Preliminary Analytical Considerations In Designing A Terrorism And Extremism Online Network Extractor

Martin Bouchard, PhD; Kila Joffres, MA; Richard Frank, PhD

International CyberCrime Research Center
School of Criminology
Simon Fraser University
mbouchard@sfu.ca, kja4@sfu.ca, rfrank@sfu.ca

Abstract. It is now widely understood that extremists use the Internet in attempts to accomplish many of their objectives. In this chapter we present a web-crawler called the Terrorism and Extremism Network Extractor (TENE), designed to gather information about extremist activities on the Internet. In particular, this chapter will focus on how TENE may help differentiate terrorist websites from anti-terrorist websites by analyzing the context around the use of predetermined keywords found within the text of the webpage. We illustrate our strategy through a content analysis of four types of web-sites. One is a popular white supremacist website, another is a jihadist website, the third one is a terrorism-related news website, and the last one is an official counterterrorist website. To explore differences between these websites, the presence of, and context around 33 keywords was examined on both websites. It was found that certain words appear more often on one type of website than the other, and this may potentially serve as a good method for differentiating between terrorist websites and ones that simply refer to terrorist activities. For example, words such as \terrorist," \security," \mission," \intelligence," and \report," all appeared with much greater frequency on the counterterrorist website than the white supremacist or the jihadist websites. In addition, the white supremacist and the jihadist websites used words such as \destroy," \kill," and \attack" in a specific context: not to describe their activities or their members, but to portray themselves as victims. The future developments of TENE are discussed.

Keywords: Web-Crawler, Extremism, Terrorism, Internet

1 Introduction

Much like any other groups and organizations, extremist and terrorist groups can be found on the Internet, including many who have their official website [21]. Conway [4] has suggested that extremists use the Internet in five general ways: information recruitment, networking, information provision, financing, and recruitment (see also [16]). The Internet's appeal follows from its ability to provide a broad reach, to provide low costs, to be timely and efficient, and to provide some degree of security and anonymity [8]. Tsfati and Weimann [19] further emphasize that the Internet is extremely well suited to terrorists for the purposes of communication as it is decentralized, uncensored, largely free of control or restrictions, and allows for worldwide access for current members or potential recruits.

Despite the identification of the Internet as a tool for terrorist groups, limited empirical research has been conducted into the nature of the terrorism-related content online. However, researchers involved with the Dark Web Project have started to build knowledge on the content and structure of websites hosting terrorism-related content [1]. For example, Zhou et al. [24] proposed a semiautomated methodology (combining the efficiency of automatic data collection and material that can be qualified as \extremism" has yet to be fully understood (the aphorism that one man's terrorist is another man's freedom fighter has yet to find a suitable empirical solution for such purposes). Thus, before undertaking the process of making TENE more intelligent, additional baseline data is needed about the structure and content of single websites in order to establish the ground rules necessary to create a valid web-crawler program.

The current study contributes to this end goal. In this chapter, we use a preliminary version of the web-crawler designed to collect the entire information contained on a single website. In particular, this chapter will focus on how TENE may help differentiate terrorist/extremist websites from other types of related websites by analyzing the context around the use of predetermined keywords found within the HTML of the webpage. We illustrate our strategy through a content analysis of four types of websites. One is a popular white supremacist website, another is a jihadist website, the third one is a terrorism-related news website, and the last one is an official counterterrorist website.

2 Methodology

TENE is a web-crawler that emerges from previous work on extracting online child pornography networks (see [10] [22] [6]). TENE operates by starting the crawling process at user-specified webpages, retrieving the pages from the Internet, analyzing them, and recursively following the links out of the pages. For the purpose of this chapter, the web-crawler starts at a page that covers material broadly associated with extremism or terrorism. Such a webpage can be found by the user, given to the web-crawler by the police, or obtained from terrorism-related literature. The starting website is then retrieved for the crawler, but there is no need to display the content in a web-browser and hence only the HTML (Hypertext Markup Language) of the webpage is retrieved. Certain statistics about the content of webpages are recorded, such as the frequency of user-specified keywords and count of images or videos. In its mature form, TENE will also follow the links found on a webpage if these links point to a webpage that contains extremism or terrorism material. These links will be subsequently explored recursively until certain criteria are met.

As the Internet is extremely large and a crawler would most likely never stop crawling, three conditional limits can be implemented into the web-crawler. These conditions help keep the crawling process under control and the network content-relevant. First, to keep the network extraction time bounded, a limit can be put on the number of pages retrieved (in our previous work on child pornography, that limit was 250,000). Second, the network size may be fixed at a specific number of websites (for example, 500). The webpages are retrieved in such a way that each website is sampled equally, or as equally as possible. Finally, in order to provide some boundaries for the crawl and guide the network extraction process to a relevant network, a set of keywords needs to be defined. For the crawler to include a given webpage in the analysis, the page has to contain a user-defined number of unique keywords.

The end result of the crawling process is knowledge about a set of webservers, including the webpages contained within them, and the links between the webpages. These results are then aggregated up to the server level, with the resulting network summarizing the content on each of the servers, count of keywords, videos, and images, and the links between each of the servers. This essentially creates a map of a terrorism network from the Internet. Note that the version of TENE used for the purpose of this chapter remains within the realm of the initial user-specified website from which it starts. This work will eventually lead to the establishment of rules allowing for automatic identification of a terrorism/extremism-related website from another.

2.1 Keywords

As previously mentioned, to keep the websites crawled in TENE topic-relevant,

Table 2. Number of Keywords per Page for each Website

Keyword	Jihadist	White Supremacist	News	Counterterrorist
Allah	3.63	0.04	1.07	0
Attack	6.38	0.89	0.87	0.6
Black	0.86	3.11	4.79	0
Bomb	0	0.28	0.8	0.53
Combat	0	0.06	0.1	0.47
Counterterrorism	0	0	0.02	4.33
Dead	0	0.15	1.60	0.07
Destroy	0.38	0.57	0.25	0
East	0	1.45	1.24	0
Fight	0	0.68	0.55	0.07
Foreign	0	0.19	0.59	0
Free Speech	0.5	0.04	0.40	0
Infidel	0.13	0.04	0	0
Intelligence	0	0.15	0.32	7.67
Islam	7.38	0.09	0.55	0
Jew	3.5	13.38	2.37	0
Jihad	2.25	0	0.06	0
Join	0.25	0.70	0.29	0.53
Kill	0.13	0.38	1.39	0.07
Live	0.25	1.55	0.91	0.07
Member	0.13	2.89	1.32	1.27
Mission	0	0.13	0.69	1.6
Obama	0	0.28	5.62	0.13
Race	0.38	19.28	11.12	0.67
Report	1	0.77	2.40	5.8
Security	0	0.17	1.04	4.07
Struggle	0.13	0.45	0.21	0
Terrorist	3.13	0.36	0.6	10.93
Train	0.25	0	0.39	0
Victim	0	0.15	0.26	0
Violence	0.63	0.38	0.55	0.13
West	1.5	1.72	1.99	0
White	0	20.98	0.70	5.27

Table 3. Keywords most strongly associated with specific websites

Website(s)	J	WS	N	C	J & N
Keywords	Allah Attack Infidel Islam Jihad	Jew White	Dead Foreign Kill Obama	Combat Counterterrorism Intelligence Security Terrorist	Free Speech Train
Website(s)	WS & S	N & C	WS, N & C	J, WS & N	J, WS, N & C
Keywords	Black East Fight Live Race Struggle Victim	Mission Report	Bomb Member	Destroy Violence West	Join

J = Jihadist; WS = White Supremacist; N = News; C = Counterterrorism

used by various combinations of two or three web-sites. For instance, both the jihadist and the news website used the words \free speech" and \train" more frequently than other web-sites. Similarly, the white supremacist and the news website shared several words, including \black," \East," \fight," \live," \race," struggle," and \victim." The news and counterterrorist website employed the terms \mission" and \report" more frequently than other websites. In some instances, all but one website used a particular word at similar rates. For example, the white supremacist website, \fight," \live," \race,"

{ \The western and American print and electronic media are continuously spitting venom against Islam,"
{ \The Muslims have already suffered too much of violence and tyranny but now the non-Muslim world plans to eliminate the Muslims once for all simply because strong Islam is something intolerable for the non-Muslim forces," and
{ \The enemy has already declared a war against Islam".

In addition, the word \kill" (used 0.13 times per page) was further used to emphasize \the killing of innocent people" at the hands of the U.S.

Similarly, when the white supremacist website used words such as \attack," \bomb," \dead" and \destroy," it was to emphasize the victimization of \whites" at the hands of other \races". For example, \attack" (appearing 0.89 times per page) was used in the context of attacks on freedom of speech, various terrorist attacks against Americans, non-white immigrants as \attacks on freedom," and so on. The word \destroy" (occurring 0.57 times per page) was often used in the same way, describing how:

{ \Mexicans will destroy America,"
{ \Jewish heritage week will destroy American heritage," and
{ \multiculturalist movement will destroy the fabric of White America."

The term \fight" (used 0.68 times per page) was also employed in a similar manner, with the white supremacist website noting that white individuals must \fight for the security and survival of [their] people," fight against \white racism," and fight against organized crime by fighting multiculturalism. Finally, the word \kill" was used to discuss various killings of \Whites" by \Blacks," although it was also used it in other contexts, including engaging in Holocaust denial by questioning the killing of Jewish people and discussing the killing of Saddam Hussein. Overall, a tendency emerged for these websites to use certain words in ways that emphasizes their role as victims with a sympathetic cause rather than as aggressors with a violent agenda. This allows websites to set the stage for encouraging action, further propaganda, and/or for recruitment purposes.

Some of the same words were also used from vastly different perspectives. For instance, each of the websites used the word \terrorist" to refer to American activities in the Middle East, but approached the issue from a different angle. Within the jihadist website, the word was used in reference to \U.S. terrorist attacks" against Islamic nations, to describe the \terrorist war against Muslim ummah [\community"]" launched by the U.S., to describe the \U.S. and its western allies' attack on Afghanistan as the worst kind of terrorism," and to describe America's allies as \terrorists." At the same time, the website seeks to dissociate the word terrorism from Muslims, by emphasizing that the \Noble Quran" does not preach terrorism and that mosques are not training grounds for terrorists, despite western \propaganda" to this effect.

The white-supremacist also uses the word \terrorism" largely to describe America's actions in the Middle East, criticizing the U.S. policy surrounding the

\War against Terrorism." For instance, the website states that \Covert Operations are a huge part of the CIA [and] are simply state- sponsored terrorism," further arguing that \The only way we can end our War against Terrorism, is to end the US practice of conducing Terrorism under the guise of Covert Ops." The website also criticizes the use of detention centers such as Guantanamo Bay to hold so-called \terrorists" or suspected terrorists". The general consensus appears to be that the U.S. should focus more on the state of its nation within its borders rather than outside. As such, while both websites condemn American actions in the Middle East, they do so from different stand points and for different reasons.

Overall, the different types of websites can be differentiated by the different frequency of keywords used. However, both jihadist and white supremacist websites use various words for the same purpose (e.g., portraying themselves as victims) and use the same words for different purposes (e.g., attacking or defending the West or for recruitment or general discussion).

5 Discussion

The \National Strategy for Homeland Security" report in the U.S. emphasized that science and technology were important counter-terrorist tools [14]. It has been suggested that the use of information technology will increase national safety [13] by assisting in intelligence gathering through the collection and analysis of terrorism-related data [3]. This renders the creation of web-crawlers designed to detect and extract networks of extremist or terrorist web-sites a valuable enterprise, as these can build knowledge related to the content (i.e., group activities, recruitment processes, propaganda materials, etc.) of such websites and the affiliations of these groups.

This exploratory study used TENE, a specially designed web-crawler, to explore certain content aspects of two extremist websites (a white supremacist website and a jihadist website), with comparisons made to non-extremist websites (a counterterrorist website and a news website). It was found that all websites could be identified by certain keywords; for instance, the jihadist website used words such as \Allah," \Islam," and \Attack" at greater rates than the other websites. In addition, the extremist websites used language in specific ways, with words such as \attack," \destroy," and \dead" being used to emphasize the group's role as a victim. It should be stressed that the number of websites selected is small; as such, this project is entirely exploratory in nature, designed to provide preliminary information on how extremist websites might be identified by a web-crawler and how they might be used by terrorist or extremist groups.

Past studies have also explored the presence of extremist groups on the Internet and developed tools to collect extremist websites [25] [3] [7] [17] [19] [20]. Some organizations, including SITE institute, the Anti-Terrorism Coalition, and the Middle East Media Research Institute (MEMRI) have used manual analysis techniques to collect and monitor extremist websites. The Artificial Intelligence Lab uses automated processes for collection building. The Dark Web project has combined both manual and automated processes to build and analyze collections with the goal of combining the efficiency of automated techniques with the accuracy of manual ones. The TENE project seeks to extend past work on ism and the spread of propaganda. We also envision TENE as an integral part

17. Thomas, T.L. (2003). Al Qaeda and the Internet: The danger of Cyberplanning'. Parameters, 33, 112-123.
18. Tong, S., & Koller, D. (2002). Support vector machine active learning with applications to text classification. Journal of Machine Learning Research, 2, 45-66.
19. Tsfati, W., & Weimann, G. (2002). www.terrorism.com: Terror on the Internet. Studies in Conflict & Terrorism, 25(3), 317-332.
20. Weimann, G. (2004). www.terror.net: How modern terrorism uses the Internet. Special Report, US Institute of Peace. Retrieved from http://www.usip.org/pubs/specialreports/sr116.pdf
21. Weimann, G. (2006). Terror on the Internet: The new Arena, The New Challenges. Washington, D.C.: United States Institute of Peace.