



Link & Structure Analysis Of Dark Websites

SELVI.C

Assistant professor

Department of computer science,
Veltech Ranga Sanku Arts College, Avadi
Selvichinnasamy07@gmail.com

Abstract—This chapter presents basic technologies and tools utilized by Mining, a specialized space of the thus known as Structure Mining (web mining applied to the Web).net Link stays at the crossroads of knowledge Retrieval, info Extraction and Content Mining. each info Retrieval {and info and knowledge and knowledge} Extraction play necessary roles to find and extract valuable information out of structured data, before it's appropriate to be processed by net mining applications. Exploring higher these techniques is extraordinarily necessary to deal with the quantity of accessible knowledge within the info Overload Era. These theses specialize in the study of presence of no of links, file size and its kind in designated dark net sites gift within the web. It additionally analyses the link kind, and its structure and updated content.

Keywords: net Mining,
Structure Analysis.

Link

Analysis,

I. INTRODUCTION

T HIS 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 Researchers find that monitoring extremist and hate groups' Web sites and analyzing their usage and content have become time consuming and challenging. In response, this study describes the development of automated or semi-automated methodologies for capturing, classifying, and organizing domestic extremist Web site data and using them for analysis. We found that by analyzing the hyperlink structures and content of domestic extremist Web sites and constructing social network maps, their inter-organizational structure and cluster affinities could be identified.



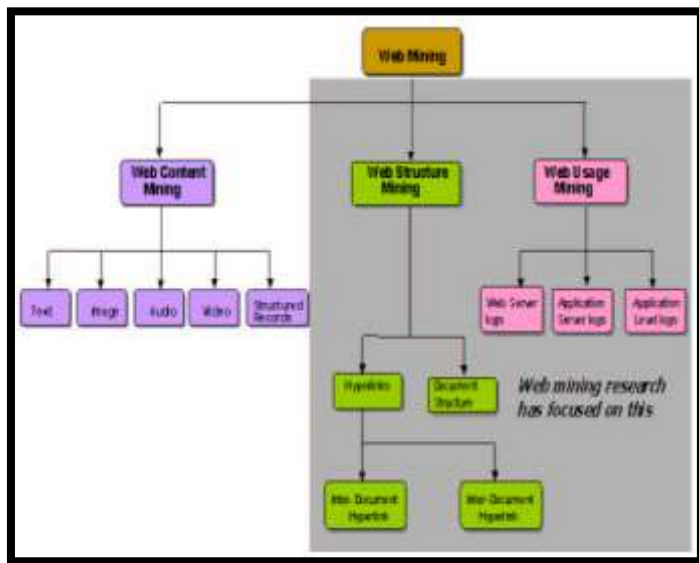
II. LITERATURE SURVEY

WEB MINING:

Web mining is the application of data mining techniques to extract knowledge from Web data, where **at least one of structure (hyperlink) or usage (Web log) data is used in the mining process** (with or without other types of Web data).

The simplicity and the quickness of the search facility, makes it the most successful search engine. Earlier search engines concentrated on the Web content to return the relevant pages to a query. Google was the first to introduce the importance of the link structure in mining the information from the web. Page Rank, that measures an importance of a page, is the underlying technology in all Google search products. The usage data can also be split into three different kinds on the basis of the source of its collection: on the server side, the client side, and the proxy side.

Web Mining can be broadly divided into three distinct categories, according to the kinds of data to be mined:



1. Web Content Mining

Web Content Mining is the process of extracting useful information from the contents of Web documents. Content data corresponds to the collection of facts a Web page was designed to convey to the users. It may consist of text, images, audio, video, or structured records such as lists and tables. Text mining and its application to Web content has been the most widely researched. Some of the research issues addressed in text mining are, topic discovery, extracting association patterns, clustering of web documents and classification of Web Pages. Research activities in this field also involve using techniques from other



disciplines such as Information Retrieval (IR) and Natural Language Processing (NLP). While there exists a significant body of work in extracting knowledge from images - in the fields of image processing and computer vision - the application of these techniques to Web content mining has not been very rapid.

2. Web Structure Mining:

The structure of a typical Web graph consists of Web pages as nodes, and hyperlinks as edges connecting between two related pages. Web Structure Mining can be regarded as the process of discovering structure information from the Web. This type of mining can be further divided into two kinds based on the kind of structural data used.

Hyperlinks

A Hyperlink is a structural unit that connects a Web page to different location, either within the same Web page or to a different Webpage. A hyperlink that connects to a different part of the same page is called an **Intra-Document Hyperlink**, and a hyperlink that connects two different pages is called an **Inter-Document Hyperlink**. There has been a significant body of work on hyperlink analysis, of which provides an up-to-date survey.

Document Structure

In addition, the content within a Web page can also be organized in a tree-structured format, based on the various HTML and XML tags within the page. Mining efforts here have focused on automatically extracting document object model (DOM) structures out of documents.

3. Web Usage Mining

Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site. Web usage mining itself can be classified further depending on the kind of usage data considered.

Link Analysis

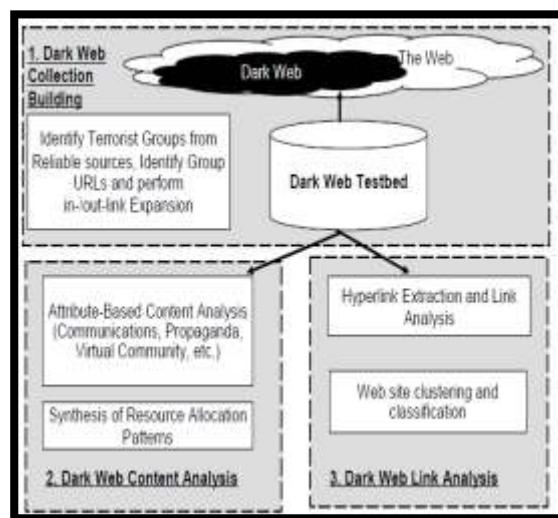


Figure: Link Analysis.

**A Modern Approach to Designing Implementation
and Reinforcement of Quality Management**

System Organised by

ZES's, Zeal Institute of Management and Computer Application,
Narhe, Pune, Maharashtra, India
on 21st November 2019



Our goal here is to shed light on the infrastructure of extremist and hate Web sites and to perform a sophisticated content analysis. We believe the exploration of hidden communities over the Web can give insight into the nature of relationships between Web sites from the same group, as well as relationships between Web sites of different extremist groups. In addition, hyperlinks between Web sites constitute an important cue for estimating the content similarity of any pair of Web sites in our collection. Hence, we employ this cue to confirm our initial manual classification of the Web sites. Uncovering hidden Web communities involves calculating a similarity measure between all pairs of Web sites in our collection. We define similarity to be a real-valued multivariable function of the number of hyperlinks in Web site “A” pointing to Web site “B”, and the number of hyperlinks in Web site “B” pointing to Web site “A”. In addition, a hyperlink is weighted proportionally to how deep it appears in the Web site hierarchy. For instance, a hyperlink appearing at the homepage of a Web site is given a higher weight than hyperlinks appearing at a deeper level. Thus, the similarity between Web sites is

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

“A” and “B” is calculated as follows:

where $\text{lv}(L)$ is the level of link L in the Web site hierarchy, with homepage as level 0 and the level increased by 1 with each level down in the hierarchy. The similarity matrix is then fed to a multidimensional scaling (MDS) algorithm which generates a two dimensional graph of the Web sites.

Link Structure Analysis

From the study we outline several generic concepts that we expect to be relevant across Web sites. They enable us to introduce semantics for describing site structure and devise a parsing technique that would be applicable to most of the sites. First, the links appear to be grouped by various types of associations. Second, among link groups we can single out those with a particular function, i.e., the links that enable navigation or high level organization of the site. The pages may have ‘content areas’ that contain groups of links or individual links that provide access to another HTML page or to any other type of content, such as PDF, PS, DOC files, and similar. Based on this, we set our objectives to define operational definitions of the three types of concepts that we use in the site structure representations.

Organizational and navigational link blocks.

A Modern Approach to Designing Implementation and Reinforcement of Quality Management

System Organised by

ZES's, Zeal Institute of Management and Computer Application,
Narhe, Pune, Maharashtra, India
on 21st November 2019



These are blocks of links typically repeated across pages with the same layout and underpinning the organization of the site. They are often link lists uninterrupted by other content elements such as text. For simplicity we refer to them as structural link blocks.

The content link blocks.

are expected to be grouped by content association. However, they are not likely to be repeated across pages but rather point to information resources.

Isolated links.

Besides the link blocks, pages often contain links that are not part of a link group and may be only loosely related to each other, for example, by their location within the same paragraph of text. With regards to the structure analysis algorithm, our objective is to achieve properties that current representations do not have.

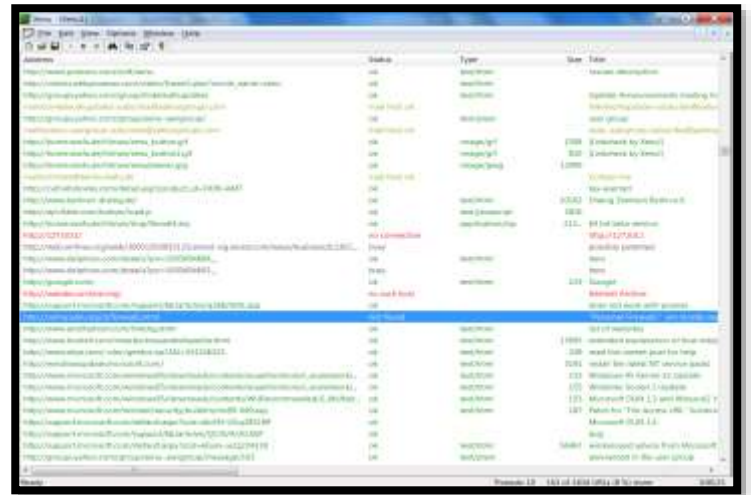
Locality

The enable us to identify both the global site structure and the local structure around individual pages. Pages that are deemed important by the site author are usually accessible via navigation menus or a site map. However, a large proportion of pages are not accessible from the main navigational structure. In order to help the users orientate themselves it is important to identify relevant site context for each page.

ETHODOLOGY

In this research on collect to the Dark websites, making various problem analysis there are having mining concentrates on such as, First, finding the Broken links, Second, The ordered by link and its pages. Third, Local links on broken page make the orphan files web contents of same category. Most of the researchers in the help of Xenu's link mining concentrate more on valid URLs can submit to a search engine.

- Apply to all jobs checked: you don't want to have change these settings for every project. Ask for password or certificate when needed will allow you to spider hidden parts of the site.



**A Modern Approach to Designing Implementation
and Reinforcement of Quality Management**

System Organised by

ZES's, Zeal Institute of Management and Computer Application,

Narhe, Pune, Maharashtra, India

on 21st November 2019



▪ Redirections as errors should be off. While I do consider redirections errors for the most part, they are less urgent to address than broken links, especially internal ones. FTP and gopher URLs. Should be checked. If you have these links, it would be good to know if they are working or not.

downloads the whole file or just touches it to make sure there is something on the other end.

▪ Valid text URLs will give you a full list of all the URLs in your site. Site Map will create a sort of sitemap based on site structure. It's generally not been satisfactory for modern sophisticated dynamic site. Statistics will give you a very good summary of your scan.

Web Server Data

They correspond to the user logs that are collected at Web server. Some of the typical data collected at a Web server include IP addresses, page references, and access time of the users.

Application Server Data

External URL's by analysis result of the presence of URL link and its Percentage calculated, in this chapter

These errors are created as a result of a connection failing to be established to a Web server. They do not have a response code but will contain a brief description of why a connection could not be established.

Xenu's Label	Contents
Pending	Link placed in the queue to check list.
Ok	Correct links.
Skip type	Link does not verify.
No Connection	Error, the can't connect to the server.
No such host	Mistakes not resolve the host. There is probably a problem.
Size	Size of the file pointed to by the link, in bytes.
Level	File level in the server tree.
Error	A message is displayed more descriptive.
Duration	Time to download the file.
Not found	Link is Not Present.
UTF-8 server	Unicode is an industry standard for character encoding of text documents. Unicode has two variants UTF-8 and UTF-16.

RESULT TYPES	URL'S	URL'S PER
Ok	198	84.62
Timeout	33	14.1
Not found	1	0.43



Skip type	2	0.85
Total	234	100

Table in LTTE URL's Result Type.

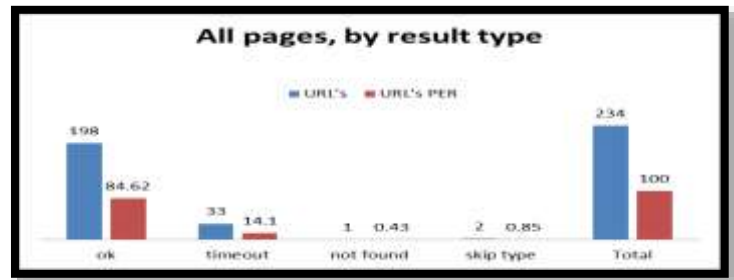


Figure. All Pages by Result Type for Ltte.

URL's Result Type's:

The above table and chart shows the analysis result of the presence of URL link and its Percentage. It is observed that LTTE websites consists of 234 Links, But 198 links analyzed by the used tool value is 84.62% , Timeout Type is internal 33 link analysis for 14.1%, skip type is External links 2 and value 0.85% , analyzed and its 1 link is not present for Not found value 0.43%.

The query results produced by search engines are distributed across different documents that may be connected with hyperlink. How search engine can recognize efficiently such a distributed results

The selected collecting websites in the dark websites which one terrorist organization take in the extract tool hand with specific situation a through correct internal link put the running option after that split levels, bytes, In-Link and Out-Link with status having appear database monitoring Location. In this broken link after click shown in the blow Figure.





Figure. Extract Link.

IV CONCLUSION & FUTURE WORK

As the Web and its usage continues to grow, so grows the opportunity to analyze Web data and extract all manner of useful knowledge from it. The past One year have seen the emergence of Web mining as a rapidly growing area, due to the efforts of the research community as well as various organizations that are practicing it. In this paper we have briefly described the key computer science contributions made by the field, the prominent successful applications, and outlined some promising areas of future research. Our hope is that this overview provides a starting point for fruitful discussion.

As the Web and its usage grows, it will continue to generate evermore content, structure, and usage data, and the value of Web mining will keep increasing. Outlined here are some research directions that must be pursued to ensure that continue to develop Web mining technologies that will enable this value to be realized then Future Problem its dark websites connect the same and different interactivity, ideological and new tactics with make Clustering.

APPENDIX

A variety of dark sites and create can be used to websites in link. In general the choice of the best way depends upon the problem to be solved. The Analysis makes concern that calculates the strongest similarity between the dataset given with the large database needs to work like an Xenu Tool. Hence chose the mining technique with The will be used to form the interactivity of dark web.

REFERENCES

- [1] Grishman R. (1997), "Information Extraction: Techniques and Challenges", International Summer School, SCIE-97.
- [2] Ciaramita, M., Gangemi, A., Ratsch, E., Saric, J., Rojas, I., Unsupervised Learning of Semantic Relations between Concepts of a Molecular Biology Ontology, in: Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI'05), pp. 659-664.
- [3] V. Burris, E. Smith, A. Strahm, "White Supremacist Networks on the Internet," *Sociological Focus*, vol. 33, 2:215-235.
- [4] H. Chen, J. Qin, E. Reid, et al., "Dark Web Portal: Collecting and Analyzing the Presence of Domestic and International Terrorist Groups on the Web," *IEEE Intelligent Transportation System Conference*, Washington, D.C., 2004.
- [5] P. B. Gerstenfeld, D.R. Grant, C. Chiang, "Hate Online: a Content Analysis of Extremist Internet Sites," *Analysis of Social Issues and Public Policy*, vol. 3, 1:29-44.

**A Modern Approach to Designing Implementation
and Reinforcement of Quality Management**

System Organised by

ZES's, Zeal Institute of Management and Computer Application,
Narhe, Pune, Maharashtra, India
on 21st November 2019



- [6] A. T. Gustavson, D.E. Sherkat, "Elucidating the Web of Hate: the Ideological Structuring of Network Ties Among Right Wing Hate Groups on the Internet," Annual Meetings of the American Sociological Association, 2004.
- [7] M. Whine, "Far Right on the Internet," *Governance of Cyberspace*. B. Loader, ed., London: Routledge, 1997
- [8] E. E. Reber, R. L. Michell, and C. J. Carter, "Oxygen absorption in the earth's atmosphere," Aerospace Corp., Los Angeles, CA, Tech. Rep. TR-0200 (4230-46)-3, Nov. 1988.
- [9] J. H. Davis and J. R. Cogdell, "Calibration program for the 16-foot antenna," Elect. Eng. Res. Lab., Univ. Texas, Austin, Tech. Memo. NGL-006-69-3, Nov. 15, 1987.
- [10] WangBin, LiuZhijing, , ' Web Mining Research' , International Conference On Computational Intelligence and Multimedia Applications, IEEE (2003).
- [11] Hiroyuki Kawano, 'Web Archiving Strategies by using Web Mining Techniques', IEEE (2003).
- [12] **Michael Azmy (2012)** Web Content Mining Research: A Survey.
- [13] Kosala, R &Blockeel, H (July 2000)*Web Mining Research: A Survey*. SIGKDD Explorations.
- [14] **Mofreh A. Hogo (2010)** Evaluation of E-Learners Behavior using Different Fuzzy Clustering Models: A Comparative Study.
- [15] Leon Andretti Abdillah(2009),Managing information and knowledge sharing cultures in higher education institutions.
- [16] Raymond Kosala, HendrikBlockeel, "Web Mining Research: A Survey", Newsletter of the SpecialInterest Group (SIG) on Knowledge Discovery & Data Mining, ACM, 2000.
- [17] R. Cooley, B. Mobasher, J. Srivastava, "Web Mining: Information and Pattern Discovery on theWorld Wide Web",Proc. of the ICTAI '97.
- [18] JaideepSrivastava, Robert Cooley, MukundDeshpande, Pnag-Ning Tan, "Web Usage ining:Discovery and Applications of Usage Patterns from Web Data".
- [19] Osmar R. Zaiane, Man Xin, Jiawei Han, "Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs, Advances in Digital Libraries".



- [20] D. Gibson, J. Kleinberg, and P. Raghavan, “Inferring Web Communities from Link Topology,” Proc. 9th ACM Conference on Hypertext and Hypermedia, ACM Press, 1998, pp. 225–234.
- [21] Lizhen Liu, Junjie Chen, Hantao Song, ‘The Research of Web Mining’, Proceedings of the 4th World Congress on Intelligent Control and Automation June 10-14, 2002, Shanghai, P.R.China ,IEEE.
- [22] Chakrabarti, B.Dom, D.Gibson, J. Kleinberg, R. Kumar, P. Raghavan,S. Rajagopalan, and A. Tomkins, Mining the Link Structure of the World Wide Web, IEEE Computer, Vol. 32, pp. 60-67, 1999.
- [23] P Ravi Kumar, and Singh Ashutoshkumar, Web Structure Mining Exploring Hyperlinks and Algorithms for Information Retrieval, American Journal of applied sciences, 7 (6) 840-845 2010.
- [24] Ding, X. He, P. Husbands, H. Zha, and H. Simon, Link analysis: Hubs and authorities on the world. Technical report: 47847,2009.
- [25] Pooja Sharma, PawanBhadana, “Weighted Page Content Rank For rdering Web Search Result”, International Journal of Engineering Science and Technology, Vol 2, 2010.