

A Focused Crawler for Dark Web Forums

Tianjun Fu, Ahmed Abbasi, and Hsinchun Chen

Artificial Intelligence Lab, Department of Management Information Systems

The University of Arizona, Tucson, Arizona 85721, USA

Phone: 520-621-2748, Fax: 520-621-2433

{futj@email.arizona.edu, abbasi@uwm.edu, hchen@eller.arizona.edu}

Abstract

The unprecedented growth of the Internet has propagated the escalation of 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 technique 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 a hundred 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.

1. Introduction

The Internet acts as an ideal method for information and propaganda dissemination (Whine, 1997; Gustavson et al., 2004). 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 et al., 2002). This problematic facet of the Internet is often referred to as the Dark Web (Chen, 2006). An important component of the Dark Web is extremist forums hidden deep within the Internet. Many have stated the need for collection and analysis of Dark Web forums (Burriss et al., 2000; Schafer, 2002). Dark Web materials have important implications for intelligence and security informatics related

application (Chen, 2006). The collection of such content is also 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 domain specific collection building. The latter are referred to as focused or topic driven crawlers (Chakrabarti et al., 1999; Pant et al., 2002).

There is a need for a focused crawler that can collect Dark Web forums. Many previous focused crawlers have focused 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 are dynamic and often require memberships. They are part of the Hidden Web (Florescu et al., 1998; Raghavan & Garcia-Molina, 2001) which is not easily accessible through normal web navigation or standard crawling. There are also *multilingual* web mining considerations. More than 30% of the web is in non-English languages (Chen & Chau, 2003). Consequently, the Dark Web also encompasses numerous languages. Another important concern is *content richness*. Dark web forums contain rich content used for routine communication and propaganda dissemination (Abbasi & Chen; 2005; Zhou et al., 2005; Qin et al., 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 (text based) files.

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. 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 in order to improve accessibility. The crawler uses language-independent features for crawl space URL ordering in order to negate any complications attributable to the presence of numerous languages. We also incorporate iterative collection of incomplete downloads and relevance feedback for improved multimedia collection.

The remainder of the paper is organized as follows. Section 2 presents a review of related work on focused and hidden web crawling. Section 3 describes research gaps and our related research questions. Section 4 describes a research design geared towards addressing those questions. Section 5 presents a detailed description of our Dark Web forum spidering system. Section 6 describes 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. This section also highlights the Dark Web forum collection statistics for data gathered using the proposed system. Section 7 presents a case study conducted to illustrate the value of the collected dark web forums for content analysis while Section 8 contains concluding remarks.

2. 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 are also relevant to collection of Dark Web forums. Some of these characteristics are specific to focused and/or hidden web crawling while others are 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.

2.1 Accessibility

Most search engines cover what is referred to as the “publicly indexable Web” (Lawrence & Giles, 1998; Raghavan & Garcia-Molina, 2000). This is the part of the web easily accessible with traditional web crawlers (Sizov et al., 2003). As noted by Lawrence and Giles (1998), 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). A study conducted in 2000 found that the invisible web contained 400-550 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 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 et al., 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 et al., 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 (Raghvan & 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, email address, etc. can be automated with standardized responses. Other 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.

2.2 Collection Type

Previous focused crawling research has been geared towards 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 portal building. Srinivasan et al. (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 has also been work on collecting weblogs. BlogPulse (Glance et al., 2004) is a blog analysis portal. The site contains analysis of key discussion topics/trends for roughly 100,000

spidered weblogs. Such blogs can also be useful for marketing intelligence (Glance et al., 2005). 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., 2005). Furthermore, web forums require information extraction wrappers for derivation of meta data (e.g., authors, messages, timestamps, etc.). Wrappers are important for data analysis and incremental crawling (respidering 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., 2005) is a system for harvesting messages from online forums. It has two components: a crawler and a wrapper. Limanto et al. (2005) developed a web forum information extraction engine that includes a crawler, wrapper generator, and extractor (i.e., application of generated wrapper). Yih et al. (2004) created an online forum mining system composed of a crawler and information extractor for mining deal forums. There has been no prior research on collecting Dark Web forums.

2.3 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 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).

2.4 URL Ordering Features

Aggarwal et al. (2001) pointed out 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 in/back links and out links (Cho et al., 1998; Pant et al., 2002). Sibling links (Aggarwal et al., 2001) consider sibling pages

(ones with shared parent in link). Context graphs (Diligenti et al., 2000) derive back links 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 (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 are also language dependent and can be harder to apply in situations where the collection is composed of pages in numerous languages. Other studies have also used *URL/anchor text*. Word tokens found within the URL anchor have been used effectively to help control the crawl space (Cho et al., 1998; Ester et al., 2001). URL tokens have also been incorporated in previous focused crawling research (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.

2.5 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 and Post, 1994). Breadth first (BFS) (Cho et al., 1998; Ester et al., 2001; Najork and Wiener, 2001) is one of the simplest strategies. It has worked fairly well in comparison with more sophisticated best-first search strategies (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.

Best-first uses some criterion for ranking URLs in the crawl space, such as *link analysis* or *text analysis*. Numerous link analysis techniques have been used for URL ordering. Cho et al. (1998) evaluated the effectiveness of Page Rank and back link 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 while Chakrabarti et al. (1999) used a modified 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 (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 et al. (1998a; 1998b) used a genetic algorithm to order the URL crawl space for the collection of topic specific web pages based on bag-of-word representations of pages.

2.6 Collection Update Procedure

Two approaches for collection updating are periodic and incremental crawling (Cho and Garcia-Molina, 2000). *Periodic crawling* entails building of a brand new collection for updating. This is commonly done since it's often easier than figuring out which pages to refresh. Periodic crawling is inefficient from a spidering perspective (more time consuming). However multiple versions of a collection may improve overall recall. *Incremental crawling* gathers new and updated content. In the case of web sites, this often requires some form of change frequency estimation (Cho & Garcia-Molina, 2003) in order to determine which pages need to be updated. For web forums, this entails fetching only those threads that have been updated (Yih et al., 2003; Glance et al., 2005) since we only want to fetch newly posted messages. This requires the use of a wrapper that can parse out the "last updated" dates for threads and compare them against the previous collection to determine which pages need to be collected.

2.7 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. There have only been a few studies that 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 bag-of-word (BOW), link, or URL token features coupled with a best-first search strategy for crawl space URL ordering. Furthermore, most prior research also 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 paper.

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 Web	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., 2005)	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	Surface Web	Recipe Sites,	Indexable	Web Page Text	Best-First: Tree Edit

(Li et al., 2006)		Blogs, and Web Forums	Files Only		Distance Similarity Scores
-------------------	--	--------------------------	------------	--	-------------------------------

3. Research Gaps and Questions

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

3.1 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 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.

3.2 Content Richness

Most previous research has focused on indexable (text based) files. Large multimedia files large (e.g., videos) can be hundreds of MB. This can cause connection timeouts or excessive server loads, resulting in partial/incomplete downloads. Furthermore, the challenges in indexing multimedia files pose problems. It's 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 in specific (Chen, 2006), there is a need to capture multimedia files.

3.3 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.

3.4 Research Questions

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

- 1) How effectively can Dark Web forums be identified and accessed for collection purposes?
- 2) How effectively can Dark Web content (indexable and multimedia) be collected?
- 3) Which collection update procedure (periodic or incremental) is more suitable for Dark Web forums? How can recall improvement further enhance the update process?
- 4) How can analysis of extracted information from Dark Web forums improve our understanding of these online communities?

4. Research Design

4.1 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 in order to minimize the amount of irrelevant web pages collected.

We also propose the use of 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.

4.2 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 would also like to evaluate our system's ability to access Dark Web forums. This translates into measuring the percentage of attempted forums accessed.

4.3 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.

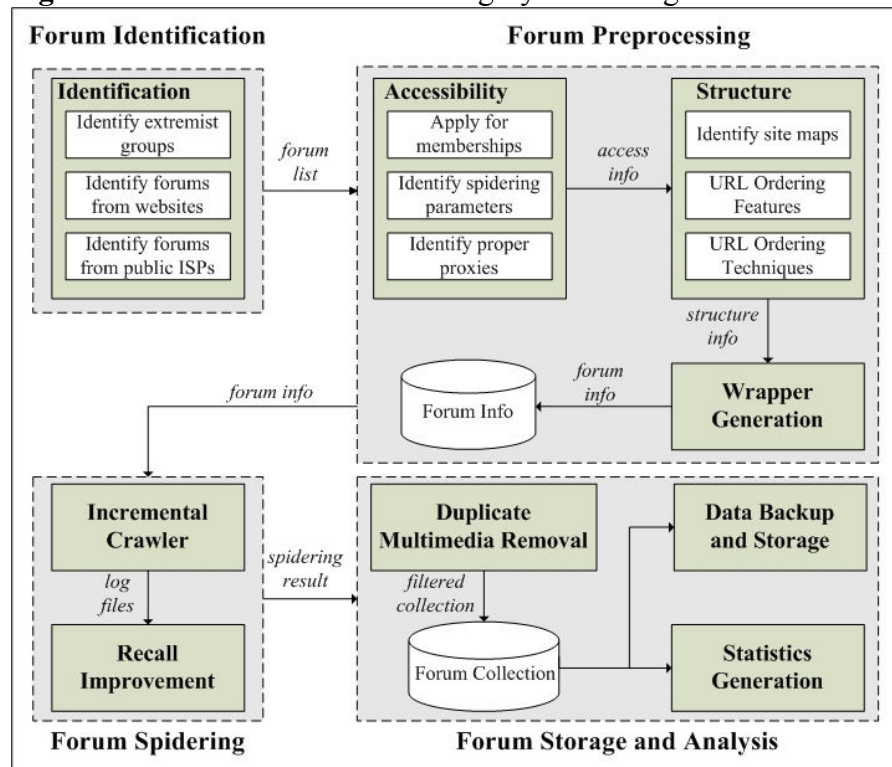
5. System Design

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

Figure 1: Dark Web Forum Crawling System Design



5.1 Forum Identification

The forum identification phase has three components.

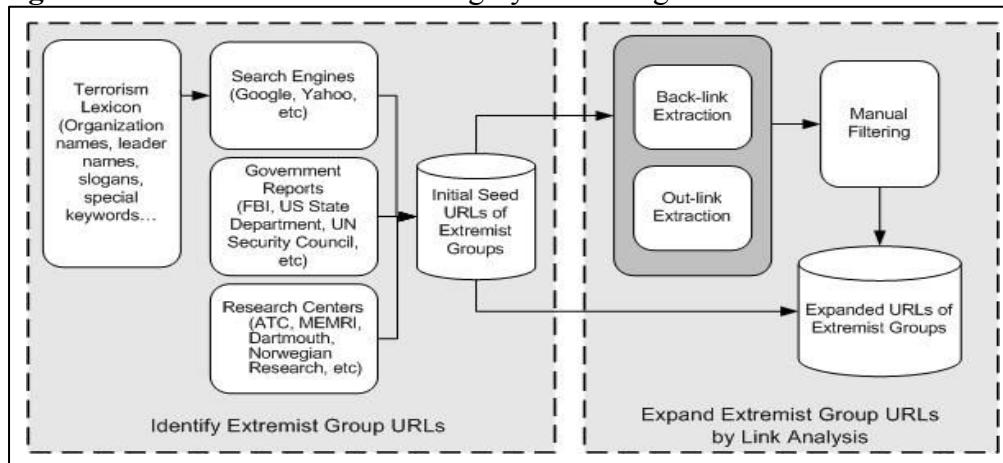
Step 1: Identify extremist groups

Sources for the US 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 United States Committee for a Free Lebanon (USCFAFL), Counter-Terrorism Committee (CTC) of the UN Security Council (UN), US 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 P. R. 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 socio-political 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.

Step 2: Identify forums from extremist websites

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: Firstly 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. Secondly 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 back-link search using Google, which has been shown to be effective in prior research (Diligenti et al., 2000), and also search out-links. The identified web forums are manually checked.

Figure 2: Dark Web Forum Crawling System Design



Step 3: 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 MSN groups, AOL Groups, etc. Public ISPs are searched with our Dark Web domain lexicon for a list of potential forums.

The above 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. In order to develop proper features and techniques for managing the crawl space.

5.2 Forum Preprocessing

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

5.2.1 Forum Accessibility

Step 1: Apply for membership

Many Dark Web forums do not allow anonymous access (Zhou et al., 2006). In order 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 emails to get access privilege. For such forums, human expertise is invaluable. Nevertheless, in some cases, access cannot be attained.

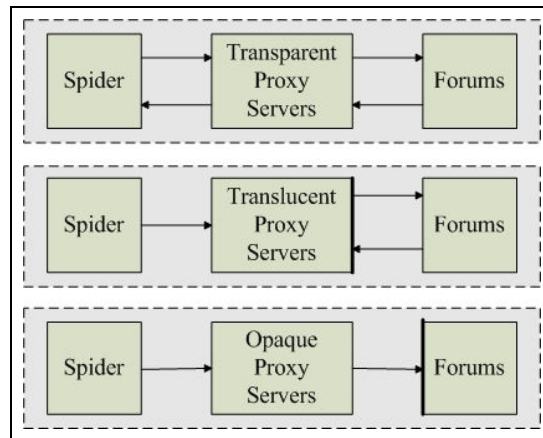
Step 2: Identify appropriate spidering parameters

Spidering parameters such as number of connections, download intervals, timeout, speed, etc., 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 download of larger files from slow servers. However we may still be blocked based on our IP address. Therefore, we use proxies to increase not only our recall but also our anonymity.

Step 3: Identify appropriate proxies

We use three types of proxy servers, as shown in Figure 3. *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 (preferred) hide your IP address and do not even 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.

Figure 3: Proxies Used for Dark Web Forum Crawling



5.2.2 Forum Structure

Step 1: Identify site maps

Forums typically have hierarchical structures with boards, threads, and messages (Yih et al., 2004; Glance et al., 2005). They also contain considerable additional information such as message posting interfaces, search, and calendar pages. We firstly identify the site map of the

forum based on the forum software packages. Glance et al. (2005) noted that although there are only a handful of commonly used forum software packages, they are highly customizable.

Step 2: URL Ordering Features

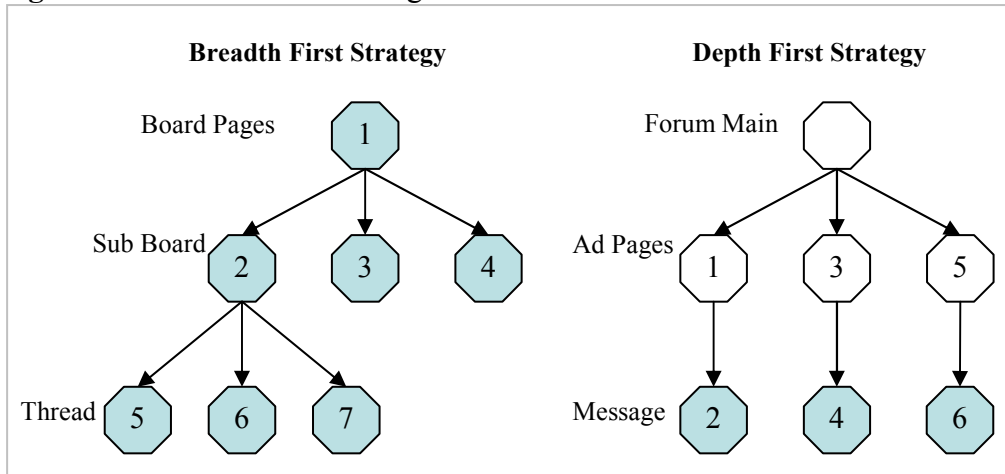
Our spidering system uses two types of language independent URL ordering features, URL tokens and page levels. With respect to *URL tokens*, for web forums, we're interested in URLs containing words such as "board," "thread," "message" etc. (Glance et al., 2005). Additional relevant URL tokens include domain names of third party file hosting web sites. These third parties often contain multimedia files. File extension tokens (e.g. ".jpg" and ".wmv") are also important. URLs that contain phrases such as "sort=voteavg" and "goto=next" are also found in relevant pages. However these are not unique to board, thread, and message pages, hence such tokens 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 are also 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. In order 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.

Step 3: 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. Breadth first (BFS) is used for board page forums while depth first (DFS) for internet service provider (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 in order to get to the message pages (typically the ad pages have a link to the actual message page). Figure 4 illustrates how the BFS and 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. DFS is necessary for ISP forums since these forums often require traversing an ad page in order to get to the message page.

Figure 4: URL Traversal Strategies



5.2.3 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., 2005). Board pages tell us when each thread was last updated with new messages. Using this information, one may respider only those thread pages containing new postings.

5.3 Forum Spidering

Figure 5 below 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 weren’t downloaded. It determines whether the URLs should be respidered.

Figure 5: Spidering Process

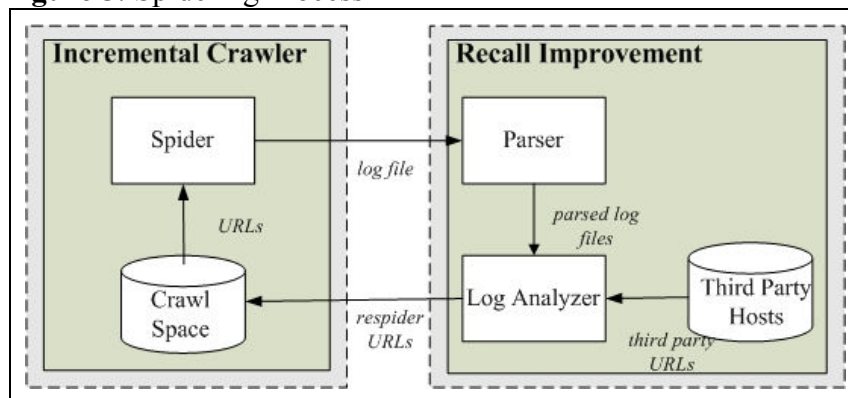
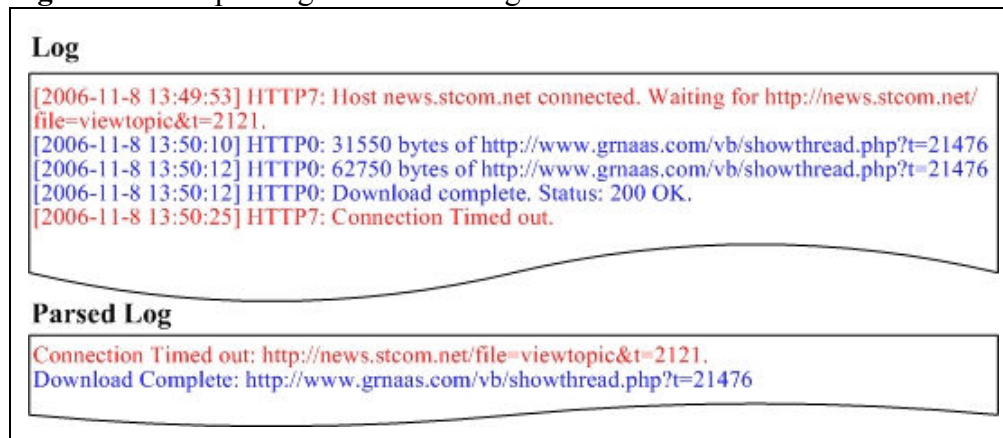


Figure 6 shows sample entries from the original and parsed log. The original log file shows the download status for each file (URL). The parsed log shows the overall status as well as the reason for download failure (in the case of undownloaded files). Blue colored entries relate to downloaded files while 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 respidering list) while “Connection Timed Out” URLs are respidered. 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 timeout.

Figure 6: Example Log and Parsed Log Entries



5.4 Forum Storage and Analysis

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

5.4.1 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.
- Non-standard files: Unrecognized file types.

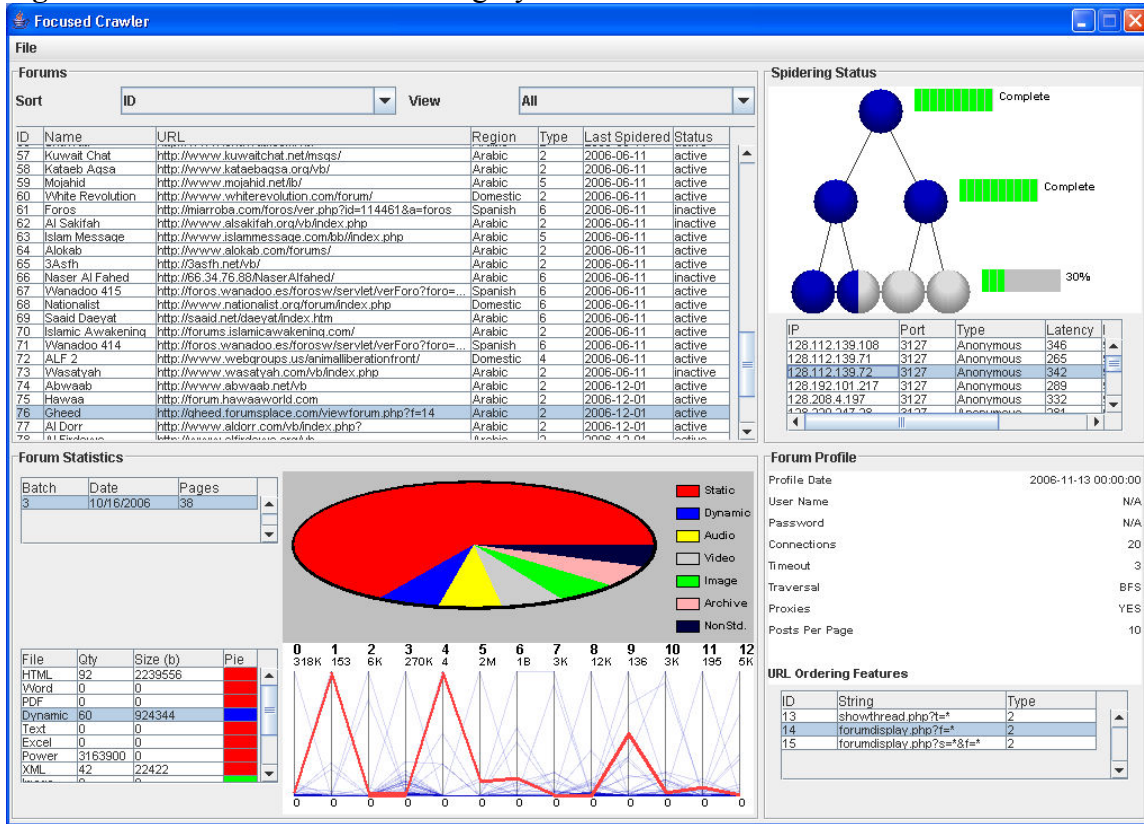
5.4.2 Duplicate Multimedia Removal

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 their meta data encoded into the file, instead of their suffixes (file extensions). It compares files based on their MD5 values, which are the same for duplicate video files collected from various Internet sources. MD5 (Message-Digest algorithm 5) is a widely-used cryptographic hash function with a 128-bit hash value. Therefore it can more accurately differentiate multimedia files with other types of files.

5.5 Dark Web Forum Crawling System Interface

Figure 7 shows the interface for the proposed Dark Web Forum spidering system. The interface has four major components. The "Forums" panel in the top left 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, sub-board, and thread pages collected for the current forum being spidered. The "Forum Statistics" panel in the bottom left shows the quantity and size of the various file type collected for each forum, using tables, pie charts, and parallel coordinates. The "Forum Profile" panel in the bottom right 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.

Figure 7: Dark Web Forum Crawling System Interface



6. Evaluation

We conducted two 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 approach in comparison with standard spidering without any accessibility mechanism.

The second 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.

6.1.1 Forum Accessibility Experiment

Table 2 below 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 internet service providers and virtually all of the attempted stand alone forums. The overall results (over 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 biggest 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 (18 forums).

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%

Table 3 shows the p-values for the pair wise t-test 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 at alpha = 0.01 for total performance as well as both forum types.

Table 3: Dark Web Forum Accessibility Statistics

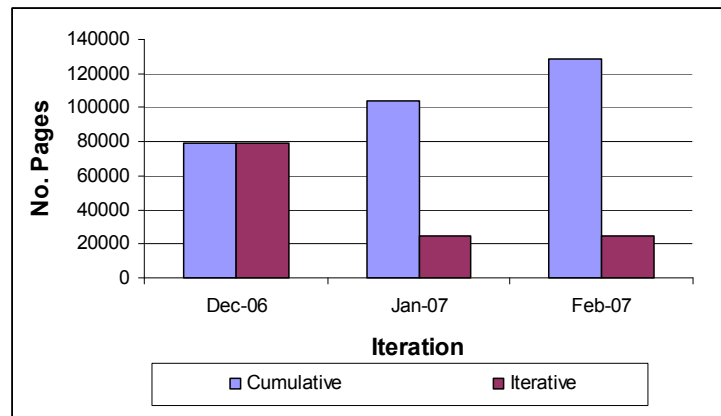
	Human Assisted Accessibility vs. Standard Spidering
Hosted Forums	< 0.001*
Stand Alone Forums	< 0.001*
Total Forums	< 0.001*

6.1.2 Forum Collection Update Experiment

In order 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 three month period between December 2007 and February 2007. Figure 8 shows the number of cumulative web pages and the amount of new pages appearing in the 40 test bed forums across the three month period. There were approximately 128,000 unique web pages in the test bed, 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 a periodic, incremental,

and incremental + RI collection update procedure. The periodic crawler collected all pages in each iteration (the cumulative amounts in Figure 8) while the incremental crawler only collected the new pages for each iteration (the iterative amounts in Figure 8). 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 also collected the new pages but used a recall mechanism that allowed improperly retrieved pages to be re-fetched 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.

Figure 8: Number of Web Pages in Test Bed across 3 Months/Iterations



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

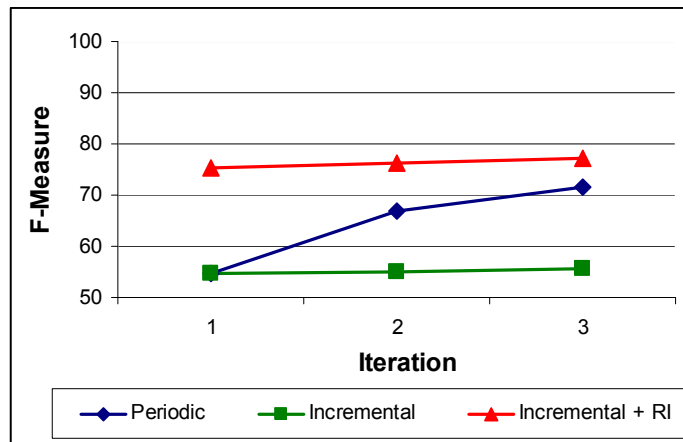
Table 4 shows the experimental results for the three collection procedures. The incremental + RI method achieved the highest precision, recall, and F-measure in a more efficient manner than the periodic approach. The incremental update without recall improvement was the most efficient time-wise however 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.

Table 4: Macro-Level 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

Figure 9 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. We can see 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 retrieves such pages immediately, it maintains a consistently higher level of performance as compared to the other two methods.

Figure 9: Results by Iteration for Various Collection Update Procedures



6.1.3 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 4/2005 and 12/2006. The spider collected indexable, multimedia, archive (e.g., .zip, .rar), and non-standard files (e.g., those with unknown/unrecognized file extensions).

Table 5 below shows the number of forums collected per region. The collection consists of stand alone and hosted forums. In general, the Middle Eastern groups tend to make greater use of stand alone forums while the U.S. domestic forums are more evenly distributed between hosted and stand alone forums.

Table 5: Dark Web Forum Collection Statistics

	Hosted Forums	Stand Alone Forums	Total Forums
Middle Eastern	21	50	71
Latin American	6	3	9
US Domestic	16	13	29
Total	43	66	109

Table 6 shows the detailed collection statistics categorized by file types. Our system was able to collect a rich assortment of indexable and multimedia files. It's interesting to note the large quantities of dynamic and multimedia files. Static HTML files, which were predominant on the Internet ten 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) are also present more often. This is partially attributable to the use of various forum software packages that generate dynamic thread pages (typically .php files).

Table 6: Dark Web Forum Collection File Statistics

	# of Files	Volume (Bytes)
Indexable Files	3,001,194	140,878,063,124
HTML Files	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
Non-Standard Files	443,244	17,303,588,746
Total	3,868,988	185,017,574,960

7. Dark Web Forum Case Study

In order to provide insight into the utility of our collection for content analysis of Dark Web forums, we conducted a detailed case study. Our case study involved topical and interactional analysis of 8 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 8 domestic supremacist forums. Table 7 provides the number

of authors and messages for each forum in the test bed, with a total of 650 authors and approximately ten thousand message postings.

Table 7: Domestic Supremacist Forum Test Bed

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

7.1 Topical Analysis

Evaluation of key topics of discussion can provide insight into the groups’ content as well as the inter-relations between the various forums. The vector-space model (tf x idf) was used to determine the word vectors for each author. The word vectors consisted of bag-of-words after stop/function words were removed. We then constructed a $n \times n$ 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 and Chitson, 1992; Leuski and 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 10 shows the annotated author MDS projections based on discussion topic similarities. 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 8 provides descriptions of each of these topic clusters.

Figure 10: Topical MDS Projections for Domestic Supremacist Forum Authors

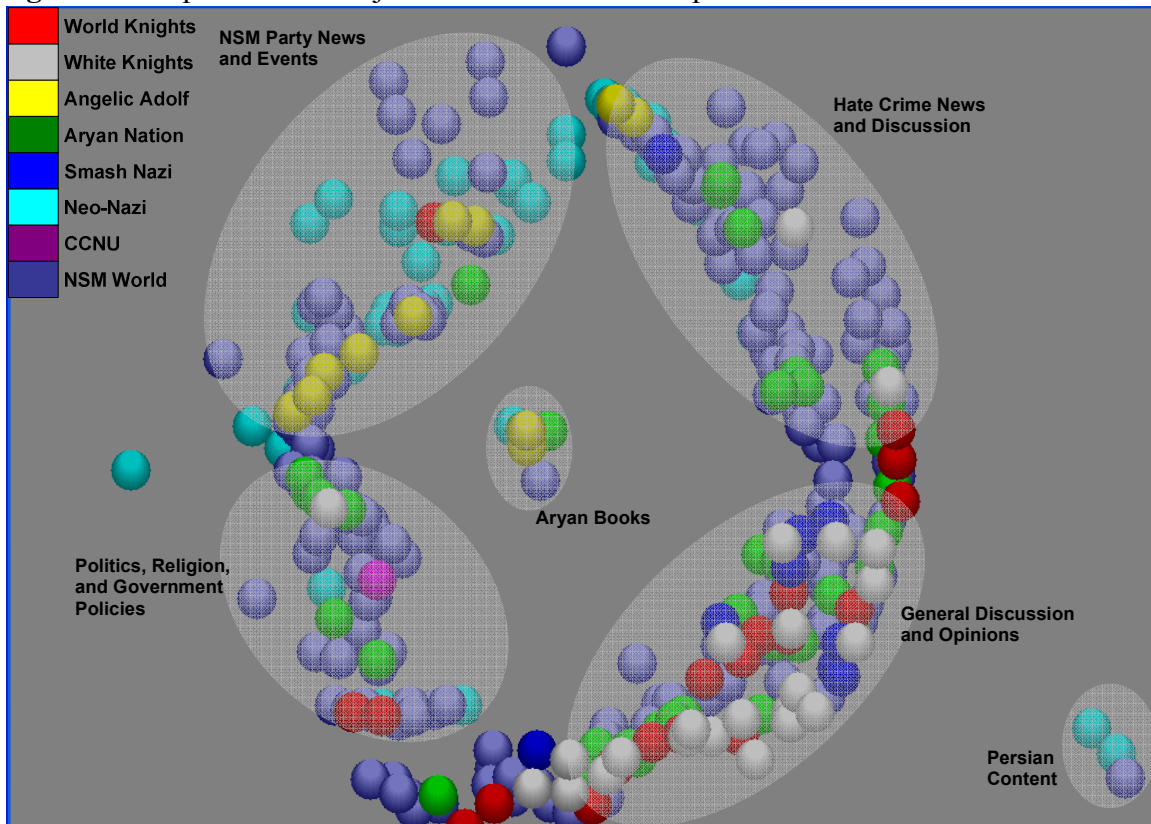


Table 8: Description of Major Discussion Topics in Test Bed Forums

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 inter-racial 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).

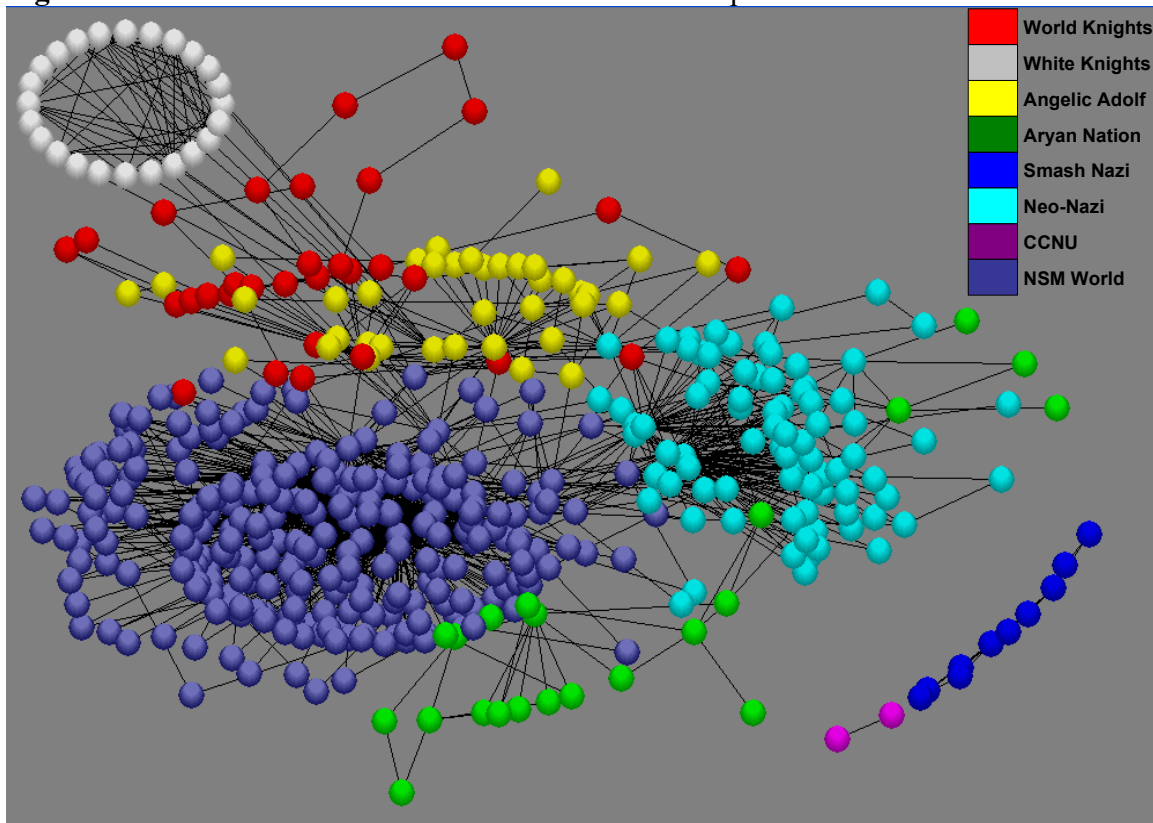
Based on Figure 10 and Table 8, it appears that the NSM World, Neo-Nazi, and Angelic Adolf forums all have ties with the National Socialist Movement (NSM) party. Members of these groups are avidly discussing issues relating to the party. The NSM World forum is the largest in size (in terms of members 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 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.

7.2 Interaction Analysis

Evaluation of participant interaction can provide insight into the interrelations between various forums. We constructed the author interaction network across the 8 test bed forums. The interaction network shows whom each individual's messages are directed towards; as well as additional forum members that are referenced in the message text. Figure 11 shows the author interaction network for the 650 authors in our test bed. Each circle (network node) denotes an author while the circle color indicates the author's forum affiliation. The lines (links) between author nodes indicate interaction between those two authors. As mentioned above, 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.

Figure 11: Author Interaction Network for Domestic Supremacist Forums



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 are also the two smallest forums in our test bed, with two and ten 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 are also consistent with previous Dark Web site analysis studies that found considerable linkage between various U.S. domestic supremacist web sites (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 (Burriss et al., 2000).

8. 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, in order 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 in order 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 the 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 in order 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.

9. Acknowledgements

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 2007

Library of Congress

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

References

- Abbasi, A. and 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., and Yu, P. S. (2001). Intelligent Crawling on the World Wide Web with Arbitrary Predicates. In *Proceedings of the 10th World Wide Web Conference*, Hong Kong.
- Baeza-Yates, R. (2003). Information Retrieval in the Web: Beyond Current Search Engines. *International Journal of Approximate Reasoning*, 34, 97-104.
- Barbosa, L. and Freire, J. (2004). Siphoning Hidden-Web Data through Keyword-Based Interfaces. In *Proceedings of the SBBD*.
- Bergman, M. K. (2000). The Deep Web: Surfacing Hidden Value. *BrightPlanet.com*,
- Burris, V., Smith, E., and Strahm, A. (2000). White Supremacist Networks on the Internet. *Sociological Focus*, 33(2), 215-235.
- Chakrabarti, S., Van Den Berg, M., and Dom, B. (1999). Focused Crawling: A New Approach to Topic-Specific Resource Discovery. In *Proceedings of the Eight World Wide Web Conference*, Toronto, Canada.
- Chalmers, M. and Chitson, P. (1992). Bead: Explorations in Information Visualization. In *Proceedings of the 15 Annual International ACM/SIGIR Conference*, 330-337.
- Chau, M. and Chen, H. (2003). Comparison of Three Vertical Search Spiders. *IEEE Computer*, 36(5), 56-62.
- Chen, H. Chung, Y., Ramsey, M., and 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., and Yang, C. (1998b). An Intelligent Personal Spider (Agent) for Dynamic Internet/Intranet Searching. *Decision Support Systems*, 23(1), 41-58.
- Chen, H. and Chau, M. (2003). Web Mining: Machine Learning for Web Applications. *Annual Review of Information Science and Technology*, (37), 289-329.
- Chen, H. (2006). *Intelligence and Security Informatics for International Security: Information Sharing and Data Mining*, London, Springer Press, 2006.
- Cheong, F. C. (1996). *Internet Agents: Spiders, Wanderers, Brokers, and Bots*. Indianapolis, IN: New Riders Publishing.

- Cho, J., Garcia-Molina, H., and Page, L. (1998). Efficient Crawling Through URL Ordering. In Proceedings of the 7th World Wide Web Conference, Brisbane, Australia.
- Cho, J and 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.
- Cho, J. and Garcia-Molina, H. (2003). Estimating Frequency of Change. *ACM Transactions on Internet Technology*, 3(3), 256-290.
- 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.
- Diligenti, M., Coetzee, F. M., Lawrence, S., Giles, C. L., and Gori, M. (2000). Focused Crawling Using Context Graphs. In Proceedings of the 26th Conference on Very Large Databases, Cairo, Egypt.
- Ester, M., Grob, M., and Kriegel, H. (2001). Focused Web Crawling: A Generic Framework for Specifying the User Interest and for Adaptive Crawling Strategies. In Proceedings of the International Conference on Very Large Databases.
- Florescu, D., Levy, A. Y., and Mendelzon, A. O. (1998). Database Techniques for the World-Wide Web: A Survey. *SIGMOD Record*, 27(3), 59-74.
- Glance, N., Hurst, M., and Tomokiyo, T. (2004). BlogPulse: Automated Trend Discovery for Weblogs. In Proceedings of the 13th International World Wide Web Conference, New York, New York.
- Glance, N., Hurst, M., Nigam, K., Siegler, M., Stockton, R. and Tomokiyo, T. (2005). Analyzing Online Discussion for Marketing Intelligence, In Proceedings of the 14th International World Wide Web Conference, Chicago, Illinois.
- Glaser, J., Dixit, J., and Green, D. P. (2002). Studying Hate Crime with the Internet: What Makes Racists Advocate Racial Violence? *Journal of Social Issues*, 58(1), 177-193.
- Gustavson, A.T. and Sherkat, D.E. (2004). 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.
- Heydon, A. and Najork, M. (1999). Mercator: A Scalable, Extensible Web Crawler. In Proceedings of the International Conference on the World Wide Web, 219-229.

- Lage, J. P., Da Silva, A. S., Golgher, P. B., and Laender, A. H. F. (2002). Collecting Hidden Web Pages for Data Extraction. In Proceedings of WIDM.
- Lawrence, S. and Giles, C. L. (1999). Searching the World Wide Web. *Science*, 280(5360), 98.
- Leuski, A. and Allan, J. (2000). Lighthouse: Showing the way to Relevant Information. In Proceedings of the IEEE Symposium on Information Visualization, 125-130.
- Limanto, H. Y., Giang, N. N., Trung, V. T., Huy, N. Q., and He, J. Z. Q. (2005). An Information Extraction Engine for Web Discussion Forums. In Proceedings of the 14th International Conference on the World Wide Web, Chiba, Japan.
- Lin, K. and Chen, H. (2002). Automatic Information Discovery from the “Invisible Web.” In Proceedings of the International Conference on Information Technology: Coding and Computing.
- Najork, M. and Wiener, J. L. (2001). Breadth-First Search Crawling Yields High-Quality Pages. In Proceedings of the World Wide Web Conference, Hong Kong.
- Ntoulas, A., Zerkos, P., and Cho, J. (2005). In Proceedings of the Joint Conference on Digital Libraries, Denver, Colorado.
- Pant, G., Srinivasan, P., and Menczer, F. (2002). Exploration versus Exploitation in Topic Driven Crawlers. In Proceedings of the WWW Workshop on Web Dynamics.
- Raghavan, S. and Garcia-Molina, H. (2001). Crawling the Hidden Web. In Proceedings of the 27th International Conference on Very Large Databases.
- 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., and 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, Berlin, Germany.
- Srinivasan, P., Mitchell, J., Bodenreider, O., Pant, G., and Menczer, F. (2002). Web Crawling Agents for Retrieving Biomedical Information. In Proceedings of the International Workshop on Agents in Bioinformatics (NETTAB), Bologna, Italy.
- Whine, M. (1997). *The Governance of Cyberspace: Politics, Technology, and Global Restructuring*. Routledge, London, U.K.
- Yih, W., Chang, P., and Kim, W. (2004). Mining Online Deal Forums for Hot Deals. In Proceedings of the Web Intelligence Conference.

Zhou, Y., Reid, E., Qin, J., Chen, H., and Lai, G. (2005). U.S. Extremist Groups on the Web: Link and Content Analysis. *IEEE Intelligent Systems*, 20(5), 44-51.