Web Mining Techniques and Issues: A Survey

Priyanka Sharma[1], Prof.(Dr)R.K Gupta[2]
[1, 2] Dept of Computer Science& Engineering
[1, 2] Madhav Institute of Technology & Science, Gwalior

Abstract- *Due to the huge amount of information available on the web, the World Wide Web has becoming one of the most important resources for extracting the information and knowledge discoveries. Many Organizations rely on these websites to attract new customers and retain the existing one. Web log file can be used to analyze customer's surfing pattern. Customer click stream data can be act as a vital source to analyze customer's path to website. Click stream data can be captured and maintain in web log file. As the web is growing fast, the users get easily missing in the web's rich hyper structure. The primary goal of the web site owner is to provide the relevant information and also good quality recommendation based on the web log entries in order to attract customer and increase the sales of company. In this paper we surveyed various Webmining techniques that categorize users and pages by analyzing users' behavior, the content of pages and order of URLs accessed techniques and hence rationalize the recommendations. This paper has attempted to provide research in the rapidly growing area of web mining. We also suggest three web mining categories and also tabulate techniques and other aspect emphasize by various researcher. There is a cause- effect relationship between challenges and opportunities web mining. Better the recommendation (A challenge) result into improved sales of product (An opportunity).*

Keywords- Web mining, Data mining, Content mining, web usage mining.

I. INTRODUCTION

With the large number of companies using the Internet to distribute and collect information, knowledge discovery on the web has become an important research area [20]. With the explosive growth of information sources available on the World Wide Web, it has become necessary for organizations to discover the usage patterns and analyze the discovered patterns to gain an edge over competitors. Jespersen et al [21] proposed a hybrid approach for analyzing the visitor click stream sequences. A combination of hypertext probabilistic grammar and click fact table approach is used to mine Web logs, which could be also used for general sequence mining tasks. Mobasher et al [19] proposed the web personalization system, which consists of offline tasks related to the mining if usage data and online process of automatic Web page customization based on the knowledge discovered. LOGSOM (LOGSOM, a system that utilizes Kohonen's self-organizing map (SOM) to organize web pages into a two-dimensional map) proposed by Smith et al [22], utilizes a self-organizing map based solely on the users' navigation behavior, rather than the content of the web pages. LumberJack proposed by Chi et al [23] builds up user profiles by combining both clustering of user sessions and traditional statistical traffic analysis using k–means algorithm. Joshi et al [24] used relational online analytical processing approach for creating a Web log warehouse using access logs and mined logs. A comprehensive overview of web usage mining research is found in [18,19,20]. The web mining is a result of hybridization of the two areas i.e. data mining and second one is World Wide Web(WWW).It can be able to be mostly defined as the finding and investigation of useful information from WWW. Web mining is the make use of of data mining performance to without human intervention discover and mine information from Web documents and services. The Web mining research is a come together research area from several research community, such as database, IR, and AI research community especially from machine learning and NLP Measuring the interestingness of users and discovered patterns is an active and important area of web mining research. Better quality recommendation system will not only help to satisfy the preference of customers for product but also help to augment the sales of company and chances of user's revisit increase many fold. Poor quality of recommendations suffers from two types of error i.e. False negative and False Positive. False negative: these are the items not recommended even though customer's likes it. False Positive: these are the items recommended to customers even though customers dislike it. In an E-commerce domain goal of company should be minimize these errorsThe World Wide Web (Web) is a popular and attractive medium to distribute information today. The Web is huge, dissimilar, and dynamic and thus raises the scalability, compact disk data, and sequential issues respectively. These factor give rise to the requirement of creating server-side and client-side bright systems that can successfully mine for knowledge. The knowledge comes not only from the content of the pages themselves, but also from the exclusive rareness of the Web, such as its hyperlink organization and its diversity of content and languages. Analysis of these characteristics often reveals interesting patterns and new knowledge. Such knowledge can be used to

improve users' good organization and effectiveness in searching for information on the Web.

II. RELATED WORK

*NeeruMago[1]*focuses to provide an up-to-date survey of the rapidly growing area of Web mining. With the growth of Web-based applications, especially electronic commerce, there is significant interest in analysing Web contents, its structure and usage of data to better understand and apply the knowledge to better serve users.

Garofalakis et al [13] review some data maining techniques and the algorithms for web mining that specifically takes into account the hyperlink information.

Chakarbarti[12] provides a survey of data mining for hypertext. His paper main emphasis on statistical techniques like NPL for web content across supervised, semi supervised and unsupervised learning also on social network analysis techniques for web structure mining

*Mohinder Singh and Navjot Kaur [2]*focus on representation issue on the process and learning algorithm which is based on page ranking method.

*Monika Yadav and Mr. Pradeep Mittal [3]*deals with a preliminary discussion of WEB mining, few key computer science contributions in the field of web mining , the prominent successful applications and outlines some promising areas of future research.

*Ketul B. Patel ,Jignesh A. Chauhan , Jigar D. Patel [4]*discusses web mining in e-commerce, the categories of web mining, pattern discovery techniques to find out interesting patterns, issues of web mining in e-commerce and application of web mining in e-commerce.

*Dr. S. Vijiyarani and Ms. E. Suganya [5]*discussed about the research issues and challenges in web mining and alsoprovided detailed review about the basic concepts of web mining, web content mining, structure mining, usage mining, tools, algorithms and types.

P. Lopes, and B. Roy "Dynamic Recommendation System Using Web Usage Mining", Using traditional web usage mining techniques in an enhanced manner valuable patterns and hidden knowledge can be discovered and focuses on providing real time dynamic recommendation to all the visitors of the website irrespective of been registered and unregistered.

Khushbu Patel [11]works on survey on the existing techniques of web mining and the issues which are related to it and also reports the summary of various techniques of web mining approached from the following angles like Feature Extraction, Transformation and Representation and Data Mining Techniques in various application domains.

III. CATEGORIES OF WEB MINING

Web mining can be categorized into three main areas. (i) Web Content Mining.(ii)Web Structure Mining and (iii) Web Usage Mining.

3.1 Web Content mining:

Web content mining is the mining, extraction and integration of useful data, information and knowledge from Web page content. The heterogeneity and the lack of structure that permits much of the ever-expanding information sources on the World Wide Web, such as hypertext documents, makes automated discovery, organization, and search and indexing tools of the Internet and the World Wide Web such as Lycos, Alta Vista, WebCrawler, ALIWEB ,MetaCrawler, and others provide some comfort to users, but they do not generally provide structural information nor categorize, filter, or interpret documents. In recent years these factors have prompted researchers to develop more intelligent tools for information retrieval, such as intelligent web agents, as well as to extend database and data mining techniques to provide a higher level of organization for semi-structured data available on the web. The agentbased approach to web mining involves the development of sophisticated AI systems that can act autonomously or semi-autonomously on behalf of a particular user, to discover and organize web-based information. It uses two main approaches i.e. Unstructured text mining and Semi structured mining approach.These approaches can be differentiated by two different point of view i.e. Information Retrieval View (IR) and Database View (DB).

3.2 Web structure Mining:

Web structure mining is the process of using graph theory to analyze the node and connection structure of a web site. According to the type of web structural data, web structure mining can be divided into two kinds i.e. Extracting patterns from hyperlinks in the web: a hyperlink is a structural component that connects the web page to a different location andMining the document structure: analysis of the tree-like structure of page structures to describe HTML or XML tag usage [2]. Web structure mining helps the users to retrieve the relevant documents by analyzing the link structure of the Web. The challenge for Web structure mining is to deal with the

structure of the hyperlinks within the Web itself. Link analysis is an old area of research. However, with the growing interest in Web mining, the research of structure analysis had increased and these efforts had resulted in a newly emerging research area called Link Mining [16], which is located at the intersection of the work in link analysis, hypertext and web mining, relational learning and inductive logic programming, and graph mining. There is a potentially wide range of application areas for this new area of research, including Internet. The Web contains a variety of objects with almost no unifying structure, with differences in the authoring style and content much greater than in traditional collections of text documents. The objects in the WWW are web pages, and links are in-, out- and co-citation (two pages that are both linked to by the same page). Attributes include HTML tags, word appearances and anchor texts [8]. Web structure mining could be used to discover authority sites for the subjects (authorities) and overview sites for the subjects that point to many authorities (hubs).Two algorithms that have been proposed to lead with those potential correlations: HITS [14] and Page Rank [15].

A. Page rank algorithm Page rank algorithm [15] is link analysis algorithm [14] that was discovered by Larry page. This algorithm is used by Google internet search engine. In this algorithm numerical weight is assigned to each element of hyperlink set of document such as World Wide Web, with the purpose of measuring the relative importance of that particular set in that hyperlink. Page rank is a probability distribution algorithm used to represent the person's randomly clicking on links will arrive at any particular page. A probability is expressed as a numeric value between 0 and 1. That numerical value is defined as damping factor. It is represented as d and usually its value set to be 0.85. Also C (A) is the number of link going out of that particular page and is known as back link. Page rank of any page is Calculated by:

PR (A) = (1-d) + d (PR (T1)/C (T1) + ... + PR (Tn)/C (Tn)) (1)

Where PR (A) is page rank of particular web page A, d is damping factor.
PR (T1) is page link with main page PR (A)
C is out-linking.

The page rank of a page is divided evenly among it's out-links to distribute to the ranks of the pages they are addressing. Page Rank can be calculated using a simple iterative algorithm, and corresponds to the principal Eigen vector of the normalized link matrix of the web. Page Rank algorithm requires very less time to calculate the rank of millions of pages. Main Limitation of this algorithm is its lesser efficiency since it uses only one parameter i.e. back link.

Hyperlink Induced Topic Search Algorithm (HITS) [14] Hyperlink-Induced Topic Search (HITS) is a link analysis algorithm which helps in rating Web pages also known as Hubs and authorities and is developed by Jon Kleinberg. It was a precursor to Page Rank. The idea behind Hubs and Authorities stemmed from a particular insight into the creation of web pages when the Internet was originally forming; that is, certain web pages, known as hubs, served as large directories that were not actually authoritative in the information that it held, but were used as compilations of a broad catalog of information that led users directly to other authoritative pages. In other words, a good hub represented a page that pointed to many other pages, and a good authority represented a page that was linked by many different hubs It concludes two main values for a page: 1. Page authority, which estimates the value of the content of the page. 2. Page hub value, which estimates the value of its links to other pages.

3.3 Web Usage Mining:

It focuses on techniques that could predict user behavior while the user interacts with the Web. As mentioned before, the mined data in this category are the secondary data on the Web as the result of interactions. These data could range very widely but generally we could classify them into the usage data that reside in the Web clients, proxy servers and servers [14]. The Web usage mining process could be classified into two commonly used approaches [15]. The first approach maps the usage data of the Web server into relational tables before an adapted data mining technique is performed. The second approach uses the log data directly by utilizing special pre-processing techniques. Web usage mining is the application of data mining techniques to discover usage pattern from Web data, in order to understand and better serve the needs of Web-based applications [18]. Web usage mining consists of three phases, namely preprocessing, pattern discovery, and pattern analysis. General Architecture of Web usage mining Process is presented in Figure 1 [18]. Robert Cooley et al. [18] proposes that the web mining process can be divided into two main parts. The first part includes the domain dependent processes of transforming the Web data into suitable transaction form. This includes preprocessing, transaction identification, and data integration components. The second part includes some data mining and pattern matching techniques such as association rule and sequential patterns. In the absence of cookies or dynamically embedded session Ids in the URIs, the combination of IP address can be used as a first pass estimate of unique users. This estimate can be refined using the referrer field as described in [18]. Some

authors have proposed global architectures to handle the web usage mining process. Cooley et al [18] proposed a site information filter, named WebSIFT that establishes a framework for web usage. The WebSIFT performs the mining in distinct tasks. The first state is preprocessing in which user sessions are inferred from log data. The second searches for patterns in the data by making use of standard data mining techniques, such as association rules or mining for sequential patterns. In the third stage an information filter bases on domain knowledge and the web site structures is applied to the mining patterns in search for the interesting patterns. Links between pages and the similarity between contents of pages provide evidence that the pages are related. This information is used to identify interesting patterns, for example, itemsets that contain pages not directly connected are declared interesting. In Cooley et al [18] the authors propose to group the itemsets obtained by the mining stage in cluster of URL references. These clusters are aimed at real time web page personalization. A hypergraph is inferred from the mined itemsets where the nodes correspond to pages and the hyperedges connect pages in aitemset. The weight of a hyperedge is given by the confidence of the rules involved. In Buchner et al. [17] a new approach, in the form of process, is proposed to find marketing intelligence from Internet data. An n-dimensional web log data cube is created to store the collected data. Domain knowledge is incorporated into the data cube in order to reduce the pattern search space. They proposed an algorithm to extract navigation patterns from the data cube. The patterns conform to pre-specified navigation templates whose use enables the analyst to express his knowledge about the field and to guide the mining process. This model does not store the log data in compact form, and that can be major drawback when handling very large daily log files. Information on how customers are using a Web site is critical for marketers of electronic commerce businesses. Data preprocessing consists of data filtering, user identification, session/transaction identification, and topology extraction. Data filtering filters out some noise, i.e., unsuccessful requests, automatically downloaded graphics, or requests from robots, to get more compact training data. Now people use some heuristic rules to identify user, such as IP address, cookies, etc. Preprocessing consists of converting the usage, content, and structure information contained in the various available data sources into the data abstractions necessary for pattern discovery.

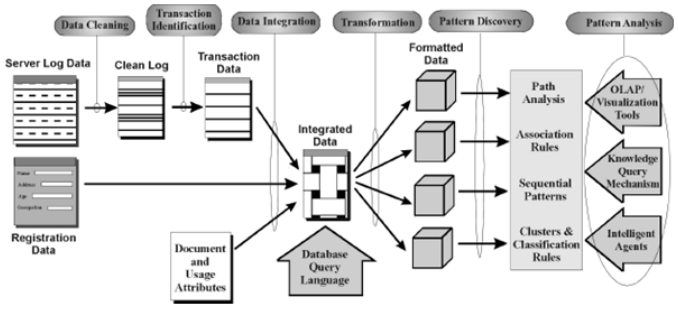

Figure1: General Architecture for web usage mining [3]

On concluding the above discussion we can summarize the following:

Web Content Mining		
1. Web Content Mining	*2. Web structure Mining*	*3. Web Usage Mining*
Text and Multimedia Document	Hyperlink Structure	Web log record.

Table 1: Web content Mining overview

Web Mining				
	Web Content Mining		**Web Structure Mining**	**Web Usage Mining**
	IR View	DB View		
View of Data	Unstructured and Semi-Structured	Semi-structured and website as DB	Link structure	Interactivity
Main Data	Text Document and Hyper text as document	Hypertext document	Link structure	Server log and Browser log
Techniques	Statistical (NLP) and Machine learning	ILP and Modified Association rule	Proprietary algorithm	Machine Learing and Statistical

Table 2: Detailed analysis of Web mining

IV. CONCLUSION

This paper has attempted to provide research in the rapidly growing area of web mining. We also suggest three web mining categories and also tabulate techniques and other aspect emphasize by various researcher. In this paper a general overview of Web usage mining is presented in introduction section. Web usage mining is used in many areas such as e-Business, e-CRM, e-Services, e-Education, e-Newspapers, e-Government, Digital Libraries, advertising, marketing, bioinformatics and so on. The major classes of recommendation services are based on the discovery of navigational patterns of users.With the growth of Web-based applications, specifically e-commerce, there is significant interest in analyzing Web usage data. As the web mining area is growing fast, there is a lot of demand for web usage mining and there is a need to develop a common framework like J2EE and .NET. Cross Industry Standard Process for Data Mining, the CRISP-DM project has developed an industry and tool-neutral Data Mining process model [CRISP-DM] for data

mining. Similar Process model or framework needs to be developed for creating an interest among the new researchers or business strategists and developers. We need a systematic web-site design methodology to create new web pages, or modify existing web pages, such that different user's navigation patterns could be better mapped to answers to a set of specific questions.

REFERENCES

[1] NeeruMago, "Web Mining: Intelligent way of mining Web based data", Apeejay Journal of Computer Science And Applications, Vol. (3), January, 2015.

[2] Mohinder Singh and Navjot Kaur, "A Review on Various Web mining Techniques with Purposed Algorithm of K-means Web Ranking", International Journal of Computer Science and Mobile Computing, IJCSMC, Vol. 2, Issue. 4, April 2013.

[3] Monika Yadav and Mr. Pradeep Mittal, "Web Mining: An Introduction",International Journal of Advanced Research in Computer Science and Software Engineering,Volume 3, Issue 3, March 2013.

[4] Ketul B. Patel ,Jignesh A. Chauhan , Jigar D. Patel "Web Mining in E-Commerce: Pattern Discovery, Issues and Applications" International Journal of P2P Network Trends and Technology- Volume1Issue3- 2011.

[5] Dr.S.Vijiyarani and Ms. E. Suganya, "RESEARCH ISSUES IN WEB MINING", International Journal of Computer-Aided Technologies (IJCAx) Vol.2, No.3, July 2015.

[6] Kun Chang Lee and Sangjae Lee, " Interpreting the web-mining results by cognitive map and association rule approach", Information Processing and Management 47 (2011) .

[7] WenlongRen and Jianzhuo Yan, " An Improved CMAC Neural Network Model for Web Mining", 2015 8th International Symposium on Computational Intelligence and Design.

[8] NyomanKarna, IpingSupriana, NurMaulidevi, "Social CRM using Web Mining for Indonesian Academic Institution", 2015 International Conference on Information Technology Systems and Innovation (ICITSI) Bandung – Bali, November 16 – 19, 2015.

[9] RoyaHassanian-esfahani and Mohammad-javadKargar, "A Survey on Web News Retrieval and Mining", 2016 Second International Conference on Web Research (ICWR).

[10] Dr. Sanjay Kumar Dwivedi and BhupeshRawat, "A Review Paper on Data Preprocessing: A Critical Phase in Web Usage Mining Process", 2015 International Conference on Green Computing and Internet of Things (ICGCIoT).

[11] Khushbu Patel, AnuragPunde, KavitaNamdev, Rudra Gupta and MohitVyas," Detailed study of Web Mining approach- A survey", International Journal of Engineering Science & Research Technology, Patel 4(2) : February 2015

[12] S.chakarbarti ." Data Mining for Hyper text- A tutorial survey", ACM SIGKDD Explorations , 1(2): 1-11-2000.

[13] M.N Garofalakis, R. Rastogi, S. Seshadri and K Shim, " Data Mining and the web: Past, Present and Future", In Workshop of Web Information and Data Management, 1999 pp 43-47, 1999.

[14] Broder, A., R. Kumar, F. Maghoul, P. RaghavanandS. Rajagopalan et al., 2000.Graph structure in the web Computing.

[15] Xing, W. and A. Ghorbani, 2004. Weighted PageRank algorithm. Proceeding of the 2nd Annual Conference on Communication Networks and Services Research, May 19-21, IEEE Computer Society, Washington DC., USA., pp: 305-314. DOI: 10.1109/DNSR.2004.1344743.

[16] Web usage mining by BamshadMobasher. Page No 449-483.

[17] A.G. Büchner, M. Baumgarten, S.S. Anand, M.D. Mulvenna, J. G. Hughes, Navigation Pattern Discovery from Internet Data, in WEBKDD, San Diego, CA 1999.

[18] Robert Cooley, BamshadMobasher, and Jaideep Srivastava, Web Mining: Information and Pattern Discovery on the World Wide Web (A Survey Paper) (1997), in Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97), November 1997.

[19] BamshadMobasher, Robert Cooley, Jaideep Srivastava, Creating Adaptive Web Sites Through Usage-Based Clustering of URLs, in Proceedings of the 1999 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX'99), November 1999.

[20] Soren E. Jespersen, JesperThorhauge, Torben Bach Pederson, A Hybrid Approach to Web Usage Mining,Technical Report 02-5002, Department of Computer Science Aalborg University, July 2002.

[21] Jespersean S.E., Throhauge J., and Bach T., A hybrid approach to Web Usage Mining, Data Warehousing and Knowledge Discovery, (DaWaK'02), LNCS 2454, Springer Verlag Germany, pp73-82, 2002.

[22] Smith K.A. and Ng A., Web page clustering using a self-organizing map of user navigation patterns, Decision Support Systems, Volume 35 , Issue 2 (May 2003) Special issue: Web data mining, Pages: 245 – 256.

[23] Chi E.H., Rosien A. and Heer J., LumberJack:IntelligentDiscovey and Analysis of Web User Traffic Composition. In Proceedings of ACM-

SIGKDD Workshop on Web Mining for Usage Patterns and User Profiles, Canada, ACM press, 2002.

[24] Joshi K. P., Joshi A., Yesha Y., Krishnapuram, R.,Warehousing and Mining We Logs, Proceedings of the 2ndACM CIKM Workshop on Web Information and Data Management, pp. 63-68, 1999.