station, TX.

Last, M., Markov, A., & Kandel, A. (2006). Multi-Lingual Detection of Terrorist Content on the Web. Paper presented at the Proceedings of the PAKDD'06 International Workshop on Intelligence and Security Informatics, Singapore, Springer, Berlin / Heidelberg, pp. 16–30.

Marshall, B., McDonald, D., Chen, H., & Chung, W. (2004). EBizPort: collecting and analyzing business intelligence information. Journal of the American Society for Information and Science and Technology, 55(10), 873–891.

Middle East Media Research Institute. (2004). Jihad and Terrorism Studies Project. Retrieved March 2004, retrieved March 26, 2008 from http://www.memri.org/jihad.html.

Mladenic, D. (1998). Turning Yahoo into an automatic web page classifier. Paper presented at the Proceedings of the 13 European Conference on Artificial Intelligence, Brighton, UK.

Nasukawa, T., & Nagano, T. (2001). Text analysis and knowledge mining system. IBM Systems Journal, 40(4), 967–984.

Newman, M. (2004, May 11). Video appears to show beheading of American civilian. The New York Times.

Popp, R., Armour, T., Senator, T., & Numrych, K. (2004). Countering terrorism through information technology. Communications of the ACM, 47(3), 36–43.

Project for the Research of Islamist Movements. (2004). PRISM, 2004, retrieved March 26, 2008 from http://www.e-prism.org.

Sageman, M. (2004). Understanding terror networks. Philadelphia, PA: University of Pennsylvania Press.

Strickland, L.S., & Hunt, L.E. (2005). Technology, security, and individual privacy: New tools, new threats, and new public perceptions. Journal of the American Society for Information Science and Technology, 56(3), 221–234.

Technical Analysis Group. (2004). Examining the cyber capabilities of Islamic terrorist groups. Hanover, NH: Institute for Security Technology Studies at Dartmouth College.

Thomas, T.L. (2003, Spring). Al Qaeda and the Internet: The danger of cyberplanning. Parameters, 112–123.

Trybula, W.J. (1999). Text mining. In M.E. Williams (Ed.), Annual review of information science and technology (Vol. 34, pp. 385–419). Medford, NJ: Information Today, Inc.

Tsfati, Y., & Weimann, G. (2002). retrieved March 26, 2008 from http://www.terrorism.com/, Terror on the Internet. Studies in Conflict & Terrorism, 25, 317–332.

Xu, J., & Chen, H. (2005). Criminal network analysis and visualization. Communications of the ACM, 48(6), 100–107.

Young, F.W. (1987). Multidimensional scaling: History, theory, and applications. Hillsdale, NJ: Lawrence Erlbaum Associates.

Zhu, B., & Chen, H. (2005). Chapter 4: Information Visualization. In B. Cronin (Ed.), Annual Review of Information Science and Technology (Vol. 39, pp. 139–177). Medford, NJ: Information Today, Inc.

DOI: 10.1002/asi