A Novel Approach for Social Network Analysis & Web Mining for Counter Terrorism

Prof. G. A. Patil
Computer science & engineering department,
D. Y. Patil College of Engineering & Technology, Kolhapur.
Kolhapur, India
gasunikita@yahoo.com

Prof. K. B. Manwade
Computer science & engineering department,
Ashokrao Mane Group of Institution, Vathar.
Vathar tarf Vadgaon, India
mkarveer@gmail.com

Mr. P. S. Landge
Computer science & engineering department,
D. Y. Patil College of Engineering & Technology, Kolhapur.
Kolhapur, India
pramodslandge@gmail.com

***Abstract*-Terrorists and extremists are increasingly utilizing Internet technology as an effective mode to enhance their ability to influence the outside world. Lack of multilingual and multimedia terrorist/extremist collections and advanced analytical methodologies; limit our experiential understanding of their Internet usage. To address this research gap, we explore an integrated approach for identifying and collecting terrorist/extremist Web contents and to discover hidden relationships among communities.**

It has been shown in the literature that content analysis gives more insight of technical sophistication, content richness; whereas the link analysis focuses on the web interactivity. A dark web attribute system has made the sincere effort on identifying and comparing terrorist website with genuine web sites by using content and link analysis still there is scope in the same area as proposed in [1].

This proposed work focuses on identifying & analyzing new web page attributes. It is aimed to compare different terrorist/extremist sites with genuine sites and accordingly prepare metrics which can be further used for identification of other sites of terrorist/extremist groups. Also proposed work focus on to visualize and analyze hidden domestic terrorism communities and intercommunity relationships among all web sites in our collection.

Keywords: -Web content analysis; Web usage analysis; Web collection building, Social Network.

I. INTRODUCTION

Terrorist organizations have generated thousands of Web sites that support psychological warfare, fundraising, recruitment, coordination, and distribution of propaganda materials. The level of technical sophistication of the Islamic terrorist organizations' Web sites has increased according to Katz, who monitors Islamic fundamentalist Internet activities. The rapid proliferation and increased sophistication of Web sites and online forums run by terrorist/extremist organizations are indications of the growing popularity of the Internet in terrorism campaigns. They also indicate that there is a vast pool of sympathizers that such organizations have attracted, with some applying their IT expertise as contributions to the cause.

The Web has evolved towards multimedia-rich content delivery, end user personal content generation, and community-based social interactions. Due to the freedom and convenience of publishing in Weblogs, [2] [3] this form of media provides an ideal environment as a propaganda platform for extremist or terrorist groups to promote their ideologies. Criminals may also make use of the virtual environment to organize crimes such as money laundering and drugs trafficking without being easily identified. As a result, it is important to understand the social network of the bloggers in order to assess the risks that may threaten the national security.

II. LITERATURE SURVEY

In social networks analysis the main task is usually about how to extract social networks from different communication resources [4]. The data used for building social networks is relational data, which can be obtained and transferred from different resources including the web, email communication, Internet relay chats, telephone communications, organization and business events, etc.

In recent years, there have been studies of how terrorists use the Web to facilitate their activities. The first step towards studying terrorists' tactical use of the Web is to build a high-quality Dark Web [5] collection.

The rapid expansion of the web is causing the constant growth of information, leading to several problems such as increased difficulty of extracting potentially useful knowledge. Web content mining [6] confronts this problem gathering explicit information from different web sites for its access and knowledge discovery. Web mining is concerned with the use of data mining techniques to automatically discover and extract information from World Wide Web documents and services. Web content mining approach to extract information from web based databases.

The DWAS is an effective tool to analyze the technical sophistication of terrorist/extremist groups' Internet usage and could contribute to an evidence based understanding of the applications of Web technologies in the global terrorism phenomena. DWAS is used to visualize and analyze hidden domestic terrorism community and intercommunity [7] relationship. The DWAS helps in identifying the groups that are considered by authoritative sources as terrorist/extremist groups. The sources include government agency reports, authoritative organization reports and studies published by terrorism research centers. Also DWAS identify a set of seed terrorist group URLs from the authoritative sources and the terrorism keyword lexicon to query major search engines on the Web. After identifying the seed URLs, out-links and in-links of the seed URLs were automatically extracted using link-analysis programs. Once the terrorist/extremist Web sites are identified, a program is used to automatically download all their contents. The DWAS framework focuses on the attributes that could help us better understand the level of advancement and effectiveness of terrorists' Web usage, namely, technical sophistication attributes, content richness attributes (an extension of the traditional media richness attributes), and Web interactivity attributes.

Still DWAS have scope for improvement in identifying and analyzing new attributes for content analysis, applying new data mining algorithm for link analysis as suggested in [1][7][8].

A. Terrorism and the internet

In social networks analysis the main task is usually about how to extract social networks from different communication resources. The data used for building social networks is relational data, which can be obtained and transferred from different resources including the web, email communication, Internet relay chats, telephone communications, organization and business events, etc. For example, email communications are a rich source for extracting and constructing social networks. By means of email social networks extraction, the relationship between email senders and receivers can be transformed by measuring the frequency of email communication and taking the communication behavior (such as reply, forward, etc.) into account. The transformed relational data can then be used for social networks construction.

Terrorist groups have sought to replicate or supplement the communication, fundraising, propaganda, recruitment, and training functions on the Internet by building Web sites with massive and dynamic online libraries of speeches, training manuals, and multimedia resources that are hyperlinked to other sites that share similar beliefs.

B. Dark web collection building

The first step towards studying terrorist's tactical use of the web is to build a high-quality Dark web collection. Web collection building is the process of gathering and organizing unstructured information from pages and data on the Web. We propose to use a semi-automated approach to collecting Dark web contents.

The data is identified for the terrorist name and their URLs of terrorist groups from Government report such as FBI, US State department and research centers MEMRI, ATC etc. Web crawler is further used to automatically extract the URL out-link and in-link. Afterwards, the similarity between web site A and B will be calculated by using following formula,

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

Where lv (L) is the level of link L in the web sites hierarchy, with the homepage as level 0 and each lower level in the hierarchy is increased by one.

Graphs are generated in order to visualize the link associations and further analyzing those using Java universal network graph (JUNG). The link tree would be having a seed and its directly associated URLs in the first level. The next levels (i.e. second and third level) depict further expansion of the association to gather enough no. of URL for the study.

Then applies the robust filtering method to find essential websites and remove the unwanted websites, which would be useful for further analysis purpose. Extending and implementing the web crawler to download the entire web document within these sites, enables the parameter gathering.

C. Dark web content analysis

In order to reach an understanding of the various facets of terrorist Web usage and communications, a systematic analysis of the Web sites' content is required. Due to analyze purpose we need to find out web interactivity attributes, media richness attributes, and technical sophistication attributes. For example,' use of HTML frames'',' use of Java scripts,'' etc. The media richness attributes measure how well the Web sites use multimedia to deliver information to their users, e.g., ''hyperlinks,'' ''images,'' ''video/audio files,'' etc.

Web interactivity has been widely adopted by researchers in e-Government and e-Education domains to evaluate how well Web sites facilitate the communications among Web site owners and users. The Web interactivity attributes can be consisting into three categories: one-to one-level interactivity, community-level interactivity, and transaction-level interactivity.

The one-to-one-level interactivity attributes measure how well the Web sites support individual users to give feedback to the Web site owners (e.g., provide email contact, provide guest book functions, etc.). The community- level interactivity attributes measure how well the Web sites support the two-way interaction between site owners and multiple users (e.g., use of forums, online chat rooms, etc.). The transaction-level interactivity measures how well users are allowed to finish tasks electronically on the Web sites (e.g., online purchasing, online donation, etc).

III. MODIFIED DARK WEB ATTRIBUTE SYSTEM FOR COUNTER TERRORISM

The aim of this proposed work is, to identify the terrorists name and their web sites and then download the web site contents for analysis purpose. The various steps involved in this such as, identify terrorist groups, identify terrorist group URLs, and expand terrorist URL set through link and forum analysis. Also to cluster the related websites and define new set of Web interactivity attributes for calculating the web content to understand the level of advancement and effectiveness of terrorists' Web usage. These new set of attributes as follows,

TABLE I. WEB INTERACTIVITY ATTRIBUTES

ATTRIBUTES

WI attributes	Weights	WI attributes	Weights
Community-level interactivity		Transaction-level Interactivity	
flash files			
Videoconference	5	Online Recruitment	4.5
flash files			
E- Tendering	4.5	Comment	2.75
flash files			
Contact	1.25		
flash files			
Email	1.75		
List	2.25		

TABLE II. CONTENT RICHNESS

CR attributes	Scores
Flash	No. of
Videoconference	No. of
Online Recruitment	No. of
E-tendering	No. of

TABLE III. TECHNICAL SOPHISTICATION ATTRIBUTES

TS attributes	Weights	TS attributes	Weights
Basic HTML techniques		Advanced HTML	
Use of Style	2.5	Use of Ajax	4.5
Use of Meta element	2.5	Use of XML	3.5
Use of Label	1	Use of GrviL	3
Use of Menu	2	Dynamic web programming	
Use of Object	3	Use of Java	2.5
Use of Param	1.5	Use of Java Script	3.5
Use of Span	3	Use of Python	6
Form	1.2	Use of SMSL	5.5
Frame	2	Use of Script	3.5
Table	2		
Embedded multimedia			
Use of Slide	3		
Use of Applet	1		

The proposed Dark Web system architecture and various modules are shown in figure 1.

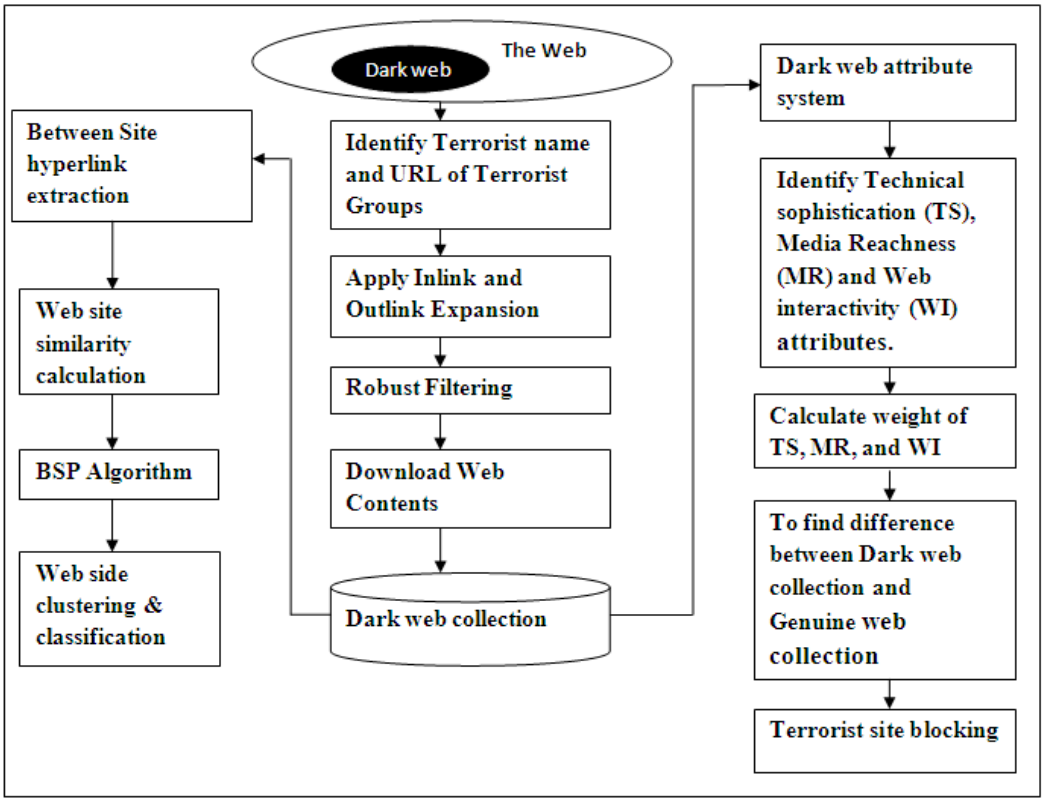


Figure 1. Modified Dark web Attribute system architecture

A Modified dark Web Attribute system

The Modified dark Web Attribute system will have the following modules

Module1. Dark web collection

Identify the terrorist name and URLs of terrorist groups from dark web. Then using link analysis program to automatically extract the URL out-link and in-link. The robust filtering method will be applied to identify essential terrorist websites. Then, using an automatic web crawling toolkit called spidersRUs download all the web document within these sites.

Module2. Content analysis of terrorist web sites

Identifying Technical Sophistication (TS), Media richness (MR) and Web interactivity (WI) attribute. Then calculate the weight of that attribute. Finally find out benchmark comparison result between terrorist websites and genuine websites.

When data from all websites belonging to a cluster is aggregated and the normalized content level is calculated into six dimensions. Each dimension represents a normalized activity scale between 0 and 1, showing the degree of activity on the dimension. The activity scale of cluster c on dimension d is calculated by the following formula and n is the total number of attributes in dimension d, while m is the total number of web sites belonging to cluster c.

$$\text{Activity Scale } (c, d) = \sum W\, i, j / m \text{ x } n$$

Module3. Link analysis

To find relationship among different web sites for the same group and the interaction with other extremist group, first step is to calculate similarity between all web site pair in the collection. Similarity can be defined as real value multivariable function of the number of hyperlink in web site A pointing to web site B and the number of hyperlink in site B pointing to site A. Hyperlink appear at sites homepage has a higher weight than hyperlink appearing at a deeper level. The similarity between web site A and B will be calculated by using following formula,

$$\text{Similarity } (A, B) = \sum_{\substack{\text{All links} \\ \text{Between A and B}}} \left(\frac{1}{1 + \text{lv}(L)} \right)$$

Where lv(L)= he level of link L in the web sites hierarchy. This varies from level 0 to N by increased by 1.

Module4. Clustering of terrorist web sites

The Business system planning (BSP) clustering algorithm will be used to form the clusters of terrorist web sites.

IV. EXPERIMENTAL RESULT

I. Dark web collection & Link Analysis

Following accomplishments were achieved while doing experimental work on two modules; Data was identified for terrorist organizations from Government report such as FBI, US State department and research centers MEMRI, ATC etc. The URLS have been crawled for link analysis, identifying their similarity, the graphs were plotted to study their relationship further. The content was downloaded for shortlisted URLs. There are twenty seven types of attributes that were selected for analysis purpose. Five major attribute groups were formed as, Technical Sophistication attribute (by grouping menu, meta, label, style, span, form, frame and table attributes), Advanced technical sophistication (by grouping java script and script attributes), Dynamic web programming (by grouping Java, PHP, ASP attributes), Content Richness (by grouping Flash, image, audio, video, no. of hyperlink, no. of download document attributes), and Web interactivity attributes (by grouping list, contact, email, comment, videoconference, online recruitment and E-tendering attributes).

Figure2. Shows results of tree generated for first level. This tree depicts URLs which are directly linked to the given seed URL.

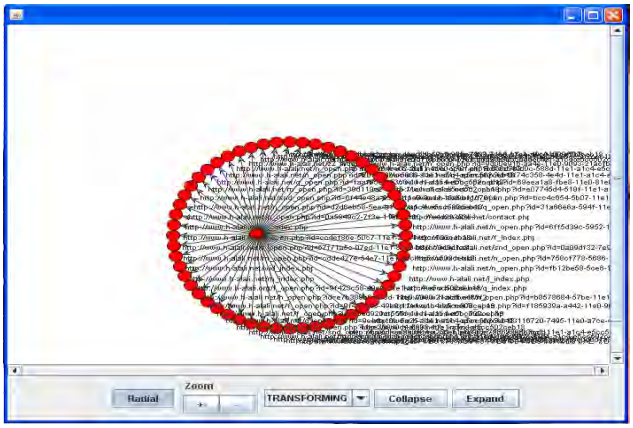


Figure 2. Link Tree for first Level

Figure 3. Shows results of tree generated for second level. This tree indicates hierarchy of linking for two level URLs in website.

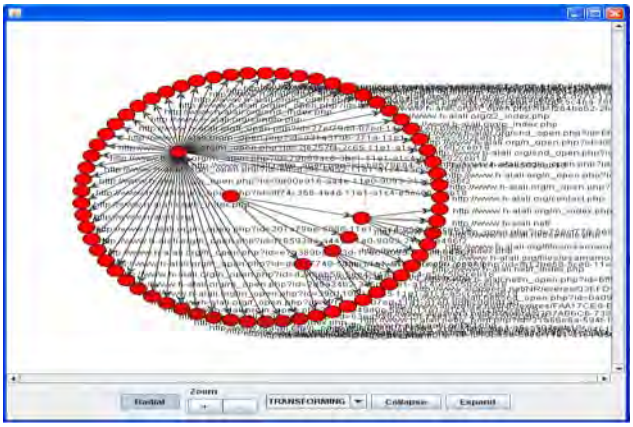


Figure 3. Link Tree for second Level

Figure 4. Shows results of tree generated for third level. This tree indicates hierarchy of linking for third level URLs in website.

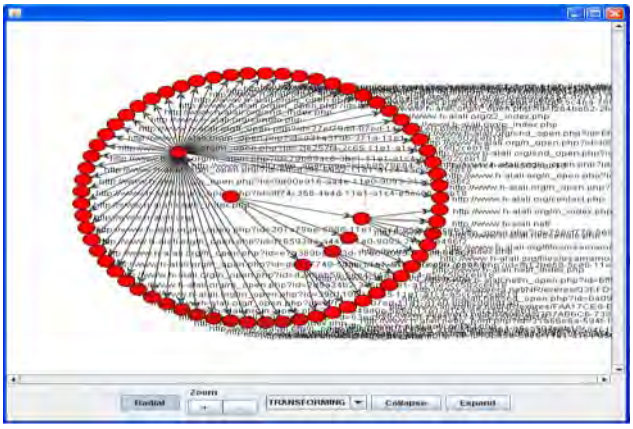
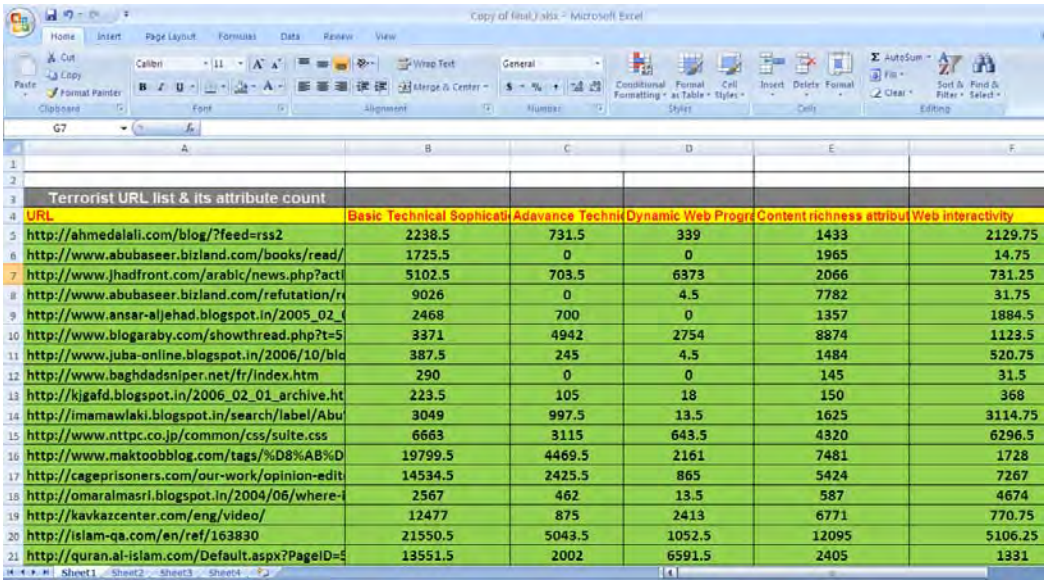


Figure 4. Link Tree for third Level

II. Content analysis of terrorist web sites

For creation of the benchmark the pool was increased to 30 URLs. Content was downloaded for these websites, using the experimental setup that is in place.

The attribute counts were taken, for the information gathered. Further the calculations were carried out for each attribute according to its weight assigned. Given below is a snapshot 1 of the attributes gathered and their calculations.



Terrorist URL list & its attribute count					
URL	Basic Technical Sophicati	Adavance Techni	Dynamic Web Progr	Content richness attribut	Web interactivity
http://ahmedalali.com/blog/?feed=rss2	2238.5	731.5	339	1433	2129.75
http://www.abubaseer.bizland.com/books/read/	1725.5	0	0	1965	14.75
http://www.jhadfront.com/arabic/news.php?actl	5102.5	703.5	6373	2066	731.25
http://www.abubaseer.bizland.com/refutation/r	9026	0	4.5	7782	31.75
http://www.ansar-aljehad.blogspot.in/2005_02_	2468	700	0	1357	1884.5
http://www.blogaraby.com/showthread.php?t=5	3371	4942	2754	8874	1123.5
http://www.juba-online.blogspot.in/2006/10/blo	387.5	245	4.5	1484	520.75
http://www.baghdadsniper.net/fr/index.htm	290	0	0	145	31.5
http://kjgafd.blogspot.in/2006_02_01_archive.ht	223.5	105	18	150	368
http://imamawlaki.blogspot.in/search/label/Abu	3049	997.5	13.5	1625	3114.75
http://www.nttpc.co.jp/common/css/suite.css	6663	3115	643.5	4320	6296.5
http://www.maktoobblog.com/tags/%D8%AB%D	19799.5	4469.5	2161	7481	1728
http://cageprisoners.com/our-work/opinion-edit	14534.5	2425.5	865	5424	7267
http://omaralmasri.blogspot.in/2004/06/where-i	2567	462	13.5	587	4674
http://kavkazcenter.com/eng/video/	12477	875	2413	6771	770.75
http://islam-qa.com/en/ref/163830	21550.5	5043.5	1052.5	12095	5106.25
http://quran.al-islam.com/Default.aspx?PageID=5	13551.5	2002	6591.5	2405	1331

Snapshot-1: Attribute count for a terrorist website.

The entire information is combined to form the total of each URLs attribute count. A consolidated sheet has been created which represents data for the entire group.

Also, 8 genuine websites were identified and their data was gathered for comparison. Snapshot 2 shows the attribute counts for the second Group of genuine websites.

Genuine URL	Basic Technical	Adavance Technical	Dynamic Web Programming	Content richness	Web interactivity
http://india.gov.in/	12304.5	5766	17777.5	11991	11453.75
http://www.australia.gov.au/sti	12649	5743	3299	13668	10849
http://www.usa.gov/Citizen/	16767.5	6594	11908.5	9337	18778.25
http://www.direct.gov.uk/en/	22090	5526.5	12976	11920	19592.75
http://www.infosys.com/	33073	15831	55184	23506	32772
http://www.cdac.in/	11618	4774	8421	7474	5227.8
http://www.rcom.co.in/	23252	5376	82.5	12596	18007
http://www.tcs.com/	21531	10199	47588	6376	19994
MEAN	19160.625	7476.1875	19654.5625	12108.5	17084.31875
MEDIAN	19149.25	5754.5	12442.25	11955.5	18392.625
STANDARD DEVIATION	7334.063309	3768.997991	20450.31919	5273.158446	8243.977025
Relative Standard Deviation (RDS)(%)	38.27674363	50.41336899	104.0487123	43.54922944	48.25464302
Confidence	5082.153194	2611.734365	14171.08779	3654.045229	5712.679643
Max	24242.77819	10087.92186	33825.65029	15762.54523	22796.99839
Min	14078.47181	4864.453135	5483.474709	8454.454771	11371.63911

Snapshot 2: Consolidated attribute counts for the Genuine URL list.

TABLE IV. ATTRIBUTES COUNT IN TERRORIST & GENUINE WEBSITES

		Basic Technical Sophication	Adavance Technical Sophistication	Dynamic Web Programming	Content richness attribute	Web Interactivity
GROUP - I:	Max	12431.40912	3770.333842	5008.259172	8321.253332	5170.006754
Terrorist/Extremist	MEAN (AVG)	8878.75	2636.15625	3123.21875	5926.78125	3267.832813
Websites Count :	Min	5326.090879	1501.978658	1238.178328	3532.309168	1365.658871
GROUP - II:	Max	24242.77819	10087.92186	33825.65029	15762.54523	22796.99839
Genuine Websites	MEAN (AVG)	19160.625	7476.1875	19654.5625	12108.5	17084.31875
Count :	Min	14078.47181	4864.453135	5483.474709	8454.454771	11371.63911

The Table IV shows the count for the attributes in terrorist/extremist websites & genuine websites. Different statistical calculations were carried out on the data, in order to finalize the benchmark.

The average result (Mean):

x is calculated by summing the individual results and dividing this sum by the number (n) of individual values:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + x_4 + \ldots.}{n}$$

The standard deviation:

It is a measure of how precise the average is, that is, how well the individual numbers agree with each other. It is a measure of a type of error called random error - the kind of error people can't control very well. It is calculated as follows:

$$\text{standard deviation, } S = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + \ldots}{n - 1}}$$

The relative standard deviation (RSD):

It is an often time more convenient. It is expressed in percent and is obtained by multiplying the standard deviation by 100 and dividing this product by the average.

$$\text{relative standard deviation, RSD} = 100S/\bar{x}$$

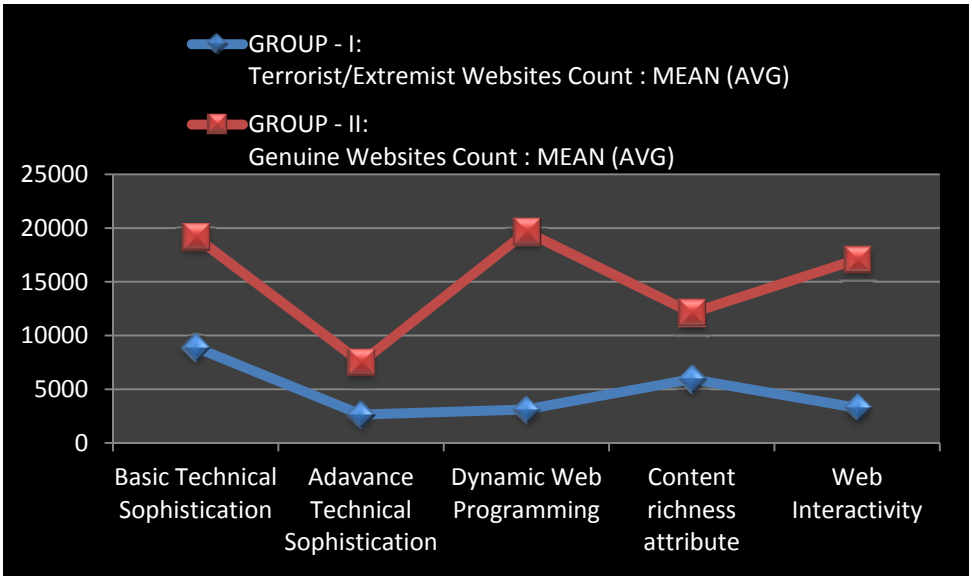
Confidence Interval:

In statistics, a confidence interval (CI) is a particular kind of interval estimate of a population parameter and is used to indicate the reliability of an estimate. It is an observed interval (i.e. it is calculated from the observations), in principle different from sample to sample, that frequently includes the parameter of interest, if the experiment is repeated. How frequently the observed interval contains the parameter is determined by the confidence level or confidence coefficient.

$$C.I. = \bar{X} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

where $\bar{X}$ = the sample mean
σ = the population standard deviation
$Z_{\frac{\alpha}{2}}$ = the Z value for the desired confidence level α (obtained from an Area Under the Normal Curve table)

Z is the significance level used to compute the confidence level. The confidence level equals 100*(1 - alpha) %, or in other words, an alpha of 0.05 indicates a 95 percent confidence level. Here, we have considered z as 0.05 which signifies 95% confidence.



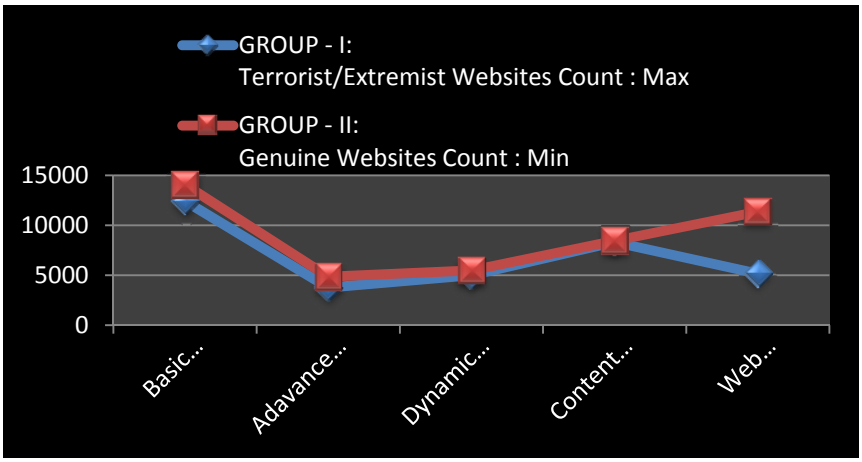
Graph No. 1: Comparison between averages of counts for all the attributes, for both the groups.

The Graph No.1 shows clear difference between averages of both the groups. The mean count for genuine websites groups is higher for every attribute.

Further, the Graph No. 2 were plotted using following formula, and two different bands were observed for terrorists and genuine websites.

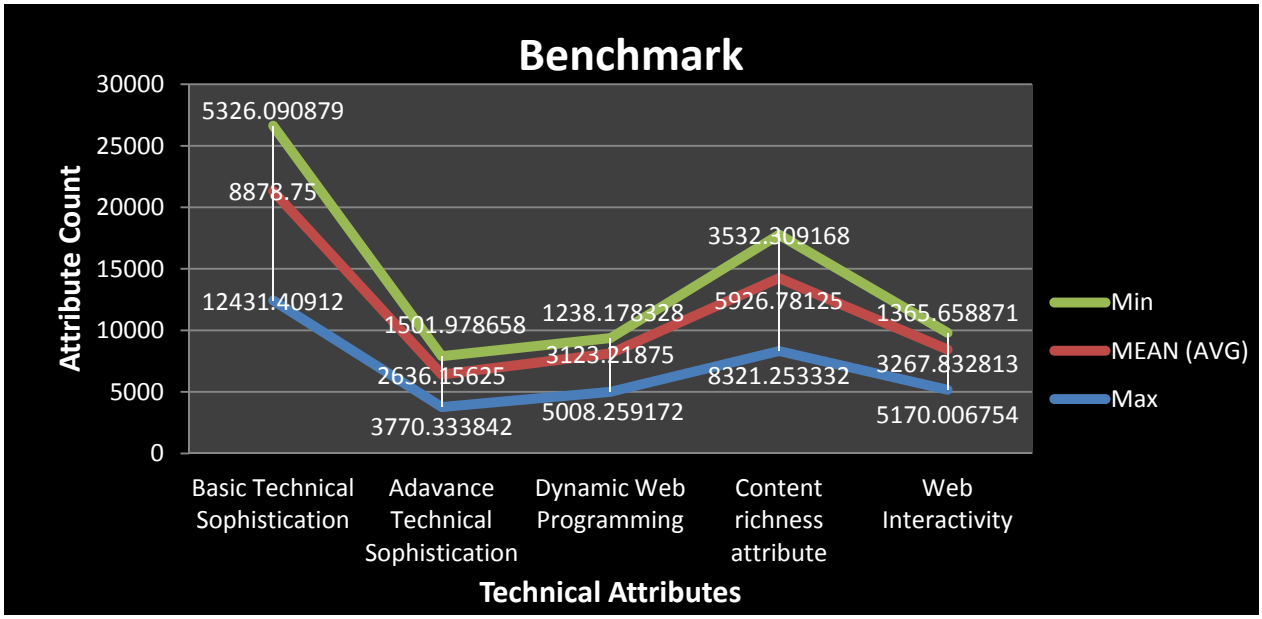
$$\text{mean} \pm \text{confidence}$$

Following Graph No. 2 was plotted to study the difference/ gap between the two bands, Hence, the lower edge of the upper band and upper edge of the lower band have been considered, in below graph.



Graph No. 2. AVG+CONFIDENCE (Group-I) and AVG-CONFIDENCE (Group-II)

Here, it is clearly visible that, there is a distinguished gap between web usage of terrorists, and that of the genuine organizations. However, the only overlapping can be seen for Content Richness, which emphasizes the focus that has been put on to audio visual content by the terrorists to appeal to the viewers.



Graph No. 3. Benchmark of terrorist websites

The Graph No. 3 shows the benchmark of terrorist websites. A benchmark has been devised from the data and its mean, standard deviation and confidence interval. A new website can be checked against this benchmark to decide its primary inclusion in the terrorists/extremists list, for further analysis like clustering and blocking.

V. CONCLUSION AND FUTURE WORK

Terrorists and extremists are increasingly utilizing Internet technology to enhance their ability to influence the outside world. We have focused on advanced analytical methodologies, for monitoring/collecting vital attributes being used by these groups. In this paper, we proposed new attributes for dark web to analyze the terrorist's tactical use of the internet.

Also new algorithm for link analysis is proposed to explore the other terrorist web sites. The content analysis gives more insight of technical sophistication, content richness whereas the link analysis focuses on the web interactivity.

The Terrorist/Extremist's web usage is more concentrated on content richness. Other attributes like web interactivity, are less focused, which helps them in one way communication. Average usage of Basic/Advanced Technical Sophistication and Dynamic Web Programming, is observed with the terrorist websites.

A benchmark has been devised from the data and its mean, standard deviation and confidence interval. A new website can be checked against this benchmark to decide its primary inclusion in the terrorists/extremists list, for further analysis like clustering and blocking.

Future work we plan to explore more advanced machine learning technique to detect technology and media usage pattern in terrorist web sites to gain more insight into terrorist usage. Also we plan to identified and clustering of terrorist/ extremist groups by using clustering algorithm.

REFERENCES

[1]. Jialun Qin, Yilu Zhou, Edna Reid, Guanpi Lai, Hsinchun Chen "Analyzing terror campaigns on the internet: Technical sophistication, content richness, and Web interactivity", International journal of human computer studies, Nov 2006.

[2]. Hsinchun Chen, Sven Thoms, T. 1. Fu, "Cyber Extremism in Web 2.0: An Exploratory Study of International Jihadist Groups", IEEE International Conference on Intelligence and Security Informatics, 2008.

[3]. Michael Chau, Jennifer Xu, "Mining communities and their relationships in blogs: A study of online hate groups", International journal of human computer studies, Oct 2006.

[4]. Peter A. Gloor, Jonas Krauss, Stefan Nann Kai Fischbach, Detlef Schoder, "Web Science 2.0: Identifying Trends through Semantic Social Network Analysis", International conference on computational science & engineering, 2009.

[5]. Chen, H., Qin, J., Reid, E., Chung, W., Zhou, Y., Xi, W., Lai, G, "The Dark Web Portal: Collecting and Analyzing the Presence of Domestic and International Terrorist Groups on the Web", Proceedings of International IEEE Conference on Intelligent Transportation Systems 2004.

[6]. Shohreh Ajoudanian, and Mohammad Davarpanah Jazi "Deep Web Content Mining" World Academy of Science, Engineering and Technology 492009

[7]. Jialun Qin, Yilu Zhou, Edna Reid, Guanpi Lai, Hsinchun Chen "Us Domestic Extremist Groups on the Web: Link and Content Analysis", IEEE intelligent system October / September 2005"

[8]. Sanjeev Sharma and R. K. Gupta " Improved BSP clustering Algorithm for Social Network Analysis", International journal of grid and Distributed Computing Vol. 3, September,2010