A Focused Crawler for Dark Web Forums

Tianjun Fu
Department of Management Information Systems, University of Arizona, Tucson, AZ 85721. E-mail: futj@email.arizona.edu

Ahmed Abbasi
Management Information Systems, Sheldon B. Lubar School of Business, University of Wisconsin–Milwaukee, Milwaukee, WI 53201. E-mail: abbasi@uwm.edu

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

The unprecedented growth of the Internet has given rise to the Dark Web, the problematic facet of the Web associated with cybercrime, hate, and extremism. Despite the need for tools to collect and analyze Dark Web forums, the covert nature of this part of the Internet makes traditional Web crawling techniques insufficient for capturing such content. In this study, we propose a novel crawling system designed to collect Dark Web forum content. The system uses a human-assisted accessibility approach to gain access to Dark Web forums. Several URL ordering features and techniques enable efficient extraction of forum postings. The system also includes an incremental crawler coupled with a recall-improvement mechanism intended to facilitate enhanced retrieval and updating of collected content. Experiments conducted to evaluate the effectiveness of the human-assisted accessibility approach and the recall-improvement-based, incremental-update procedure yielded favorable results. The human-assisted approach significantly improved access to Dark Web forums while the incremental crawler with recall improvement also outperformed standard periodic- and incremental-update approaches. Using the system, we were able to collect over 100 Dark Web forums from three regions. A case study encompassing link and content analysis of collected forums was used to illustrate the value and importance of gathering and analyzing content from such online communities.

Introduction

The Internet acts as an ideal method for information and propaganda dissemination (Gustavson & Sherkat, 2004; Whine, 1997). Computer-mediated communication offers a quick, inexpensive, and anonymous means of communication for extremist groups (Crilley, 2001). Extremist groups frequently use the Web to promote hatred and violence (Glaser, Dixit, & Green, 2002). This problematic facet of the Internet is often referred to as the *Dark Web* (Chen, 2006). Extremist forums hidden deep within the Internet are an important component of the Dark Web. Many researchers (e.g., Burris, Smith, & Strahm, 2000; Schafer, 2002) have stated the need for collection and analysis of Dark Web forums. Dark Web materials have important implications for intelligence and security-informatics-related application (Chen, 2006). The collection of such content also is important for studying and understanding the diverse social and political views present in these online communities.

The unprecedented growth of the Internet has resulted in considerable focus on Web crawling/spidering techniques in recent years. Crawlers are defined as "software programs that traverse the World Wide Web information space by following hypertext links and retrieving web documents by standard HTTP protocol" (Cheong, 1996, p. 82). They are programs that can create a local collection or index of large volumes of Web pages (Cho & Garcia-Molina, 2000). Crawlers can be used for general-purpose search engines or for topic-specific collection building. The latter are referred to as *focused* or *topic-driven crawlers* (Chakrabarti, Van Den Berg, & Dom, 1999; Pant, Srinivasan, & Menczer, 2002).

There is a need for a focused crawler that can collect Dark Web forums. Such efforts can create research testbeds which can enhance our understanding of these online communities. Many previous focused crawlers have concentrated on collecting static English Web pages from the "Surface Web." A Dark Web forum focused crawler faces several design challenges. One major concern is *accessibility*. Web forums

Received August 27, 2007; revised January 17, 2010; accepted January 18, 2010

 DOI: 10.1002/asi.21323

are dynamic and often require memberships. They are part of the "Hidden Web" (Florescu, Levy, & Mendelzon, 1998; Raghavan & Garcia-Molina, 2001), which is not easily accessible through normal Web navigation or standard crawling. There also are *content-richness* considerations. Dark Web forums contain rich content used for routine communication and propaganda dissemination (Abbasi & Chen; 2005; Zhou, Reid, Qin, Chen, & Lai, 2005). These forums contain static and dynamic text files, archive files, and various forms of multimedia (e.g., images, audio, and video files). Collection of such diverse-content types introduces many unique challenges not encountered with standard spidering of indexable (i.e., text-based) files. Another important consideration is *collection recall* because detected crawlers may be blocked. On a related note, a Dark Web forum crawler also must assess the merits of various *collection-update strategies*.

In this study, we propose the development of a focused crawler that can collect Dark Web forums. Our spidering system uses breadth and depth first (BFS and DFS) traversal based on URL tokens, anchor text, and link levels, for crawl space URL ordering (i.e., ranking of URLs in the spidering queue). We also utilize incremental crawling for collection updating using wrappers to identify updated content. The system also includes design elements intended to overcome the previously mentioned accessibility, multilingual, and content-richness challenges. For accessibility, we use a human-assisted approach (Raghavan & Garcia-Molina, 2001) for attaining Dark Web forum membership. Our system also includes tailored spidering parameters and proxies for each forum to improve accessibility. The crawler uses language-independent features for crawl space URL ordering to negate any complications attributable to the presence of numerous languages. We also incorporate iterative collection and relevance feedback mechanisms to facilitate enhanced collection of multimedia content.

The remainder of the article is organized as follows. First, a review of related work on focused and hidden Web crawling is given. The section following the review describes research gaps and our related research questions. We then present a research design geared toward addressing those questions. A detailed description of our Dark Web forum spidering system is presented, followed by the experimental results evaluating the efficacy of our human-assisted approach for gaining access to Dark Web forums as well as the incremental-update procedure that uses recall improvement (i.e., a mechanism that facilitates enhanced collection of Dark Web forum content). This section also highlights the Dark Web forum collection statistics for data gathered using the proposed system. A case study conducted to illustrate the value of the collected Dark Web forums for content analysis is detailed, and then concluding remarks are given.

Related Work: Focused and Hidden Web Crawlers

Focused crawlers "seek, acquire, index, and maintain pages on a specific set of topics that represent a narrow segment of the web" (Chakrabarti et al., 1999). The need to collect high-quality, domain-specific content results in several important characteristics for such crawlers that also are relevant to collection of Dark Web forums. Some of these characteristics are specific to focused and/or Hidden Web crawling while others could be relevant to all types of spiders. We review previous research pertaining to these important considerations, which include accessibility, collection type and content richness, URL ordering features and techniques, and collection-update procedures.

Before describing each of these issues in greater detail, we briefly discuss their importance for Dark Web forum crawling. Accessibility is an important consideration because Dark Web forums often require membership to access member postings (Chen, 2006). Furthermore, Dark Web forums are rich in multimedia content, including images and videos (Abbasi & Chen, 2005; Zhou et al., 2005). URL ordering features and techniques ensure that only the desired pages are collected, and in the most efficient manner (Guo, Li, Zhang, & Zhang, 2006). Different collection-update procedures have important implications for overall collection recall.

Accessibility

Most search engines cover what is referred to as the "publicly indexable Web" (Lawrence & Giles, 1999; Raghavan & Garcia-Molina, 2000). This is the part of the Web easily accessible with traditional Web crawlers (Sizov, Graupmann, & Theobald, 2003). As noted by Lawrence and Giles (1999), a large portion of the Internet is dynamically generated. Such content typically requires users to have prior authorization, fill out forms, or register (Raghavan & Garcia-Molina, 2000). This covert side of the Internet is commonly referred to as the *Hidden/Deep/Invisible Web*. Hidden Web content is often stored in specialized databases (Lin & Chen, 2002). For example, the IMDB movie-review database contains a plethora of useful information regarding movies; yet, standard crawlers cannot access this information (Sizov et al., 2003). One study found that the Invisible Web contained 400 to 550 times the information present in the traditional surface Web (Bergman, 2000; Lin & Chen, 2002).

Two general strategies have been introduced to access the Hidden Web via automated Web crawlers. The first approach entails the use of automated form-filling techniques. Several different automated query-generation approaches for querying such "Hidden Web" databases and fetching the dynamically generated content have been proposed (e.g., Barbosa & Freire, 2004; Ntoulas, Zerfos, & Cho, 2005). Other techniques keep an index of Hidden Web search engines and redirect user queries to them (Lin & Chen, 2002) without actually indexing the Hidden databases. However, many automated approaches ignore/exclude collection or querying of pages requiring login (e.g., Lage, Da Silva, Golgher, & Laender, 2002). Thus, automated form-filling techniques seem problematic for Dark Web forums where login is often required.

A second alternative for accessing the Hidden Web is a task-specific, human-assisted approach (Raghavan &

DOI: 10.1002/asi

Garcia-Molina, 2000). This approach provides a semi-automated framework that allows human experts to assist the crawler in gaining access to Hidden content. The amount of human involvement is dependent on the complexity of the accessibility issues faced. For example, many simple forms asking for name, e-mail address, and so on can be automated with standardized responses. More complex questions require greater expert involvement. Such an approach seems more suitable for the Dark Web, where the complexity of the access process can vary significantly.

Collection Type

Previous focused crawling research has been geared toward collecting Web sites, blogs, and Web forums. There has been considerable research on collection of standard Web sites and pages relating to a particular topic, often for vertical portal building. Srinivasan, Mitchell, Bodenreider, Pant, and Menczer (2002) and Chau and Chen (2003) fetched biomedical content from the Web. Sizov et al. (2003) collected Web pages pertaining to handicrafts and movies. Pant et al. (2002) evaluated their topic crawler on various keyword queries (e.g., "recreation").

There also has been work on collecting Weblogs. BlogPulse (Glance, Hurst, & Tomokiyo, 2004) is a blog-analysis portal. The site contains analysis of key discussion topics/trends for roughly 100,000 spidered Weblogs. Such blogs also can be useful for marketing intelligence (Glance et al., 2005a). Blogs containing product reviews analyzed using sentiment analysis techniques can provide insight into how people feel about various products.

Web forum crawling presents a unique set of difficulties. Discovering Web forums is challenging due to the lack of a centralized index (Glance et al., 2005a). Furthermore, Web forums require information-extraction wrappers for derivation of metadata (e.g., authors, messages, timestamps, etc.). Wrappers are important for data analysis and incremental crawling (i.e., re-spidering only those threads containing newly posted messages). Incremental crawling is discussed in greater detail in the "Collection-Update" section. There has been limited research on Web forum spidering. BoardPulse (Glance et al., 2005a) is a system for harvesting messages from online forums. It has two components: a crawler and a wrapper. Limanto, Giang, Trung, Huy, and He (2005) developed a Web forum information-extraction engine that includes a crawler, wrapper generator, and extractor (i.e., application of generated wrapper). Yih, Chang, and Kim (2004) created an online forum-mining system composed of a crawler and an information extractor for mining deal forums: forums where participants share information regarding deals or promotional events offered by online stores. The NetScan project (Smith, 2002) collected and visualized millions of pages from USENET newsgroups. RecipeCrawler (Li, Meng, Wang, & Li, 2006) is a focused crawler that collects cooking recipes from various information sources, including Web forums. RecipeCrawler uses the tree edit distance scores between Web pages to rank them in the crawl space (Li et al., 2006). Similar to BoardPulse (Glance et al., 2005a), RecipeCrawler also uses a crawler and a wrapper for extracting recipe information. Guo et al. (2006) proposed a board forum crawler that traverses board-based Web forums in a hierarchical manner analogous to that used by actual users manually browsing the forum. Their crawler uses Web page and URL token text features coupled with a rule-based ranking mechanism to order URLs in the crawl space. Each of the aforementioned Web forum crawlers only collected pages from the Surface Web. There has been no prior research on collecting Dark Web forums, which requires the use of mechanisms for improving forum accessibility and collection recall.

Content Richness

The Web is rich in indexable and multimedia files. Indexable files include static text files (e.g. HTML, Word, and PDF documents) and dynamic text files (e.g., .asp, .jsp, .php). Multimedia files include images, animations, audio, and video files. Difficulties in indexing make multimedia content difficult to accurately collect (Baeza-Yates, 2003). Multimedia file sizes are typically significantly larger than are indexable files, resulting in longer download times and frequent timeouts. Heydon and Najork (1999) fetched all MIME file types (including image, video, audio, and .exe files) using their Mercator crawler. They noted that collecting such files increased the overall spidering time and doubled the average file size as compared to just fetching HTML files. Consequently, many previous studies have ignored multimedia content altogether (e.g., Pant et al., 2002).

URL Ordering Features

Aggarwal, Al-Garawi, and Yu (2001) noted four categories of features for crawl space URL ordering. These include links, URL and/or anchor text, page text, and page levels. *Link*-based features have been used considerably in previous research. Many studies have used inlinks/backlinks and outlinks (Pant et al., 2002). Sibling links (Aggarwal et al., 2001) consider sibling pages (i.e., ones with shared parent in link). Context graphs (Diligenti, Coetzee, Lawrence, Giles, & Gori, 2000) derive backlinks for each seed URL and use these to construct a multilayer context graph. Such graphs can be used to extract paths leading up to relevant nodes (i.e., target URLs). Focused/topical crawlers often use bag-of-words (BOW) found in the Web *page text* (Aggarwal et al., 2001; Pant et al., 2002). For instance, Srinivasan et al. (2002) used BOW for biomedical-text categorization in their focused crawler. While page text features are certainly very effective, they also are language-dependent and can be harder to apply in situations where the collection is composed of pages in numerous languages. Other studies also have used *URL/anchor text*. Word tokens found within the URL anchor have been used effectively to help control the crawl space (Ester, Grob, & Kriegel, 2001). URL tokens also have been incorporated in previous focused crawling research

DOI: 10.1002/asi

(Aggarwal et al., 2001; Ester et al., 2001). Another important category of features for URL ordering is page *levels*. Diligenti et al. (2000) trained text classifiers to categorize Web pages at various levels away from the target. They used this information to build path models that allowed consideration of irrelevant pages as part of the path to attain target pages. A potential path model may consider pages one or two levels away from a target, known as tunneling (Ester et al., 2001). Ester et al. (2001) used the number of slashes "/" or levels from the domain as an indicator of URL importance. They argued that pages closer to the main page are likely to be of greater importance.

URL Ordering Techniques

Previous research has typically used breadth, depth, and best first search for URL ordering. Depth first (DFS) has been used in crawling systems such as Fish Search (De Bra & Post, 1994). Breadth first (BFS) (Cho, Garcia-Molina, & Page, 1998; Ester et al., 2001; Najork & Wiener, 2001) is one of the simplest strategies. It has worked fairly well in comparison with more sophisticated best-first search strategies on certain spidering tasks (Cho et al., 1998; Najork & Wiener, 2001); however, BFS is typically not employed by focused crawlers that are concerned with identifying topic-specific Web pages using the aforementioned URL ordering features. For focused crawling, BFS has been outperformed by various best-first strategies (Menczer, Pant, & Srinivasan, 2004).

Best-first uses some criterion for ranking URLs in the crawl space, such as *link analysis*, *text analysis*, or a combination of the two (Menczer, 2004). Numerous link-analysis techniques have been used for URL ordering. Cho et al. (1998) evaluated the effectiveness of Page Rank and backlink counts. Pant et al. (2002) also used Page Rank. Aggarwal et al. (2001) used the number of relevant siblings. They considered pages with a higher percentage of relevant siblings more likely to also be relevant. Sizov et al. (2003) used the HITS algorithm to compute authority scores whereas Chakrabarti et al. (1999) used a modified version of HITS. Chau and Chen (2003) used a Hopfield net crawler that collected pages related to the medical domain based on link weights.

Text-analysis methods include similarity scoring approaches and machine learning algorithms. Aggarwal et al. (2001) used similarity equations with page content and URL tokens. Others have used the vector space model and cosine similarity measure (Menczer et al., 2004; Pant et al., 2002; Srinivasan et al., 2002). Sizov et al. (2003) used support vector machines (SVM) with BOW for document classification. Srinivasan et al. (2002) used BOW and link structures with a neural net for ordering URLs based on the prevalence of biomedical content. Chen, Chung, Ramsey, and Yang (1998a, 1998b) used a genetic algorithm to order the URL crawl space for the collection of topic-specific Web pages based on BOW representations of pages. Chakrabarti, Punera, and Subramanyam (2002) incorporated an apprentice learner, which evaluated the utility of an outlink by comparing its HTML source code against prior training instances. Recent studies have compared the effectiveness of various machine learning classification algorithms such as Naïve Bayes, SVM, and Neural Networks, for focused crawling (Pant & Srinivasan, 2005, 2006).

Collection-Update Procedure

Two approaches for collection updating are periodic and incremental crawling (Cho & Garcia-Molina, 2000). *Periodic* Web forum crawling entails eventually updating the collection by re-spidering all forum pages (e.g., Guo et al., 2006). This is commonly done because it is often easier than figuring out which Web forum pages to refresh, especially since the if-modified-since header does not provide information about which boards, subboards, and threads within a Web forum have been updated. Although periodic crawling is simpler from a crawler design/development perspective, it makes the collection process time consuming and inefficient. Alternatively, gathering multiple versions of a collection may improve overall recall. *Incremental* Web forum crawlers gather new and updated content by fetching only those boards and threads that have been updated since the forum was last collected (Glance et al., 2005a; Li et al., 2006; Yih et al., 2004). Incremental Web forum crawlers require the use of a wrapper that can parse out the "last updated" dates for boards and threads (Yih et al., 2004). This information is typically contained in the page's body text.

Summary of Previous Research

Table 1 provides a summary of selected previous research on focused crawling. The majority of studies have focused on collection of indexable files from the Surface Web. Only a few studies have performed focused crawling on the Hidden Web. Similarly, only a few studies have collected content from Web forums. Most previous research on focused crawling has used BOW, link, or URL token features coupled with a best-first search strategy for crawl space URL ordering. Furthermore, most prior research also has ignored the multilingual dimension, only collecting content in a single language (usually English). Collection of Dark Web forums entails retrieving rich content (including indexable and multimedia files) from the Hidden Web in multiple languages. Dark Web forum crawling is therefore at the cross section of several important areas of crawling research, many of which have received limited attention in prior research. The following section summarizes these important research gaps and provides a set of related research questions which are addressed in the remainder of the article.

Research Gaps and Questions

Based on our review of previous literature, we have identified several important research gaps.

DOI: 10.1002/asi

TABLE 1. Selected previous research on focused crawling.

System name and study	Access	Collection type	Content richness	URL ordering features	URL ordering techniques
GA Spider (Chen et al., 1998a, 1998b)	Surface Web	Topic-specific Web pages	Indexable files only	BOW	Best-first: Genetic algorithm
Focused Crawler (Chakrabarti et al., 1999)	Surface Web	Topic-specific Web pages	Indexable files only	BOW and links	Hypertext classifier and modified HITS algorithm
Context Focused (Diligenti et al., 2000)	Surface Web	Topic-specific Web pages	Indexable files only	BOW and context graphs	Best-first: Vector space, Naïve Bayes, and path models
Intelligent Crawler (Aggarwal et al., 2001)	Surface Web	Topic-specific Web pages	Indexable files only	BOW, URL tokens, anchor text, links	Best-first: Similarity scores and link analysis
Ariadne (Ester et al., 2001)	Surface Web	Topic-specific Web pages	Indexable files only	BOW, URL tokens, anchor text, links, user feedback, levels	Relevance scoring and text classifier
Hidden Web Exposer (Raghavan; Garcia-Molina, 2001)	Hidden Web	Dynamic search forms	Indexable files only	URL Tokens	Rule-based: Crawler stayed within target sites
InfoSpiders (Srinivasan et al., 2002)	Surface Web	Biomedical pages and documents	Indexable files only	BOW and Links	Best-first: Vector space model and Neural Net
NetScan (Smith, 2002)	Surface Web	USENET Web forums	Indexable files only	n/a	n/a
Topic Crawler (Pant et al., 2002)	Surface Web	Topic-specific Web pages	Indexable files only	BOW	Best-n-First: Vector space model
Hopfield Net Crawler (Chau; Chen, 2003)	Surface Web	Medical-domain Web pages	Indexable files only	Links	Best-first: Hopfield Net
BINGO! (Sizov et al., 2003)	Surface and Hidden Webs	Handicraft and movie Web pages	Indexable files only	BOW and links	Best-first: SVM and HITS
BlogPulse (Glance et al., 2004)	Surface Web	Weblogs for various topics	Indexable files only	Weblog text	Differencing algorithm
Hot Deal Crawler (Yih et al., 2004)	Surface Web	Online-deal forums	Indexable files only	URL tokens, thread date	Date comparison
BoardPulse (Glance et al., 2005a)	Surface Web	Product Web forums	Indexable files only	URL tokens, thread date	Wrapper learning of site structure
Web Forum Spider (Limanto et al., 2005)	Surface Web	Web forums	Indexable files only	Web page text and URL tokens	Machine learning classifier
Board Forum Crawler (Guo et al., 2006)	Surface Web	Board Web forums	Indexable files only	Web page text and URL tokens	Rule-based: Uses URL tokens and text
RecipeCrawler (Li et al., 2006)	Surface Web	Recipe sites, blogs, and Web forums	Indexable files only	Web page text	Best-first: Tree edit distance similarity scores

Focused Crawling of the Hidden Web

There has been limited focused crawling work on the Hidden Web. Most focused crawler studies developed crawlers for the Surface Web (Raghavan & Garcia-Molina, 2001). Prior Hidden Web research mostly has focused on automated form filling or query redirection to hidden databases (i.e., accessibility issues). There has been little emphasis on building topic-specific Web page collections from these hidden sources. We are not aware of any attempts to automatically collect Dark Web content pertaining to hate and extremist groups.

Content Richness

Most previous research has focused on indexable (i.e., text-based) files. Large multimedia files large (e.g., videos) can be hundreds of megabytes. This can cause connection timeouts or excessive server loads, resulting in partial/incomplete downloads. Furthermore, the challenges in indexing multimedia files pose problems. It is difficult to assess the quality of collected multimedia items. As Baeza-Yates (2003) noted, automated multimedia indexing is more of an image-retrieval challenge than an information-retrieval problem. Nevertheless, given the content richness of the Internet in general and the Dark Web specifically (Chen, 2006), there is a need to capture multimedia files.

Collection Recall Improvement

Prior crawling research has not addressed the issues associated with collecting content in adversarial settings. Dark Web forum spidering involves avoiding detection since it could have obvious ramifications for collection recall.

DOI: 10.1002/asi

Web Forum Collection-Update Strategies

There has been considerable research on evaluating various collection-update strategies for Web sites (e.g., Cho & Garcia-Molina, 2000); however, there has been little work done on comparing the effectiveness of periodic versus incremental crawling for Web forums. Most Web forum research has assumed an incremental approach. Given the accessibility concerns associated with Dark Web forums, periodic and incremental approaches both provide varying benefits. Periodic crawlers can improve collection recall by allowing multiple attempts at capturing previously uncollected pages. This may be less of a concern for Surface Web forums, but is important for the Dark Web. In contrast, incremental crawlers can improve collection efficiency and reduce redundancy. There is a need to evaluate the effectiveness of periodic and incremental crawling applied to Dark Web forums.

Research Questions

Based on the gaps just described, we propose the following research questions:

RQ1: How effectively can Dark Web forums be *identified* and *accessed* for collection purposes?

RQ2: How effectively can Dark Web content (indexable and multimedia) be collected?

RQ3: Which *collection-update procedure* (periodic or incremental) is more suitable for Dark Web forums? How can recall improvement further enhance the update process?

RQ4: How can *analysis of extracted information* from Dark Web forums improve our understanding of these online communities?

Research Design

Proposed Dark Web Forum Crawling System

In this study, we propose a Dark Web forum spidering system. Our proposed system consists of an accessibility component that uses a human-assisted registration approach to gain access to Dark Web forums. Our system also utilizes multiple dynamic proxies and forum-specific spidering parameter settings to maintain forum access.

Our URL ordering component uses language-independent URL ordering features to allow spidering of Dark Web forums across languages. We plan to focus on groups from three regions: U.S. Domestic, Middle East, and Latin America/Spain. Additionally, a rule-based URL ordering technique coupled with BFS and DFS crawl space traversal is utilized. Such a technique is employed to minimize the amount of irrelevant Web pages collected.

We also propose utilizing an incremental crawler that uses forum wrappers to determine the subset of threads that need to be collected. Our system will include a recall-improvement procedure that parses the spidering log and reinserts incomplete downloads into the crawl space. Finally, the system features a collection analyzer that checks multimedia files for duplicate downloads and generates collection statistics at the forum, region, and overall collection levels.

Accessibility

As noted by Raghavan and Garcia-Molina (2001), the most important evaluation criterion for Hidden Web crawling is how effectively the content was accessed. They developed an accessibility metric as follows: *databases accessed/total attempted*. We intend to evaluate the effectiveness of the task-specific, human-assisted approach in comparison with not using such a mechanism. Specifically, we also would like to evaluate our system's ability to access Dark Web forums. This translates into measuring the percentage of attempted forums accessed.

Recall-Improvement Mechanism

Given the collection challenges regarding Dark Web forums, we propose the use of a recall-improvement mechanism that controls various spidering settings for enhanced collection recall. The recall-improvement component is intended to control key spidering parameters such as the number of spiders per forum, the proxies per spider, and other pertinent spidering settings. It is essentially a heuristic used to counterattack crawler detection.

Incremental Crawling for Collection Updating

We plan to evaluate the effectiveness of our proposed incremental crawler in comparison with periodic crawling. The incremental crawler will obviously be more efficient in terms of spidering time and data redundancy; however, a periodic crawling approach gets multiple attempts to collect each page, which can improve overall collection recall. Evaluation of both approaches is intended to provide additional insight into which collection-update technique is more suitable for Dark Web forum spidering.

System Design

Based on our research design, we implemented a focused crawler for Dark Web forums. Our system consists of four major components (shown in Figure 1):

- *Forum Identification*: identifies the list of extremist forums to spider;
- *Forum preprocessing*: includes accessibility and crawl space traversal issues as well as forum wrapper generation;
- *Forum spidering*: consists of an incremental crawler and recall-improvement mechanism;
- *Forum storage and analysis*: stores and analyzes the forum collection.

DOI: 10.1002/asi

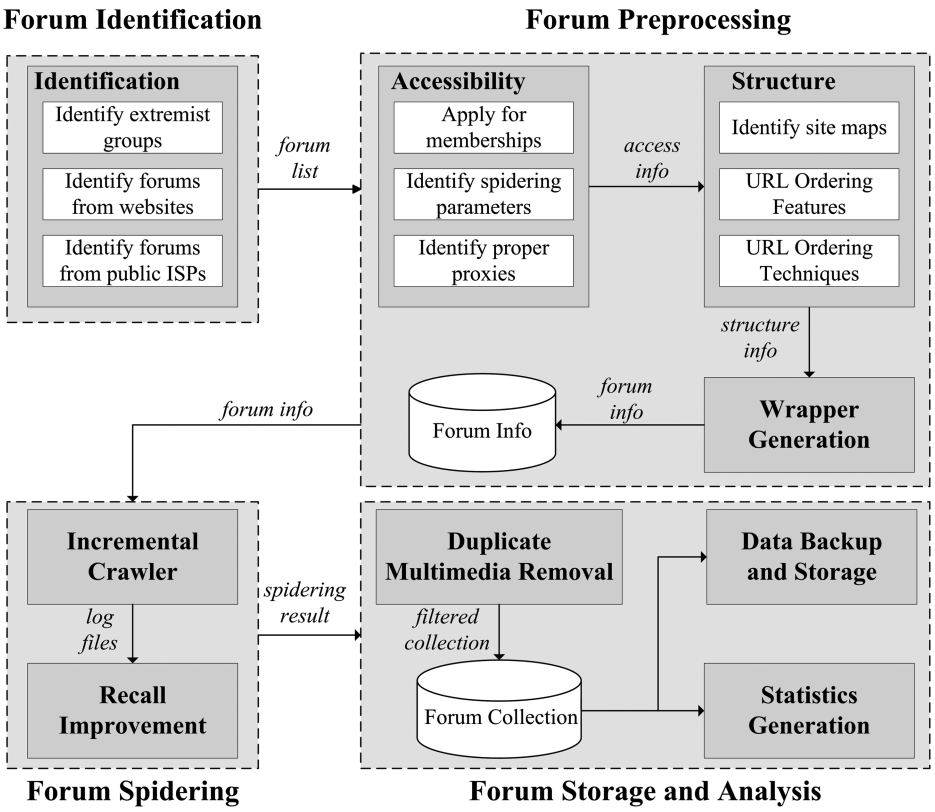

FIG. 1. Dark Web forum crawling system design.

Forum Identification

The forum identification phase has three steps.

Identify extremist groups. Sources for the U.S. domestic extremist groups include the Anti-Defamation League (ADL), FBI, Southern Poverty Law Center (SPLC), Militia Watchdog (MW), and the Google Web Directory (GD) (as a supplement). Sources for the international extremist groups include the U.S. Committee for a Free Lebanon (USCFAFL), Counter-Terrorism Committee (CTC) of the U.N. Security Council (UN), U.S. State Department report (US), *Official Journal of the European Union* (EU) as well as government reports from the United Kingdom (UK), Australia (AUS), Japan (JPN), and People's Republic of China (CHN). Due to regional and language constraints, we chose to focus on groups from three areas: North America (English), Latin America (Spanish), and the Middle East. These groups are all significant for their sociopolitical importance. Furthermore, collection and analysis of Dark Web content from these three regions can facilitate a better understanding of the relative social and cultural differences between these groups. In addition to obvious linguistic differences, groups from these regions also display different Web design tendencies and usage behavior (Abbasi & Chen, 2005), which provide a unique set of collection and analysis challenges.

Identify forums from extremist Web sites We identify an initial set of extremist group URLs, and then use link analysis for expansion purposes as shown in Figure 2. The initial set of URLs is identified from three sources: First, we use search engines coupled with a lexicon containing extremist organization name(s), leader(s)' and key members' names, slogans, and special keywords used by extremists. Second, we utilize government reports. Finally, we reference research centers. A link analysis approach is used to expand the initial list of URLs. We incorporate a backlink search using Google, which has been shown to be effective in prior research (Diligenti et al., 2000). Outlinks for initial seed URLs as well as their inlinks identified using Google also are collected. The identified Web forums are manually checked by domain experts. Only verified Dark Web forums are collected.

Identify forums hosted on major Web sites. We also identify forums hosted by other Web sites and public Internet service providers (ISPs) that are likely to be used by Dark Web groups. For example, public ISPs such as MSN groups, AOL Groups, and so o are searched with our Dark Web domain lexicon for a list of potential forums.

The aforementioned three steps help identify a seed set of Dark Web forums. Once the forums have been identified, several important preprocessing issues must be resolved before spidering. These include accessibility concerns and identification of forum structure, which is necessary to develop proper features and techniques for managing the crawl space.

Forum Preprocessing

The forum preprocessing phase has three components: accessibility, structure, and wrapper generation. The

DOI: 10.1002/asi

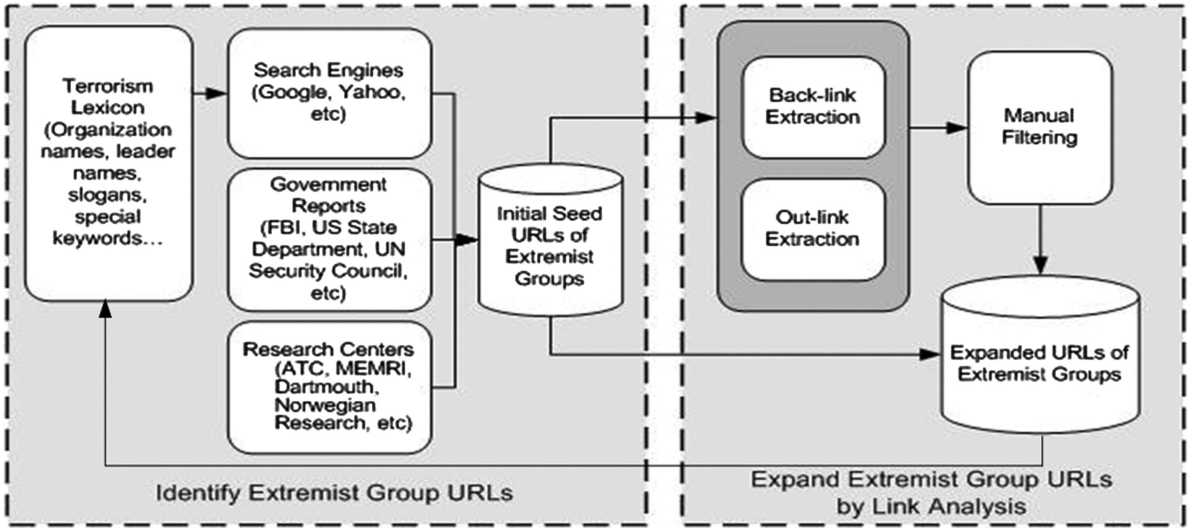

FIG. 2. Dark Web forum identification process.

accessibility component deals with acquiring and maintaining access to Dark Web forums. The structure component is designed to identify the forum URL mapping and devise the crawl space URL ordering using the relevant features and techniques.

Forum accessibility

Apply for membership. Many Dark Web forums (~30–40%) do not allow anonymous access (Zhou et al., 2006). To access and collect information from those forums, one must create a user ID and password, send an application request to the Web master, and wait to get permission/registration to access the forum. In certain forums, Web masters are very selective. It can take a couple of rounds of e-mail to get access privilege. For such forums, human expertise is invaluable. Nevertheless, in some cases, access cannot be attained. Based on our experience with hundreds of Dark Web forums, approximately 10% cannot be accessed at all.

Identify appropriate spidering parameters. Spidering parameters such as number of connections, download intervals, timeout, speed, and so on, need to be set appropriately according to server and network limitations and the various forum-blocking mechanisms. Dark Web forums are rich in terms of their content. Multimedia files are often fairly large in volume (particularly compared to indexable files). The spidering parameters should be able to handle the downloading of larger files from slow servers; however, one may still be blocked based on the IP address. Therefore, we use proxies to increase not only our recall but also our anonymity.

Identify appropriate proxies. We use three types of proxy servers. *Transparent* proxy servers are those that provide anyone with your real IP address. *Translucent* proxy servers hide your IP address or modify it in some way to prevent the target server from knowing about it; however, they let anyone know that you are surfing through a proxy server. *Opaque* proxy servers (i.e., preferred) hide your IP address and do not let anyone know that you are surfing through a proxy server. There are several criteria for proxy server selection, including the latency (the smaller the better), reliability (the higher the better), and bandwidth (the faster the better). We update our list of proxy servers periodically from various sources, including free proxy providers such as www.xroxy.com and www.proxy4free.com Additionally, the crawler uses a Web browser user agent string and does not follow the robot exclusion protocol, though nearly none of the Dark Web forums collected had a robots.txt file.

Forum structure

Identify site maps. We first identify the site map of the forum based on the forum software packages. Glance et al. (2005a) noted that although there are only a handful of commonly used forum software packages, they are highly customizable. Forums typically have hierarchical structures with boards, threads, and messages (Glance et al., 2005a; Yih et al., 2004). They also contain considerable additional information such as message-posting interfaces, search, printing, advertisement, and calendar pages (all irrelevant from our perspective). Furthermore, forums contain multiple views of member postings (e.g., sorted by author, date, topic, etc.). Collecting these duplicate views can introduce considerable redundancy into the collection, dramatically increase collection time, increase the likelihood of being detected/blocked, and result in spider traps (Guo et al., 2006). The URL ordering features and techniques are important to allow the crawler to collect only the desired pages (i.e., ones containing nonredundant message postings) in the most efficient manner.

URL ordering features. Our spidering system uses two types of language-independent URL ordering features: URL tokens and page levels. With respect to *URL tokens*, for Web forums, we are interested in URLs containing words such

DOI: 10.1002/asi

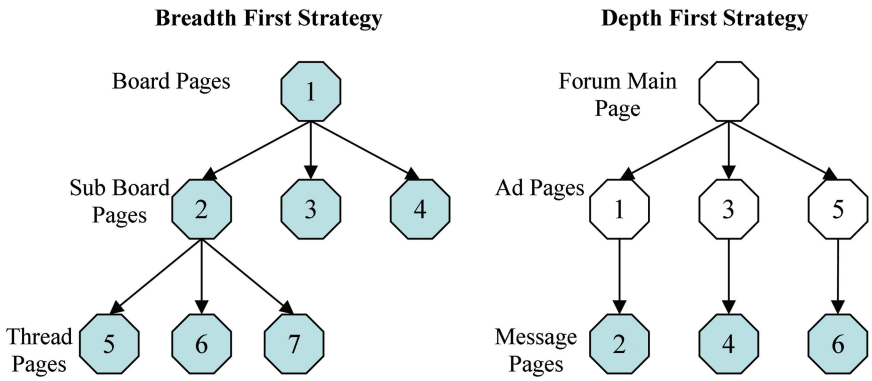

FIG. 3. URL traversal strategies.

as "board," "thread," "message," and so on (Glance et al., 2005a). Additional relevant URL tokens include domain names of third-party, file-hosting Web sites. These third parties often contain multimedia files. File-extension tokens (e.g., ".jpg" and ".wmv") also are important. URLs that contain phrases such as "sort=voteavg" and "goto=next" also are found in relevant pages; however, these are not unique to board, thread, and message pages, and such tokens thus are not considered significant. The set of relevant URL tokens differs based on the forum software being used. Such tokens are language-independent, yet software-specific.

Page *levels* also are important, as evidenced by prior focused crawling research (Diligenti et al., 2000; Ester et al., 2001). URL-level features are important for Dark Web forums due to the need to collect multimedia content. Multimedia files are often stored on third-party host sites that may be a few levels away from the source URL. To capture such content, we need to use a rule-based approach that allows the crawler to go a few additional levels. For example, if the URL or anchor text contains a token that is a multimedia file extension or the domain name for a common third-party file carrier, we want to allow the crawler to "tunnel" a few links.

URL ordering techniques. As mentioned in the previous section, we use rules based on URL tokens and levels to control the crawl space. Moreover, to adapt to different forum structures, we need to use different crawl space traversal strategies. BFS is used for board page forums whereas DFS is used for ISP forums. DFS is necessary for many ISP forums due to the presence of ad pages that periodically appear within these forums. When such an ad page appears, it must be traversed to get to the message pages (Typically, the ad pages have a link to the actual message page.) For DFS, a preset depth limit is used to avoid spider traps. Figure 3 illustrates how the BFS and the DFS are performed for each forum type. Only the colored pages are fetched while the number indicates the order in which the pages are traversed by the crawler. One level of tunneling is allowed to fetch multimedia content hosted on third-party host Web sites outside of the Web forum. A parser analyzes the URL tokens and anchor text for multimedia keywords. These include (a) the domain names for popular third-party hosts (b) multimedia file extensions such as .wmv, and .avi; (c) terms appearing in the anchor text, such as "video," "movie," and "clip." Only URLs containing attributes from the aforementioned feature categories are tunneled.

Wrapper generation. Forums are dynamic archives that keep historical messages. It is beneficial to only spider newly posted content when updating the collection. This is achieved by generating wrappers that can parse Web forum board and thread pages (Glance et al., 2005s). Board pages tell us when each thread was last updated with new messages. Using this information, one may re-spider only those thread pages containing new postings (Guo et al., 2006). Web forums generally use a dozen or so popular software for creating Web forums, including vBulletin, Crosstar, DCForum, ezBoard, Invision, phpBB, and so on. We developed wrappers based on these forums' templates, as was done by previous research (e.g., Glance et al., 2005a; Guo et al., 2006). The wrappers parse out the board pages and compare the posting dates for the most recent messages for all threads in a forum against the dates when the threads were last collected. If the thread has been updated, an incremental crawler retrieves all new pages (i.e., it fetches all pages containing messages posted since the thread was last spidered). The use of an incremental crawler via wrappers is an efficient way to collect Web forum content (Guo et al., 2006).

Forum Spidering

Figure 4 shows the spidering process. The incremental crawler fetches only new and updated threads and messages. A log file is sent to the recall-improvement component. The log shows the spidering status of each URL. A parser is used to determine the overall status for each URL (e.g., "download complete," "connection timed out"). The parsed log is sent to the log analyzer, which evaluates all files that were not downloaded. It determines whether the URLs should be re-spidered.

Figure 5 shows sample entries from the original and parsed log. The original log file shows the download status for each file (URL). The parsed log shows the overall status as well as the reason for download failure (in the case of undownloaded

DOI: 10.1002/asi

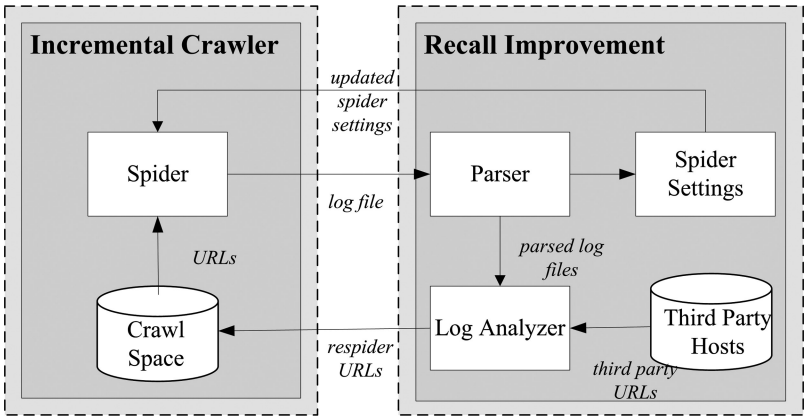

FIG. 4. Spidering process.

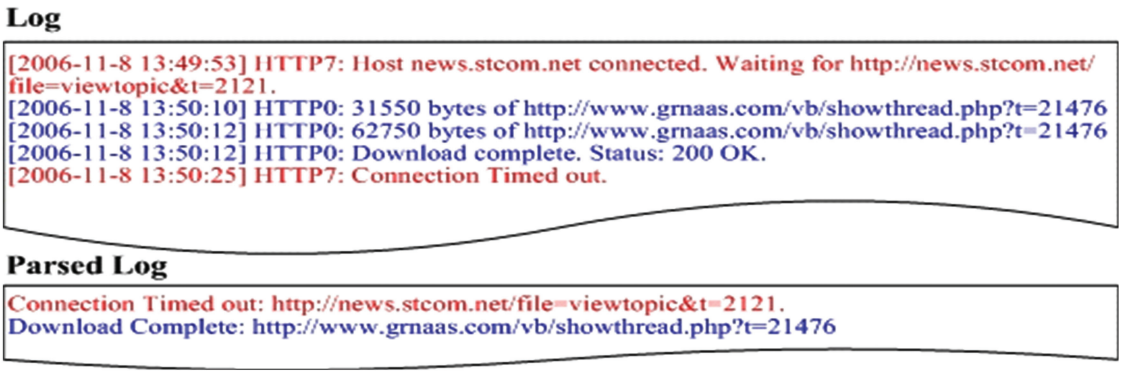

FIG. 5. Example log and parsed log entries.

files). Blue-colored entries relate to downloaded files whereas red-colored entries relate to undownloaded files. The log analyzer determines the appropriate course of action based on this cause of failure. "File Not Found" URLs are removed (not added to re-spidering list) whereas "Connection Timed Out" URLs are re-spidered. The recall-improvement phase also checks the file sizes of collected Web pages for partial/incomplete downloads. Multimedia file downloads are occasionally manually downloaded, particularly larger video files that may otherwise time out.

Once the list of re-spidering URLs has been generated, the recall-improvement mechanism adjusts important spidering settings to improve collection performance. There are several important spidering parameters that can have an impact on Dark Web forum collection recall. These include the number of spiders per forum and the total number of proxies and proxies per spider as well as the batch size (i.e., the subset of URLs to be collected at a time) and timeout interval between batches. Given the large number of potential URLs that may need to be fetched from a single forum, URLs in the crawl space are broken up into batches to alleviate forum server overload.

Spidering parameters are adjusted based on the premise that the uncollected pages (requiring re-spidering) likely failed to be retrieved due to excessive load on the forum server or as a result of being blocked by the network or forum administrator. Therefore, the recall-improvement mechanism decreases the number of spiders and URLs per batch while also increasing the number of proxies per spider and the timeout interval between batches. These spidering adjustments are made to alleviate server load and avoid blockages. The steps involved in the spidering adjustment component of the recall-improvement mechanism are shown next. The values in parentheses signify the possible range of values for that particular parameter. For instance, a new forum would initially be crawled using 60 spiders; however, if necessary, this number may eventually decrease to 1 to improve recall.

1. Decrease the number of spiders per forum by half (1–60).
2. Increase the proxy ratio (i.e., No. of proxies per spider) by 1 (1–5).
3. Decrease the number of URLs per batch by half (100–1000).
4. Increase the timeout interval between batches by 5 s (5–60).

Forum Storage and Analysis

The forum storage and analysis phase consists of a statistics generation and duplicate multimedia removal components.

DOI: 10.1002/asi

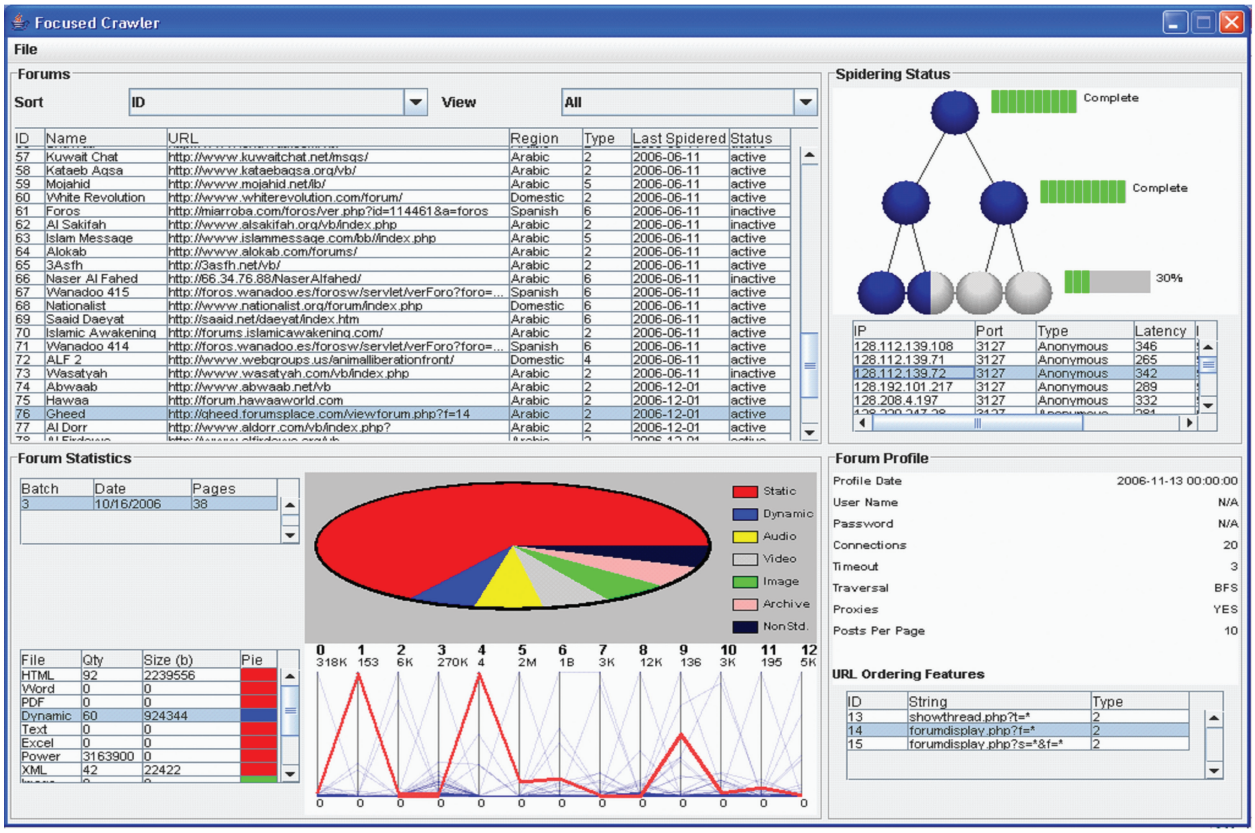

FIG. 6. Dark Web forum crawling system interface.

Statistics generation. Once files have been collected, they must be stored and analyzed. The statistics consist of four major categories:

- Indexable files: HTML, Word, PDF, Text, Excel, PowerPoint, XML, and Dynamic files (e.g., PHP, ASP, JSP).
- Multimedia files: Image, Audio, and Video files.
- Archive files: RAR, ZIP.
- Nonstandard files: Unrecognized file types.

Duplicate multimedia removal. Dark Web forums often share multimedia files, but the names of those files may be changed. Moreover, some multimedia files' suffixes are changed to other file types' suffixes, and vice versa. For example, an HTML file may be named as a ".jpg." Therefore, simply relying on file names results in inaccurate multimedia-file statistics. We use an open-source duplicate multimedia removal software tool that identifies multimedia files by the metadata encoded into the file, instead of their suffixes (i.e., file extensions). It compares files based on their Message-Digest Algorithm 5 (MD5) values, which are the same for duplicate video files collected from various Internet sources. MD5 is a widely used cryptographic hash function with a 128-bit hash value. Comparing MD5 values allows a more accurate mechanism for differentiating multimedia files than does simply comparing file names, types, and sizes. In our analysis of duplicate Dark Web multimedia files, comparing MD5 hashes found three times as many duplicates as simply relying on file names, sizes, and types.

Dark Web Forum Crawling System Interface

Figure 6 shows the interface for the proposed Dark Web Forum spidering system. The interface has four major components. The "Forums" panel in the top-left corner shows the spidering queue in a table that also provides information such as the forum name, URL, region, when it was last spidered, and whether the forum is still active. The "Spidering Status" panel in the top-right corner displays information about the percentage of board, subboard, and thread pages collected for the current forum being spidered. The "Forum Statistics" panel in the bottom-left corner shows the quantity and size of the various file types collected for each forum, using tables, pie charts, and parallel coordinates. The "Forum Profile" in the bottom-left panel shows each forum's membership information and forum spidering parameters, including the number of crawlers, URL ordering technique (i.e., BFS or DFS), and URL ordering features (e.g., URL tokens, keywords) used to control the crawl space.

Evaluation

We conducted three experiments to evaluate our system. The first experiment involved assessing the effectiveness of our human-assisted accessibility mechanism. Raghavan and Garcia-Molina (2001) noted that accessibility is the most important evaluation criterion for Hidden Web research. We describe how effectively we were able to access Dark Web forums in our collection efforts using the human-assisted

DOI: 10.1002/asi

TABLE 2. Dark Web forum accessibility statistics.

	Human-assisted accessibility			Standard spidering		
	Hosted forums	Stand-alone forums	Total forums	Hosted forums	Stand-alone forums	Total forums
Total attempted	52	67	119	52	67	119
Accessed/collected	43	66	109	25	56	71
Inaccessible	9	1	10	27	11	48
%Collected	82.69	98.51	91.60	48.08	83.58	59.66

approach in comparison with standard spidering without any accessibility mechanism.

The second experiment assessed the impact of different spidering parameter settings on collection recall. Since accessibility and recall of Dark Web forum content is a critical concern, we evaluated the impact on collection recall of using a different number of spiders per forum, proxies per spider, batch sizes, and timeout intervals between batches.

The third experiment entailed evaluating the proposed incremental spidering approach that uses recall improvement as a collection-updating procedure. We performed an evaluation of the effectiveness of periodic crawling as compared to standard incremental crawling and our incremental crawler, which uses iterative recall improvement for Dark Web forum collection updating.

For the latter two experiments, we used precision, recall, and F-measure to evaluate performance. For Web forums, relevant documents were considered to be unique Web pages containing forum postings (Glance et al., 2005a). Since Web forums are dynamic, their postings can be arranged in numerous ways (e.g., by date, by topic, by author, etc.). From a collection perspective, these views contain duplicate information: Only a single copy of each posting is desired (Guo et al., 2006). Irrelevant pages include ones containing duplicate forum postings or no forum postings at all as well as incorrectly collected pages (i.e., ones containing an HTML error code). Hence, consistent with prior forum crawling research (Glance et al., 2005a), we define precision, recall, and F-measure as follows:

a = No. of retrieved pages containing nonduplicate forum postings
b = Total no. of pages containing nonduplicate forum postings
c = No. of retrieved pages containing duplicate forum postings
d = No. of retrieved pages containing an HTML error code
e = No. of retrieved pages that do not contain forum postings

$$\text{Recall} = \frac{a}{b}$$

$$\text{Precision} = \frac{a}{a + c + d + e}$$

$$\text{F - measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Forum accessibility experiment. Table 2 presents results on our ability to access Dark Web forums with and without a human-assisted accessibility mechanism. Using the human-assisted accessibility approach, we were able to access over 82% of Dark Web forums hosted by various ISPs and virtually all of the attempted stand-alone forums. The overall results (>91% accessibility) indicate that the use of a human-assisted accessibility mechanism provided good results for Dark Web forums. In contrast, using standard spidering without any accessibility mechanism resulted in only 59.66% of the forums being accessible to collect. The largest impact of the accessibility approach occurred on the hosted forums, where lack of usage of human-assisted accessibility resulted in a 34% drop in the number of forums collected ($n = 18$).

Pairwise t tests were conducted to assess the improved access performance of the human-assisted accessibility mechanism as compared to a standard spidering scheme devoid of any special accessibility method. The improved performance was statistically significant ($\alpha = 0.01$) for total performance as well as for both forum types ($ps < 0.001$).

Spidering parameter experiment. To evaluate the effectiveness of different settings for key spidering parameters, we conducted a simulated experiment in which 40 Dark Web forums were spidered several times using different parameter settings. The parameters of interest included the number of spiders per forum, the number of proxies per spider, the number of URLs per batch, and the timeout interval between batches. The less aggressive settings were run earlier (e.g., using fewer spiders, longer timeout intervals, etc.) to decrease the likelihood of forum administrators blocking the latter spidering runs. Figure 7 shows the average percentage recall for different combinations of number of spiders and proxies per spider applied to the 40 testbed forums. Each condition was run using a constant batch size (300 URLs) and timeout interval between batches (20 s). The figure can be read as follows: When using 30 spiders per forum and one proxy per spider (i.e., 30 proxies total), the recall was slightly higher than 50%. In contrast, when using 30 spiders per forum and five proxies per spider (i.e., 150 proxies total), the recall was slightly higher than 70%.

Based on the results in Figure 7, note that the use of proxies has a profound impact on collection performance. Unlike regular forums, collection of Dark Web forums has recall of less than 30% when no proxies are used because of aggressive blocking from the forum masters. Recall constantly improves as the number of proxies per spider is increased up to four, but levels off after that point with no significant improvement when using five proxies per spider. This suggests that the use

DOI: 10.1002/asi

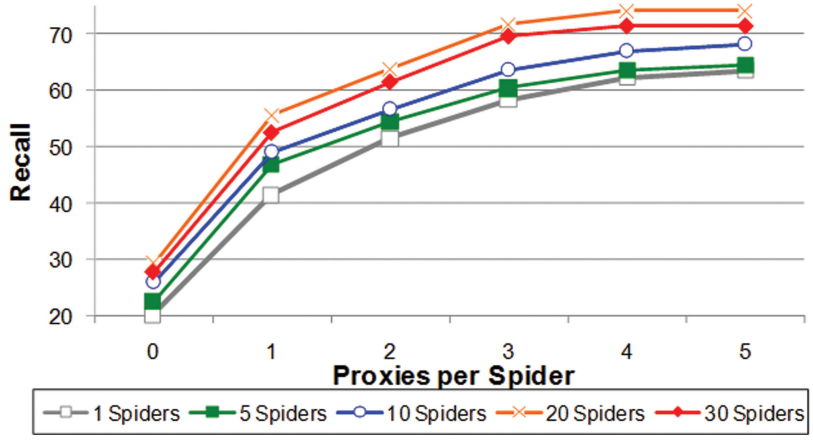

FIG. 7. Recall results for different settings of number of spiders and proxies per spider.

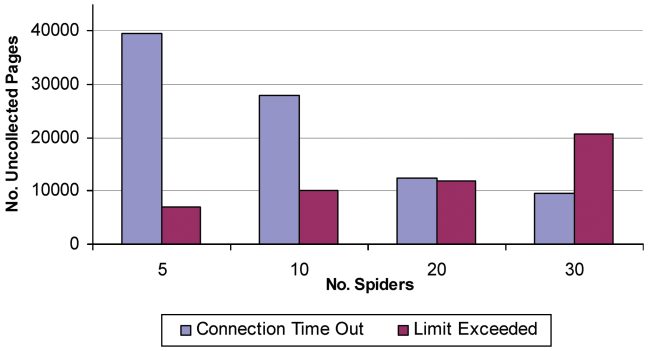

FIG. 8. Number of uncollected pages for different numbers of spiders.

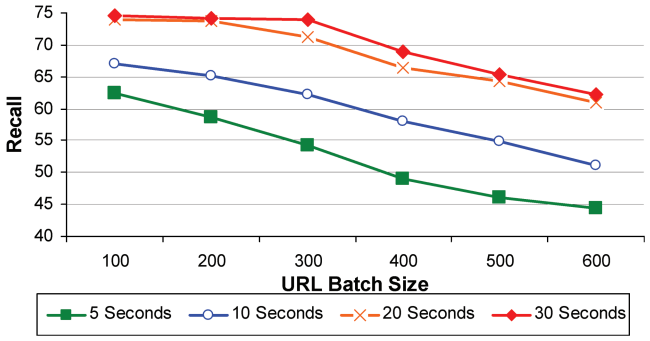

FIG. 9. Recall results for different settings of batch size and timeout interval.

of four times as many proxies as spiders per forum provides a sufficient level of anonymity.

The number of spiders per forum also impacts recall, with optimum recall attained using 20 spiders per forum. Using less spiders diminishes recall because of the extended duration required to spider the URLs in a batch, which causes the spiders to get detected. The use of more than 20 spiders (e.g., 30) decreases performance due to an excessive number of connections that can either alert the forum master and/or network administration or cause the server to overload. Thus, when selecting the number of spiders per forum, one must balance the time required to collect the pages with the amount of server load at any point in time. Using too few or too many spiders can decrease recall due to the time taken or the excessive server load, respectively. This finding was supported by an analysis of the log files when using a different number of spiders per forum. Figure 8 shows the number of uncollected pages from our 40 Web forum testbed, for different numbers of spiders when using five proxies per spider. Uncollected pages were placed into two categories based on their spider log entries. "Connection time out" pages are those that could not be collected because our spider was connected to the forum for too long. "Limit exceeded" pages are those that could not be collected because the forum blocked the spider for exceeding its download quota for a particular time period. Note that using a smaller number of spiders results in greater connection timeouts whereas the use of 30 spiders leads to increased "limit exceeded" errors. The use of 20 spiders provides the optimal balance between the two types of errors.

We also tested the impact of different batch sizes and timeout intervals (between batches) on collection recall, using the same 40 Dark Web forum testbed. For this experiment, a constant number of spiders per forum ($n = 20$) and proxies per spider ($n = 4$) were incorporated for each combination of batch and timeout interval. The results are presented in Figure 9. The diagram can be read as follows: When using a 10-s timeout interval between batches and a 200 URL batch size, recall of approximately 65% was attained.

Based on the results in Figure 9, note that both batch size and timeout interval impact recall for Dark Web forums. Not surprisingly, longer timeout intervals equate to enhanced recall; there is a 20% improvement in performance when using a timeout interval of 30 s between batches as opposed to 5-s timeout interval. Additionally, larger batch sizes also lead to deteriorating performance. When using a 30-s timeout, the drop in recall is most noticeable when increasing the batch size from 300 to 400 URLs. Although smaller batch sizes and longer timeout intervals improve recall, they also increase the spidering time. Thus, using a batch size of 300 URLs with a timeout interval of 30 s may be more favorable since it can drastically reduce spidering time with a minimal drop in recall, as compared to using a batch size of

DOI: 10.1002/asi

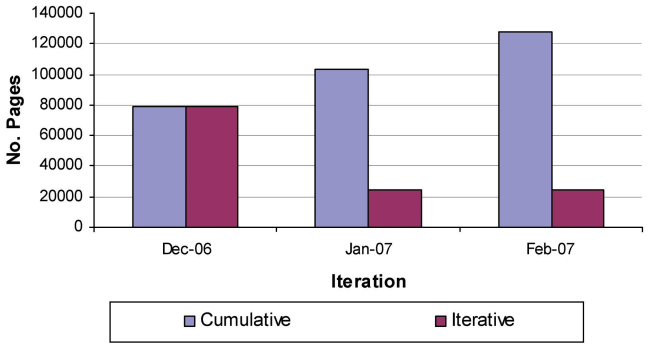

FIG. 10. Number of Web pages in testbed across 3 months/iterations.

100 or 200 URLs. The parameter-testing experiments have important implications for the spidering of Dark Web forums. Based on the results, it appears that tuning of various spidering parameters, including the number of spiders, number of proxies per spider, batch size, and timeout interval, play an integral role in recall performance.

Forum collection-update experiment. To evaluate the effectiveness of the proposed incremental crawling with recall-improvement approach (referred to as incremental + RI) for collection updating, we conducted a simulated experiment in which 40 Dark Web forums were spidered three times over a 3-month period between December 2007 and February 2007. Figure 10 shows the number of cumulative Web pages and the amount of new pages appearing in the 40 testbed forums across the 3-month period. There were approximately 128,000 unique Web pages in the testbed, which were used as the gold standard for precision, recall, and F-measure computation. We collected the pages on a monthly basis (a total of three iterations) using periodic, incremental, and incremental + RI collection-update procedures. The periodic crawler collected all pages in each iteration (the cumulative amounts in Figure 10) while the incremental crawler only collected the new pages for each iteration (the iterative amounts in Figure 10). The advantage of periodic crawling is the ability to ascertain multiple versions of a page, which can improve the likelihood of gathering pages uncollected in the previous round at the expense of collection time and server congestion. The incremental + RI procedure also collected the new pages, but used a recall mechanism that allowed improperly retrieved pages to be refetched n number of times. The recall-improvement phase, which identifies uncollected pages based on their spidering status and file size, is intended to retrieve uncollected pages in an efficient manner (i.e., without putting excessive burden on the forum servers). Consequently, a value of $n = 2$ was utilized since we have found that excessive attempts (i.e., larger values of n) typically decrease performance due to server congestion. For all experimental conditions, we used 20 spiders per forum, four proxies per spider, a batch size of 300 URLs per forum, and a timeout interval of 30 s.

Performance was evaluated using the precision, recall, and F-measures. Precision was defined as the percentage of pages

TABLE 3. Macrolevel results for different update procedures.

Update Procedure	Precision	Recall	F-measure	Time (min)
Periodic	74.32	69.03	71.58	6,101
Incremental	57.80	53.69	55.67	4,855
Incremental + RI	79.59	74.74	77.09	5,758

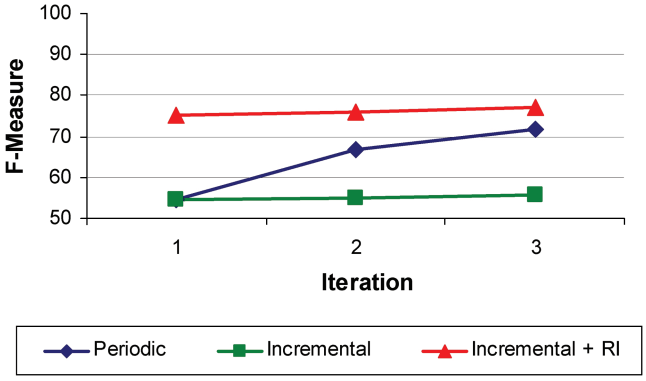

FIG. 11. Results by iteration for various collection-update procedures.

downloaded that were correctly collected. Correctly collected pages included all relevant pages completely downloaded. Incorrect pages were those that were partial/incomplete or irrelevant. Recall was defined as the percentage of relevant pages collected.

Table 3 shows the experimental results for the three collection procedures. The incremental + RI method achieved the highest precision, recall, and F-measure in a more efficient manner than did the periodic approach. The incremental update without recall improvement was the most efficient timewise; however, it only had an F-measure of roughly 55%. The results suggest that Dark Web forums require the use of a spidering strategy that entails multiple attempts to fetch uncollected pages.

Figure 11 shows the overall F-measure for the three collection-updating procedures after each spidering iteration. The diagram exemplifies the impact of making multiple attempts to collect unfetched pages. Note that the overall performance of periodic crawling improves dramatically during the second and third iterations since many of the previously uncollected Web pages are gathered. Since the incremental + IR method immediately retrieves such pages, it maintains a consistently higher level of performance, as compared to the other two methods.

Forum collection statistics. We used our spidering system for collection of Dark Web Forums in three regions. The spider was run incrementally for a 20-month period between April 2005 and December 2006. The spider collected indexable, multimedia, archive (e.g., .zip, .rar), and nonstandard files (e.g., those with unknown/unrecognized file extensions).

Table 4 shows the number of forums collected per region. The collection consists of stand-alone and hosted forums. In general, the Middle Eastern groups tended to make greater

DOI: 10.1002/asi

TABLE 4. Dark Web forum collection statistics.

	Hosted forums	Stand-alone forums	Total forums
Middle Eastern	21	50	71
Latin American	6	3	9
U.S. Domestic	16	13	29
Total	43	66	109

TABLE 5. Dark Web forum collection file statistics.

	No. of files	Volume (bytes)
Indexable files	3,001,194	140,878,063,124
HTML iles	283,578	2,942,658,681
Word files	2,108	46,649,107
PDF files	16	8,168,345
Dynamic files	2,715,354	137,178,574,841
Text files	657	2,249,471,937
Excel files	1	177,152
PowerPoint files	2	528,834
XML files	26	466,706
Multimedia files	423,749	25,833,258,770
Image files	422,155	8,554,125,848
Audio files	5,479	3,664,642,638
Video files	6,115	13,614,490,284
Archive files	801	621,721,139
Nonstandard files	443,244	17,303,588,746
Total	3,868,988	185,017,574,960

use of stand-alone forums while the U.S. domestic forums were more evenly distributed between hosted and stand-alone forums.

Table 5 shows the detailed collection statistics categorized by file types. Our system was able to collect a rich assortment of indexable and multimedia files. Note the large quantities of dynamic and multimedia files. Static HTML files, which were predominant on the Internet 10 years ago, have a minimal amount of usage in the Dark Web forums. Dynamic files outnumber static HTML files by a ratio of 10:1 while multimedia files (particularly images) also are present more often. This is partially attributable to the use of various forum software packages that generate dynamic thread pages (typically .php files).

Dark Web Forum Case Study

To provide insight into the utility of our collection for content analysis of Dark Web forums, we conducted a detailed case study. Such case studies, which have been used in prior related work (e.g., Glance et al., 2005b), are useful for illustrating the value of the collection as well as the Dark Web forum crawling system used to generate the collection. Our case study involved topical and interactional analysis of eight Dark Web forums from our collection. Topic and interaction analysis have been prevalent forms of content analysis in previous computer-mediated communication research. The dataset consisted of messages from eight domestic supremacist forums. Table 6 provides the number of authors and messages for each forum in the testbed, with a total of 650 authors and approximately 10,000 message postings.

TABLE 6. Domestic supremacist forum testbed.

Forum	Authors	Messages
Angelic Adolf	28	78
Aryan Nation	54	489
CCNU	2	429
Neo-Nazi	98	632
NSM World	289	7,543
Smash Nazi	10	66
White Knights	24	751
World Knights	35	223
Total	650	10,211

Topical Analysis

Evaluation of key topics of discussion can provide insight into the groups' content as well as the interrelations between the various forums. The vector-space model ($tf \times idf$) was used to determine the word vectors for each author. The word vectors consisted of BOW after stop/function words were removed. We then constructed an $n \times n$ matrix of similarity scores computed using the cosine measure across all 650 authors. The similarity matrix was visualized using a spring-embedding algorithm, which belongs to the family of force-directed placement algorithms. Such algorithms are common multidimensional scaling techniques in which the distance between objects is proportional to their similarity (with closer objects being more similar). Spring-embedding algorithms are a popular technique in information retrieval for viewing similarities between documents (Chalmers & Chitson, 1992; Leuski & Allan, 2000). Our implementation shows authors placed based on their cosine similarity scores. Author clusters were manually annotated with descriptions of major discussion (based on term co-occurrences).

Figure 12 shows the annotated author projections based on discussion-topic similarities, as generated by the spring-embedding algorithm. Each circle denotes an author while the circle color indicates the author's forum affiliation. The gray transparent ovals indicate author clusters based on common discussion topics. Table 7 provides descriptions of each of these topic clusters.

Based on Figure 12 and Table 7, it appears that the NSM World, Neo-Nazi, and Angelic Adolf forums all have ties with the National Socialist Movement (NSM) party. Members of these groups are avidly discussing issues relating to the party. The NSM World forum is the largest in size (in terms of members and postings), but also has the most diversity in terms of topics. This forum is the leading news source, with the most content relating to domestic and international stories and events relevant to its members. Most of the smaller forums (e.g., White Knights, World Knights, and

DOI: 10.1002/asi

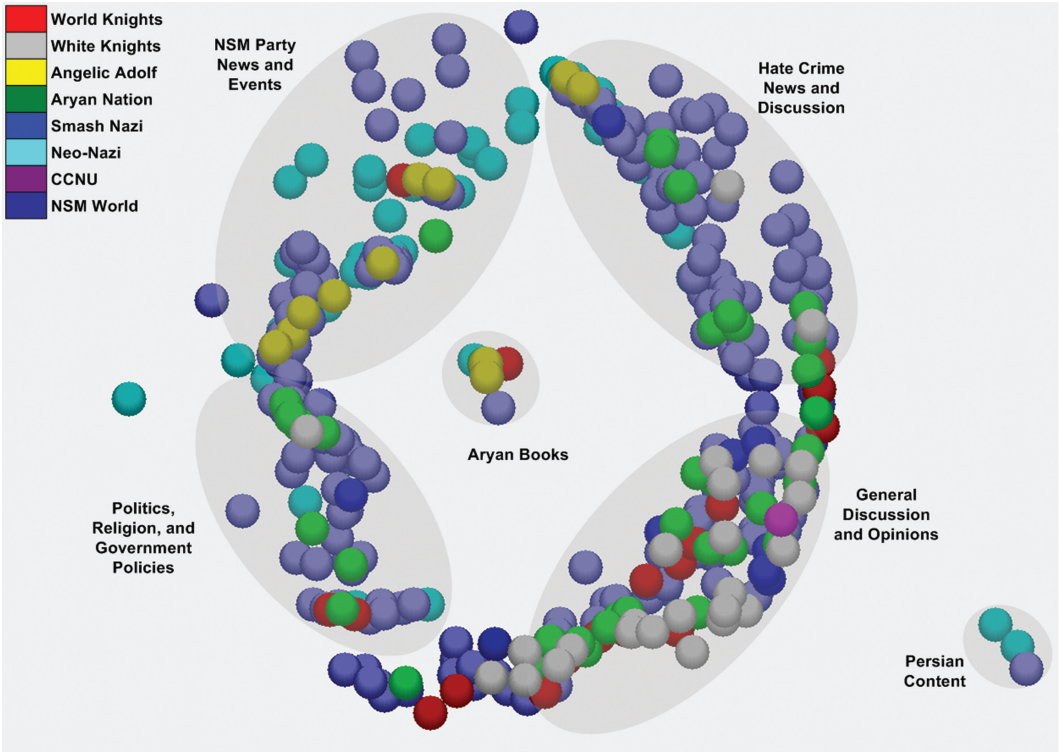

FIG. 12. Topical MDS projections for domestic supremacist forum authors.

TABLE 7. Description of major discussion topics in testbed forums.

Topic	Description
NSM Party News	News about National Socialist Movement party meetings, rallies, anniversary celebrations, and internal party politics
Hate Crime News	News about violent interracial domestic crimes involving White victims
Politics and Religion	Discussion about religious beliefs, foreign and domestic policies, and political malcontent
General Discussion and Opinions	Opinions and beliefs about different races and religions
Aryan Books	Information about the availability of literature pertaining to Aryan beliefs (including books and newsletters).
Persian Content	Content written in Farsi. There is a considerable Persian following in the Nazi groups (though the vast majority contribute in English).

Smash Nazi) are predominantly conversational forums where members discuss/argue their opinions and beliefs. Overall, there is considerable topical overlap across forums indicating that the authors of these various online communities are discussing similar matters.

Interaction Analysis

Evaluation of participant interaction can provide insight into the interrelations between various forums. We constructed the author-interaction network across the eight testbed forums. The interaction network shows to whom each individual's messages are directed, and additional forum members who are referenced in the message text. The interaction network was constructed using an interactional coherence algorithm, which analyzes message threads and outputs an interaction network (Fu, Abbasi, & Chen, 2008). Figure 13 shows the author-interaction network for the 650 authors in our testbed. Each circle (i.e., network node) denotes an author while the circle color indicates the author's forum affiliation. The lines (i.e., links) between author nodes indicate interaction between those two authors. As mentioned earlier, interaction can be in the form of direct communication between the two authors (i.e., one replying to the other's message) or via an indirect reference to the other author's screen name. A spring-layout algorithm was used to cluster authors based on link/interaction strength.

The network provides evidence of considerable interaction between members across the various forums. Cross-forum interaction occurs when a message in one forum directly addresses a member of another forum. The only forums that do not have any such cross-forum interaction are CCNU and Smash Nazi. Coincidentally, these also are the two smallest forums in our testbed, with 2 and 10 members, respectively. In contrast, members of the NSM, Neo-Nazi, and Angelic Adolf forums have considerable interaction. This is consistent with the topical analysis presented in the previous section, which also found discussion-topic similarities between members of these forums. These results also are consistent with previous Dark Web site-analysis studies that found considerable linkage between various U.S. domestic supremacist Web sites

DOI: 10.1002/asi

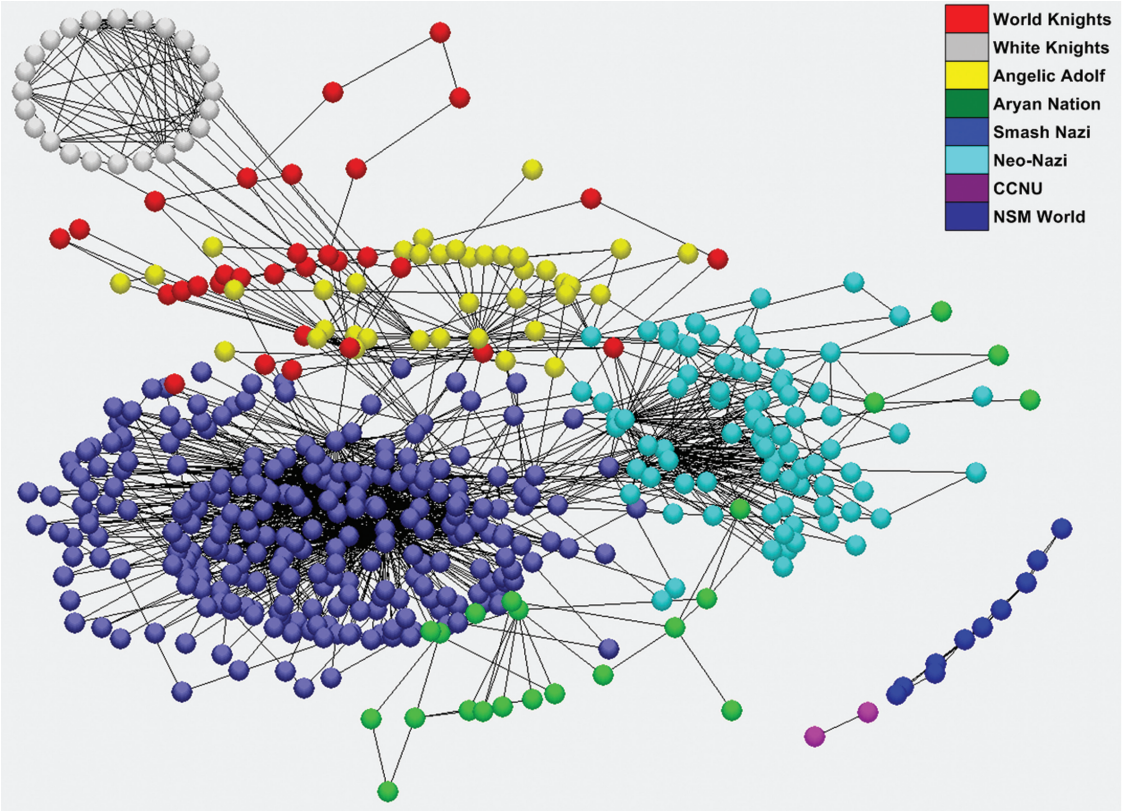

FIG. 13. Author-interaction network for domestic supremacist forums.

(Zhou et al., 2005). The case study illustrates the utility of the Dark Web forum collection for content analysis of these online communities. Synchronous efforts to collect and analyze such Web forum content are an important, yet sparsely explored, endeavor (Burris et al., 2000).

Conclusions and Future Directions

In this study, we developed a focused crawler for collecting Dark Web forums. We used a human-assisted accessibility mechanism to access identified forums with a success rate of over 90%. Our crawler uses language-independent features, including URL tokens, anchor text, and level features, to allow effective collection of content in multiple languages. It also uses forum software-specific traversal strategies and wrappers to support incremental crawling. The system uses an incremental crawling approach coupled with a recall-improvement mechanism that continually re-spiders uncollected pages. Such an update approach outperformed the use of a standard incremental-update strategy as well as the traditional periodic-update method in a head-to-head comparison in terms of precision, recall, and computation time.

The system has been able to maintain up-to-date collections of 109 forums in multiple languages from three regions: U.S. domestic supremacist, Middle Eastern extremist, and Latin groups. We also presented a case study using the collection to demonstrate its utility for content analysis. The case study provided insight into important discussion topics and interaction patterns for selected U.S. domestic supremacist forums. We believe that the proposed forum crawling system allows important entry to Dark Web forums, which facilitates better accessibility for the analysis of these online communities. The collection of such content has significant academic and scientific value for intelligence and security informatics as well as various other research communities interested in analyzing the social characteristics of Dark Web forums.

We have identified several important directions for future research. We plan to improve the Dark Web forum accessibility mechanism to attain higher access rates. We also plan to expand our collection efforts to also include Weblogs and chatting log archives. Additionally, we intend to evaluate the effectiveness of multimedia-categorization techniques to enhance our ability to collect relevant image and video content.

Acknowledgments

This research has been supported in part by the following grants: NSF Digital Government "COPLINK Center: Social Network Analysis and Identity Deception Detection for Law Enforcement and Homeland Security," October 2004–September 2007; NSF/CIA, Knowledge Discovery and Dissemination (KDD) Program "Detecting Identity Concealment," September 2005–August 2005; and Library of

Congress "Capture of Open Source Web Based Multimedia Multilingual Terrorist Content," February 2007–February 2008.

References

Abbasi, A., & Chen, H. (2005). Identification and comparison of extremist-group Web forum messages using authorship analysis. IEEE Intelligent Systems, 20(5), 67–75.

Aggarwal, C.C., Al-Garawi, F., & Yu, P.S. (2001). Intelligent crawling on the World Wide Web with arbitrary predicates. In Proceedings of the Tenth World Wide Web Conference (pp. 96–105). New York: ACM Press.

Baeza-Yates, R. (2003). Information retrieval in the Web: Beyond current search engines. International Journal of Approximate Reasoning, 34, 97–104.

Barbosa, L., & Freire, J. (2004). Siphoning hidden-Web data through keyword-based interfaces. In Proceedings of the 19th Brazilian Symposium on Databases (pp. 309–321). Brasília, Brazil:SBBD.

Bergman, M.K. (2000). The deep Web: Surfacing hidden value. Retrieved March 3, 2010, from http://quod.lib.umich.edu/cgi/t/text/text-idx?c=jep;view=text;rgn=main;idno=3336451.0007.104

Burris, V., Smith, E., & Strahm, A. (2000). White supremacist networks on the Internet. Sociological Focus, 33(2), 215–235.

Chakrabarti, S., Punera, K., & Subramanyam, M. (2002). Accelerated focused crawling through online relevance feedback. In Proceedings of the 11th International World Wide Web Conference (pp. 148–159). New York: ACM Press.

Chakrabarti, S., Van Den Berg, M., & Dom, B. (1999). Focused crawling: A new approach to topic-specific resource discovery. In Proceedings of the Eighth World Wide Web Conference (pp. 1623–1640). New York: ACM Press.

Chalmers, M., & Chitson, P. (1992). Bead: Explorations in information visualization. In Proceedings of the 15th Annual International ACM/SIGIR Conference (pp. 330–337). New York: ACM Press.

Chau, M., & Chen, H. (2003). Comparison of three vertical search spiders. IEEE Computer, 36(5), 56–62.

Chen, H. (2006). Intelligence and security informatics for international security: Information sharing and data mining. London: Springer Press.

Chen, H., & Chau, M. (2003). Web mining: Machine learning for Web applications. Annual Review of Information Science and Technology, 37, 289–329.

Chen, H., Chung, Y., Ramsey, M., & Yang, C. (1998a). A smart itsy bitsy spider for the Web. Journal of the American Society for Information Science, 49(7), 604–619.

Chen, H., Chung, Y., Ramsey, M., & Yang, C. (1998b). An intelligent personal spider (agent) for dynamic internet/intranet searching. Decision Support Systems, 23(1), 41–58.

Cheong, F.C. (1996). Internet agents: Spiders, wanderers, brokers, and bots. Indianapolis, IN: New Riders.

Cho, J., & Garcia-Molina, H. (2000). The evolution of the Web and implications for an incremental crawler. In Proceedings of the 26th International Conference on Very Large Databases (pp. 200–209). New York: ACM Press.

Cho, J., Garcia-Molina, H., & Page, L. (1998). Efficient crawling through URL ordering. Computer Networks and ISDN Systems, 30(1–7), 161–172.

Crilley, K. (2001). Information warfare: New battle fields terrorists, propaganda, and the Internet. In Proceedings of the Association for Information Management, 53(7), 250–264.

De Bra, P.M.E. & Post, R.D.J. (1994). Information retrieval in the World-Wide Web: Making client-based searching feasible. In Proceedings of the First World-Wide Web Conference (pp. 183–192). New York: ACM Press.

Diligenti, M., Coetzee, F.M., Lawrence, S., Giles, C.L., & Gori, M. (2000). Focused crawling using context graphs. In Proceedings of the 26th Conference on Very Large Databases (pp. 527–534). New York: ACM Press.

Ester, M., Grob, M., & Kriegel, H. (2001). Focused Web crawling: A generic framework for specifying the user interest and for adaptive crawling strategies. Retrieved March 3, 2010, from http://www.dbs.informatik.uni-muenchen.de/~ester/papers/VLDB2001.Submitted.pdf

Florescu, D., Levy, A.Y., & Mendelzon, A.O. (1998). Database techniques for the World-Wide Web: A Survey. SIGMOD Record, 27(3), 59–74.

Fu, T., Abbasi, A., & Chen, H. (2008). A hybrid approach to Web forum interactional coherence analysis. Journal of the American Society for Information Science and Technology, 59(8), 1195–1209.

Glance, N., Hurst, M., Nigam, K., Siegler, M., Stockton, R., & Tomokiyo, T. (2005a). Analyzing online discussion for marketing intelligence. In Proceedings of the 14th International World Wide Web Conference (pp. 1172–1173). New York: ACM Press.

Glance, N., Hurst, M., Nigam, K., Siegler, M., Stockton, R., & Tomokiyo, T. (2005b). Deriving market intelligence from online discussion. In Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (pp. 419–428), Chicago.

Glance, N., Hurst, M., & Tomokiyo, T. (2004, May). BlogPulse: Automated trend discovery for weblogs. Paper presented at the 13th International World Wide Web Conference Workshop on Weblogging Ecosystem: Aggregation, Analysis, and Dynamics, New York, NY. Retrieved March 3, 2010, from http://www.blogpulse.com/papers/www2004glance.pdf

Glaser, J., Dixit, J., & Green, D.P. (2002). Studying hate crime with the Internet: What makes racists advocate racial violence? Journal of Social Issues, 58(1), 177–193.

Guo, Y., Li, K., Zhang, K., & Zhang, G. (2006). Board forum crawling: A Web crawling method for Web forum. In Proceedings of the Conference on Web Intelligence (pp. 745–748). Washington, DC: IEEE.

Gustavson, A.T., & Sherkat, D.E. (2004, August). Elucidating the Web of hate: The ideological structuring of network ties among White supremacist groups on the Internet. Paper presented at Annual Meeting of American Sociological Association, San Francisco, CA.

Heydon, A., & Najork, M. (1999). Mercator: A scalable, extensible Web crawler. In Proceedings of the International Conference on the World Wide Web (pp. 219–229). New York: ACM Press.

Lage, J.P., Da Silva, A.S., Golgher, P.B., & Laender, A.H.F. (2002). Collecting hidden Web pages for data extraction. In Proceedings of the Fourth International Workshop on Web Information and Data Management (pp. 69–75). New York: ACM Press.

Lawrence, S., & Giles, C.L. (1999). Searching the World Wide Web. Science, 280(5360), 98.

Leuski, A., & Allan, J. (2000). Lighthouse: Showing the way to relevant information. In Proceedings of the IEEE Symposium on Information Visualization (pp. 125–130). Washington, DC: IEEE.

Li, Y., Meng, X., Wang, L., & Li, Q. (2006). RecipeCrawler: Collecting recipe data from WWW incrementally. In Proceedings of the 7th International Conference on Web-Age Information Management (pp. 263–274). Washington, DC: IEEE.

Limanto, H.Y., Giang, N.N., Trung, V.T., Huy, N.Q., & He, J.Z.Q. (2005). An information extraction engine for Web discussion forums. In Special Interest Tracks and Posters of the 14th International Conference on the World Wide Web (pp. 978–979). New York: ACM Press.

Lin, K., & Chen, H. (2002). Automatic information discovery from the "Invisible Web." In Proceedings of the International Conference on Information Technology: Coding and Computing (p. 332). Washington, DC: IEEE.

Menczer, F. (2004). Lexical and semantic clustering by Web links. Journal of the American Society for Information Science and Technology, 55(14), 1261–1269.

Menczer, F., Pant, G., & Srinivasan, P. (2004). Topical Web crawlers: Evaluating adaptive algorithms. ACM Transactions on Internet Technology, 4(4), 378–419.

Najork, M., & Wiener, J.L. (2001). Breadth-first search crawling yields high-quality pages. In Proceedings of the World Wide Web Conference (pp. 114–118). New York: ACM Press.

Ntoulas, A., Zerfos, P., & Cho, J. (2005). In Proceedings of the Fifth ACM/IEEE-CS Joint Conference on Digital Libraries (pp. 100–109). New York: ACM Press.

DOI: 10.1002/asi

Pant, G., & Srinivasan, P. (2005). Learning to crawl: Comparing classification schemes. ACM Transations on Information Systems, 23(4), 430–462.

Pant, G., & Srinivasan, P. (2006). Link contexts in classifier-guided topical crawlers. IEEE Transactions on Knowledge and Data Engineering, 18(1), 107–122.

Pant, G., Srinivasan, P., & Menczer, F. (2002, May). Exploration versus exploitation in topic driven crawlers. Paper presented at the Second World Wide Web Workshop on Web Dynamics, Honolulu, Hawaii. Retrieved March 2, 2010, from http://www.dcs.bbk.ac.uk/webDyn2/proceedings/pant_topic_driven_crawlers.pdf

Raghavan, S., & Garcia-Molina, H. (2001). Crawling the hidden Web. In Proceedings of the 27th International Conference on Very Large Databases (pp. 129–138). New York: ACM Press.

Schafer, J. (2002). Spinning the Web of hate: Web-based hate propagation by extremist organizations. Journal of Criminal Justice and Popular Culture, 9(2), 69–88.

Sizov, S., Graupmann, J., & Theobald, M. (2003). From focused crawling to expert information: An application framework for Web exploration and portal generation. In Proceedings of the 29th International Conference on Very Large Databases (pp. 1105–1108). New York: ACM Press.

Smith, M. (2002). Tools for navigating large social cyberspaces. Communications of the ACM, 45(4), 51–55.

Srinivasan, P., Mitchell, J., Bodenreider, O., Pant, G., & Menczer, F. (2002, July). Web crawling agents for retrieving biomedical information. Paper presented at International Workshop on Agents in Bioinformatics (NETTAB), Bologna, Italy. Retrieved March 3, 2010, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.16.8948&rep=rep1&type=pdf

Whine, M. (1997). The governance of cyberspace: Politics, technology, and global restructuring. London: Routledge.

Yih, W., Chang, P., & Kim, W. (2004). Mining online deal forums for hot deals. In Proceedings of the Web Intelligence Conference (pp. 384–390). Washington, DC: IEEE.

Zhou, Y., Reid, E., Qin, J., Chen, H., & Lai, G. (2005). U.S. extremist groups on the Web: Link and content analysis. IEEE Intelligent Systems, 20(5), 44–51.

DOI: 10.1002/asi