IOSR Journal of Computer Engineering (IOSRJCE)
ISSN: 2278-0661, ISBN: 2278-8727 Volume 5, Issue 4 (Sep-Oct. 2012), PP 31-36
www.iosrjournals.org

Literature Survey on Web Mining

Geeta R. Bharamagoudar[1], Shashikumar G.Totad[2], Prasad Reddy PVGD[3]
[1](*Associate Professor of Information Technology, GMRIT RAJAM, Andhra Pradesh, India)*
[2](*Professor, Department of Computer Science and Engineering, GMRIT RAJAM, Andhra Pradesh, India)*
[3](*Professor, Department of CS & SE, Andhra University Vizag, India)*

Abstract : *Web Mining is the process of retrieving non-trivial and potentially useful information or patterns from web. Web mining is universal set of Web Structure Mining, Web Usage Mining and Web content Mining. This paper describes and compares these three categories. It also provides comparative statements of various page ranking algorithms with link editing, General Utility Mining and Topological frequency Utility Mining Model by taking constraints such as Web Mining activity, topology, Process, Weighting factor, Time complexity, and Limitations etc.. This also helps in comparing WPs-Tree and WPs-Itree structures.*
Keywords: *Log File, Page Rank, Topology, Web Structure Mining, Web Usage Mining*

I. INTRODUCTION

World Wide Web or Web is the largest and popular source that is easily available, reachable and accessible at low cost, provides quick response to the users and reduces burden on the users of physical movements. The population of web is gigantic. It is made up of billions of interconnected hypertext documents/ web pages which are designed by millions of designers. Ted Nelson conceived the idea of hypertext in 1965. Web supports hypermedia documents. Hypermedia documents are the documents which contain in addition to text, image, audio and video files. The Web page is a page with Hypertext Markup Language (HTML) tags. A web site is a collection of several interrelated or intra related web pages. Web site designer can link one web page with other web page in the web with the help of links called hyperlinks. These links are used to connect one hypertext document with other hypertext document. It resembles a virtual society. It follows Client/Server Model. The client acts as a service consumer. The Server acts as a service provider. The interaction between client and server is using Hypertext Transfer Protocol (HTTP). The client can navigate through the web by means of client program called browser, e.g Internet Explorer, Google Chrome, Netscape Navigator, Mozilla etc. The client will send request to the server through the browser. The request is analyzed by the server and if the requested information is available at the server side, the information is delivered to the client. The request will be in the form of Universal Resource Locator (URL), which specifies resource available on the web universally. Tim Berners-Lee invented the Web in 1989 at CERN in Switzerland. The term World Wide Web is coined by him and the first World Wide Web server, httpd, and the first client program (a browser and editor) "WorldWideWeb" is written by him.

With vast amount information being shared worldwide, there was a requisite to find required information in a systematic and effective way. Search Engines came into existence. Six Stanford students introduced the search system Excite in 1993. MCC Research Consortium at University of Texas established EINet Galaxy in 1994. Yahoo was produced by Jerry Yang and David Filo in 1994 and offered directory search containing favorite websites. In consequent years, many search engines developed, e.g. Lycos, Inforseek, AltaVista, Inktomi, Ask Jeeves, Northernlight. etc.. Sergery Brin and Larry at Stanford University coined Google in 1998. MSN Search engine was tossed by Microsoft in spring 2005.

MIT and CERN took the lead in the formation of W3C (The World Wide Web) consortium in December 1994. The main goal of W3C was to encourage standards for the progression of the Web and allow interoperability between WWW products by producing provisions and reference software. The First International Conference on WWW was held on 1994 [21]. Many trades started on the web and turn out to be more coherent.

Data Mining is also referred as knowledge discovery in databases (KDD). It is a process of discovering useful patterns or knowledge from data sources. It is a multidisciplinary field involving artificial intelligence, statistics, information retrieval, statistics and visualization. Web mining is a process of discovering useful and intelligent information or knowledge from the web site topology, web page content and web usage data.

This paper is organized as follows. Section I describes related work. Section II compares Web Mining Categories i.e Web Structure Mining, Web Content Mining and Web usage Mining. Section III describes Page ranking algorithms along with Link Editing, General Utility Mining and Topological Frequency Utility Mining Model algorithms. Section IV describes WPs-Tree, and WPs-Itree algorithms. Section V draws conclusions and presents future developments of the proposed approach.

II. RELATED WORK

To perform any website evaluation, web visitor's information plays an important role, in order to assist this, many tools are available. Li, L,Zhang and C. And Zhang [1] expressed that Web Mining is a popular technique for analyzing website visitor's behavioral patterns in e-service systems. Jian Pei,J. Han, B.Mortazavi and Hua Zhu [8] found that Web Log Mining helps in extracting interesting and useful patterns from the Log File of the sever. H. Tao Shen, Beng Chin OOi and Kian-Lee Tan [9] suggested that HTML documents contain more number of images on the WWW. Such documents' containing meaningful images ensures a rich source of images cluster for which query can be generated. The documents which are highly needed by users can be placed near to the home page of the website. Manoj Manuja and Deepak Garg [2] suggested that the development of web mining techniques such as web metrics and measurements, web service optimization, process mining etc… will enable the power of WWW to be realized. Jing Wang and others [12] found that weakness of both frequency and utility can be overcome by General Utility Mining Model. Miller & Remington [3] revealed that the structure of linked pages has decisive impact factor on the usability. Geeta and others [4] suggested that the number of pages at a particular level, the number of forward links and the number of backward links to a particular web page reflect the behavior of visitors to a specific page in the website. However Garofalakis [5] pointed out that the number of hit counts calculated from Log File is an unreliable indicator of page popularity. Geeta & others [10] suggested that the topology of the website plays an important role in addition to log file statistics to help users to have quick response. Jia-Ching and others [11] found that Web Usage Mining helps in discovering web navigational patterns mainly to predict navigation and improve website management. Lee and others [6] proved that the web behavioral patterns can be used to improve the design of the website. These patterns also could help in improving the business intelligence.

III. WEB MINING CATEGORIES

Web mining technology provides techniques to extract knowledge from web data. Web Mining is universal set of Web Structure mining, Web Usage mining and Web Content Mining. The various web mining techniques include Classification, Association, Clustering and Personalization etc. These three categories are described as follows.

3.1 WEB STRUCTURE MINING

Web structure Mining concentrates on link structure of the web site. The different web pages are linked in some fashion. The potential correlation among web pages makes the web site design efficient. This process assists in discovering and modeling the link structure of the web site. Generally topology of the web site is used for this purpose. The linking of web pages in the Web site is challenge for Web Structure Mining. The structure of the web page is as shown below. Structure of a web page is shown below

<html>
•
•
link
</html>

3.2 WB USAGE MINING

Web usage mining extracts user's navigation patterns by applying data mining techniques to server logs, together with employing some topology of the web site, Web structure. Web Usage Mining deals with three main steps: Preprocessing, Knowledge discovery and pattern analysis. Web Usage mining has materialized as the fundamental procedure for comprehending more custom-made, user approachable and business-oriented intelligent top web services. Improvements in pre-processing of data, demonstrating, and mining techniques, applied to the web sources, have already lead to many effective applications in competent information systems, tailoring pages to individual users' characteristics or preferences , web smarter analytics tools and procedures for management of content. As the interaction between Users and Web resources exponentially increases, the need for smart web usage analysis tools will also continue to grow. As the complexity of Web applications and user's interaction with these applications increases, the need for intelligent analysis of the web usage data will also continue to grow.

The Log File of server is as shown below

- 127.0.0.1 - - [27/Jan/2003:14:09:25 +0530] "GET /doc/lic-website/lic-home.htm/ HTTP/1.1" 404 1197
- 127.0.0.1 - - [27/Jan/2003:14:09:37 +0530] "GET /doc/lic-website/lic-home.htm/ HTTP/1.1" 404 1197

The fields listed are as follows

- Address
- RFC
- Authuser
- Time Stamp

- HyperText Transfer Protocol Version Number 1.1
- Status code
- Transfer volume

The statistical information collected from Log File is as shown in Table 1.

Table 1 Momentous statistical information found from Web Logs Profile

Statistical Information	Parameters which can be measured
History of the website and users	Number of navigators to the website for a given threshold Number of hits to a particular page Time spent on each page Users visiting a particular web page Average time spent on each page The longest web pages traversal path Association between web pages Redirected or Failed or Successful hits Number of bytes transferred Various Browsers used by clients Usage of GET/POST method
Status Codes which help for troubleshooting	400 Series- Failure Eg:404-File Not Found 300 Series-Redirect 200 Series-Success 500 Series-Server failures
Grade of a Web Page can be measured based on number of hits, time spent, location of web page, in degree and out degree.	Excellent Medium Low

3.3 WEB CONTENT MINING

Web content mining discovers useful knowledge from web page contents. It helps in classifying, clustering or associating web pages according to their topics. It also aids in discovering patterns in web pages to extract useful data such as characteristics of products/items, posting of forums and dependency among the products etc. We can collect the opinions and needs of customer. The content can be improved mainly to satisfy the customer needs.

Table 2 Categories of Web Mining

Parameters	**Categories of Web Mining**		
	Web Structuring Mining	Web Usage Mining	Web Content Mining
Visualization of data	➢ Collection of interconnected web pages ➢ Hyperlink structure	➢ Client/Server Interactions statistics	➢ Unstructured documents ➢ Structured Documents
Source data	➢ Hyperlink structure	➢ Log file of server ➢ Log of Browser	➢ Hypertext Documents ➢ Text Documents
Topology	➢ How all web pages in the website are interlinked together	➢ How well the behavior of users varies for the given web site	➢ How well content of one page is related to content of another page.
Depiction	➢ Web graph	➢ Relational Table ➢ Graph	➢ Relational ➢ Edge labeled graph
Working Model	➢ Page Rank algorithms	➢ Association Rules ➢ Classification ➢ Clustering	➢ Statistical ➢ Machine Learning
Usability	➢ Web Site Personalization ➢ User Modeling ➢ Adaptation and Management	➢ Extracting users behavior ➢ Detecting outliers ➢ Categorization	➢ Relevance of web pages. ➢ Relationship between segments of text paragraphs

IV. PAGERANK ALGORITHMS

The importance of web pages can be determined using hyperlink structure of the web pages on the web. Surgey Brin, Larry Page, C. Ridings and M. Shishigin [13, 14] developed a ranking algorithm, Page Rank, used by Google. Google [15] brings more important documents up in the search results using PageRank algorithm. PageRank is the heart of Google though many other factors are considered [16]. A web page is assigned a high rank if the sum of its backlinks ranks is high. The extension of PageRank algorithm called Weighted pageRank is suggested by Wenpu and Ali Ghorbani [17]. In this algorithm, popularity of a page is measured by its number of inconimng and outgoing links.

Jaroslav Pokorny and Jozef Smizansky [19] contributed a novel ranking method of page importance ranking using WCM technique, called Page Content Rank. The significance of words in the page is one of important parameters to be considered to measure significance of page. The significance of the word is specified with respect to given query.

KleinBerg invented a WSM based algorithm called Hyperlink-Induced Topic Search (HITS) [18]. The set of pages that are relevant and popular are called authority pages for a given query. The set of pages which contain links to relevant pages including links to many authorities are called hub pages.

Page popularity is defined as number of number of hit counts (HC) the particular page. The page popularity can be estimated by counting the accesses to this page based exclusively on a given file [5]. This count may not be a accurate count. Since the page close to home page stands on a path between index page and destiny page, a page close to index will probably have more hit counts. To get accurate results, other factors need to be considered such as depth of the page (d), number of pages at the same level (n), and number of references to a particular page(r). The Relative Access (RA) factor for a web page is defined as shown in Equation (1):

$$RA = a * HC \quad \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

$a = F(d, n, 1/r)$ for all web pages of the web site.

$F = d + n / r$

Utility and frequency plays an important role in General Utility Mining model [20]. General Utility Mining Model is represented as a linear combination of both utility and frequency as shown in Equation (2):

$$GU(I) : \beta\, f(I)/F + (1-\beta)\, u(I)/U \quad \ldots\ldots\ldots\ldots\ldots\ldots (2)$$

Where $f(I)/F$ represents the fraction of frequency of item I out of total frequency, $u(I)/U$ represents ration of utility of item I to total utility, β is the weighting factor of frequency, and $(1-\beta)$ is the weighting factor of utility.

Topological Frequency utility Mining Model considers in addition to frequency and utility, the detailed topology of the website to rank a web page [10]. Table 3 shows comparison of Web page ranking algorithms.

Table 3 Comparison of web page ranking algorithms

System	PageRank	Weighted Page Rank	Page Content Rank	HITS	Link Editing	General Utility Mining	Topological Frequency Utility Mining
Web mining activity	Web structure Mining	Web structure Mining	Web content Mining	Web structure Mining & Web Content Mining	Web structure Mining &Web Usage Mining	Web Usage Mining	Web structure Mining &Web Usage Mining
Rank assigned considering	Pages on the web	Pages on the web	Pages on the web	Pages on the web	Pages in the website	Pages in the website	Pages in the website
Topology	Partial	Partial	Not Considered	Partial	Partial	Not considered	Complete
Process	The page rank for each page is computed during indexing but not during query time. The relevance of a page plays an important role to sort	The distribution of score is unequal among its out links. The computation of scores is done at indexing time.	New grades are computed on the fly for top n pages. Whenever query is posted relevant web pages will be displayed.	Figures out hub and authority grades of top n highly appropriate pages on the fly. Applicable as well as key pages are returned	The grade for each page is computed offline. The pages with high in degree and more time spent are important	The grade of each page is computed considering frequency of each page & the utility of each page.	The grade of each page is computed using Frequency, utility along with topology Parameters.

	the results.			to the users	pages.		
Input Arguments	In links	In links, Out links	Content	In links, Out links and Content	In links, Time spent on each page, depth of a page	Frequency and Utility of each page	In links, out Links, Level, Number of pages in the web site
Weighting factor	Not Considered	Not Considered	Not Considered	Not Considered	Not Considered	Considered only for a specific input	All values Ranging from 0 to 1
Time Complexity	O(log N)	<O(log N)	<O(log N)	O(log N) (higher than WPR)	O(log N)	O(log N)	O(log N)
Limitations	Computation is carried out offline. Scores are computed at indexing time not on fly. The results are displayed based on the importance of pages in the sorted order Page rank is equally distributed to outgoing links.	Less determination of the relevancy of pages to the given query.	References are not considered.	Less efficient	Partial topology	No topology considered	Less efficiency
Result Analysis	Medium	Higher than Page Rank	Approximate or equal to WPR	Less than PR	Equal to PR & less accuracy than TFU	Less Accuracy (than TFU)	High Accuracy

V. COMPARISON BETWEEN WPS-TREE AND WPS-ITREE ALGORITHMS

The Web Pages set-Tree (WPs-Tree) is a prefix-tree which represents relation by means of short and compact structure. Implementation of the WPs-Tree is based on the FP-tree data structure which is very effective in providing a compact representation of relation. The WPs-Itree [7] is an index tree, extension to FP-tree which generates a tree based on the assumed key. It allows access of selected WPs-Tree portion during the extraction task.

VI. CONCLUSION

This paper provides descriptions of various web mining activities. It provides comparison between three categories of web mining. The page ranking algorithms play a major role in making the user search navigation easier in the results of a search engine. The comparison summary of various page rank algorithms is listed in this paper which helps in best utilization web resources by providing required information to the navigators. The associated web pages information can be easily correlated from the users' behaviors. The WPs-Tree and WPs-Itree will help in providing better storage representations. The association between web pages can be found easily in an efficient way. This survey can be helpful for understanding various page ranking algorithms along with different storage representation to correlate web pages. As a future direction, the new metric can be developed which may be still better than this, so that users can have quick response, resources on the network can be used efficiently thus promoting green computing.

REFERENCES

Journal Papers:

[1] Yang, Q. and Zhang, H. , Web-Log Mining for predictive Caching, *IEEE Trans. Knowledge and Data Eng.*, 15(4), 2003,1050-1053.

[2] Manoj Manuja and Deepak Garg, Semantic web mining of Un-structured Data: Challenges and Opportunities, *International Journal of Engineering*, 5(3) ,2011, 268-276.

[3] Miller, C.S. and Remington, R. W. Implications for information Architecture , Human Computer Interaction, *Journal IEEE Web Intelligence*, 2004, 19(3), 225-271.

[4] Geeta.R.B, Shashikumar G. Totad & Prasad Reddy PVGD, Optimizing User's Access To Web Pages, *International refereed Journal JooiJA, Transactions on World Wide Web-Spring*, 2008, 8(1), 61-66.

[5] Garofalakis, Web Site Optimization Using Page Popularity, *IEEE Internet Computing*, 1999, 3(4), 22-29.

[6] Y.S.Lee, S.J Yen, and M.C.Hsiegh. A Lattice-Based Framework for Interactively and Incrementally Mining web traversal patterns, *International Journal of Web Inforrnation Systems*, 2005. 197-207.

[7] Elena baralis, Tania Cerquitelli, and Silvia Chiusano, Imine: Index Support for Item Set Mining, IEEE transactions on Knowledge and Data Engineering, 2009, 493-506.

Proceedings Papers:

[8] Jain Pei, Jiawei Han, Behzad Mortazavi_asl and Hua Zhu, Mining Access Patterns Efficiently from Web Logs, *Pacific-Asia Conference on Knowledge Discovery and Data Mining(PAKDD'00)*, Kyoto, Japan, 2000, 396-407.

[9] Heng Tao Shen, Beng Chin Ooi and Kian_Lee Tan, Giving meanings to WWW, *ACM SIGM Multimedia*, L.A,2000, 39-47.

[10] Geeta.R.B, Shashikumar G. Totad & Prasad Reddy PVGD, Topological Frequency Utility Mining Model *Springer International Conference, SocPros 12*, 2011, 505-508.

[11] Jia-ching Ying , Vincent S. Tseng, Philip S. Yu *IEEE International Conference on Data Mining workshops IEEE Computer Society*, 2009.

[12] Jing Wang, Ying Liu, Lin Zhou, Yong Shi, and Xingquan Zhu, Pushing frequency constraint to utility Mining Model, ICCS *Springer-Verlag Berlin Heidlberg, LNCS 4489*, 2007, 685-692

[13] L.page, S.Brin, R. Motwani, and T. Winograd, The Pagerank Citation Ranking: Bringing order to the Web", *Technical report*, Stanford Digital libraries SIDL-WP-1999-0120, 1999.

[14] C. Ridings and M. Shishigin, Pagerank Uncovered, *Technical report*, 2002

[15] http://www.webrankinfo.comenglish/seo-news/topic-16388.htm, 2006, Increased Google index size.

[16] Neelam Duhan, A.K. Sharma and Komal Kumar Bhatia, Page Ranking Algorithms: A Survey, IEEE International Advance Computing Conference, Patiala, 2009, 1530-1537.

[17] Wenpu Xing and Ali Ghorbani, Weighted PageRank Algorithm, *IEEE Proceedings of the second Annual conference on Communication Networks and Services Research*, 2004

[18] Kleinberg J., Authoritative Sources in a Hyperlinked Environment, Proceedings of the 23rd annual International ACM SIGR conference on Research and Development in Information Retrieval, 1998.

[19] Jaroslav Pokorny, Jozef Smizansky, Page Content Rank: An Approach to the Web Content Mining

[20] Jing Wang, Ying Liu, Yong Shi, and Xingquan Zhu, Pushing Frequency Constraint to Utility Mining Model, *ICCS 2007 Springer-Verlag Berlin Heidelberg, Part III, LNCS 4489*, 2007, 685-692.

Books:

[21] Bing Liu, *Web Data Mining* (New York, Springer International Edition, 2008)