REVIEW AND ANALYSIS OF TECHNIQUES USED FOR DERIVING ACCESS PATTERNS FROM WEB LOGS

Mr. Rinal H. Doshi[1], Ms. Richa Mehta[2] and Dr. Harshad B. Bhadka[3]

ABSTRACT

The web server – the important component is a huge, rich and dynamic information server. It has rich collection of information regarding hyperlinks and web page access usage. It makes us to gain research attraction particularly for web usage mining. This research review analysis provides research ideas and innovative application leading towards discovery user access pattern and hidden information from datasets. The task of extracting hidden patterns from web log databases can be classified as web usage mining. The data sets which are subjected for the purpose of web usage mining are to be ensured of free from noisy data, irrelevant data, inconsistent data and incomplete data with regard to the purpose of web usage mining. This introduces pre-processing stage just before the web usage mining. This review analysis focused on sequential task of data resource discovery, extraction of purpose oriented relevant fields, pre-processing- data set cleaning, web usage mining for pattern discovery and analysis of extracted patterns. This research review and analysis is focused on data cleaning and web usage mining techniques and tools for enhancing usability services.

1. INTRODUCTION:

The accumulated data at any point of time is so vast in volume and the domain of the data is highly suitable for application of web usage data mining techniques to extract unknown information in the form of web log pattern. Very interesting and predictable analysis can be carried out on these patterns.

This review study and analysis is to start the process of analyzing user web access patterns and discovering datasets from various sources like e-commerce website, UCI repositories. The discovered datasets need to be transform into purpose oriented relevant data followed by removal of inconsistency and filling up the gap with regard to incompleteness of the data. The different datasets obtain from pre-processing tasks are integrated in a purpose oriented data format. These integrated data sets are ready to the subjected for web usage data mining. The data mining techniques like clustering, classification, association mining etc. are to be used for the purpose of web data mining and the extracted pattern are analysed to ascertain the interests of customers with regard to their specific requirement of products and services. This analysis is targeted to improve the structure, service domain and inventory of the products to satisfy the requirement of customers with improve level of satisfaction dynamically. This will provide extended support to improve web application design, decision support

[1] *Research Scholar, School of Engineering,R.K. University, Rajkot, L.C. Institute of Technology, Bhandu, India. rinal.doshi@lcit.org*
[2] *Lecturer, SKPIMCS, Gandhinagar, Gujarat. triumphricha@gmail.com*
[3] *Director- MCA, C.U.Shah College, Surendranagar. harshad.bhadka@yahoo.com*

system design with optimize management decision. [3] [4]

2. STUDY, ANALYSIS AND REVIEW OF LITERATURE:

Authors used web usage mining and statistical analysis for evaluating the usage of website. They used graph mining for complex web browsing behavior of users, for example in parallel browsing user opens several pages in new tabs or windows at the same time, they can simulate navigation path as a graph and apply graph mining to discover web usage pattern. Through statistical analysis they are analyzing the information about page browsing time of website and its user. They considered website as a graph and user navigation path as traversal on it. They can find web usage pattern more accuracy, which helps administrators to evaluate current usage of website. There are three phases for graph mining. Phase 1 is related to monitoring and recording user's web browsing behavior including client side web browsing behavior, phase 2 is related to converting recorded web usage data into graph structure, and phase 3 is related to applying graph mining method to discover web usage patterns. [1]

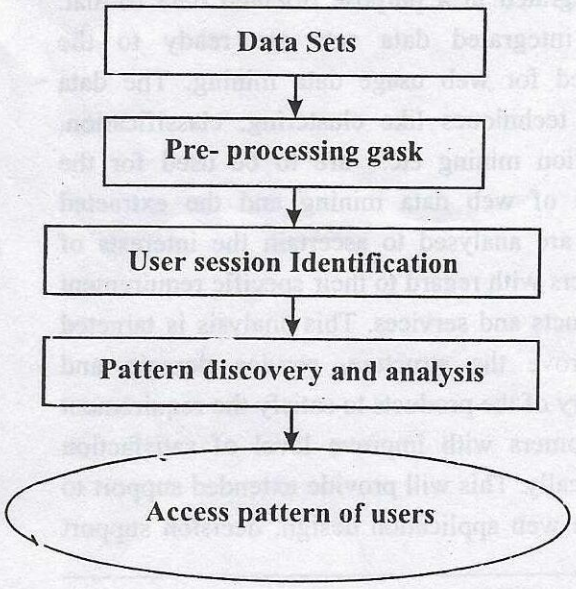


Figure 1: Process of User Access Pattern

Authors tried to give a clear understanding of the data preparation process and pattern discovery process. Figure 1 shows how to find user access pattern from web logs. Website designer can study the visitor's activities through the web analysis and find patterns on the visitors' behaviors. This analysis is obtain by way of application of data mining techniques on the data of web log records to extract predictive patterns by way of transformation and interpretation. Web usage patterns and data mining can be the basis for a great deal of future research. They expect to design a model which will be done the pattern analysis using the artificial neural network. [2]

Most of researcher talk about Server log file. Table 1 gives the details of the Log File Format.

The log file format includes basic things such as:

Table 1: Log File Format for Microsoft IIS Server

The IP address of client (user)
User Name (UN)
Request Date (RD)
Request Time (RT)
Byte Sent by client and server (BS)
Byte Received by client server (BR)
HTTP Status Code (HSC)
Window Status Code (WSC)
Server Name and IP address (SN & IP)
Time taken (TT)

Authors mainly focus on data pre-processing stage of the first phase of web usage mining with activities like field extraction and data cleaning algorithms. Field extraction algorithm performs the process of separating fields from the single line of the log file. Data cleaning algorithm eliminates inconsistent or unnecessary items in the analyzed data. This system removes accesses to irrelevant

items and failed requests in data cleaning. After that necessary items remain for purpose of analysis. Speed up extraction time when users' interested information is retrieved and users' accessed pages is discovered from log data. The information in these records is sufficient to obtain session information. This proposed work can apply data mining techniques for finding user access patterns from pre-processed web logs. [3]

Author mainly works on web mining which applies the data mining, the artificial intelligence, chart technology and traces user's visiting characteristics and extracts the user's pattern. Through the web usage mining, enterprise may discover the influence factor of customer's purchase behavior. It also produces the design mentality of the E-commerce website application algorithm, which is simple, effective and easy to realize and construct a low cost B2C website. The data abstraction required for pattern discovery is obtained by way of pre-processing approach which takes care of converting usage structure and content records of web logs. This research describes the web using pattern algorithm, which is helpful to understand customer's behavior and complete the website design which improve customer's relations, improves the requirement of system performance and improve the website design. [4]

In this research work authors demonstrates study about how to mine the useful information on the web data and give the user specific knowledge and comparison about different data mining techniques. They describe current, past and future of web mining. They introduce retrieval information from online resources on the web and discovery of user access patterns from web servers. They describe web usage mining as an improvement of data mining drawback. This paper describes the web mining sub-task and process. [5]

This research work gives an overview of web mining and web mining techniques. Web contains a dynamic and rich collection of hyperlink information and web page access and the usage of information and provides numeric sources for data mining. Main goal of this work is to discover pattern of access and hidden information from huge collection documents from the web.

Authors focus on introduces the mining tools and application and roles of e-commerce. Through web mining, the future of the company can analyze customer behavior, analyze potential target market and optimize business model e-commerce website and get reliable market feedback and reducing the company's operations cost. [6]

Authors presents the web usage mining system KOINOTITIES, which uses data mining techniques for the construction of user communities on the web which have similar interest and navigational behavior. The information produce by the system can either be used by the administrator, in order to improve the structure of the site. [7]

Here Author combines traditional education with internet technology which produces network education. In this paper FeiQiong Rong tells web usage mining is an important branch of web data mining and its application. Using the web usage mining in network education and can find user interest, the relevance of page, expected location and important page and other useful information, so as to provide a basis for meeting student's individual need in network education. To achieve this server log passed through pre-treating of data which include database access sequence and transactional log. After these works use data mining techniques such association rule, sequence pattern classification and clustering. Lastly obtain

the analysis of pattern which helps to improvement of teaching website. [8]

In this paper Authors presented a clustering web usage data from msnbc.com which is useful in finding the user access pattern and the order of visits of the hyperlinks of the each user. Web data clusters are formed using on MSNBC web navigation dataset. Here authors find interesting user access patterns from web logs using clustering data mining techniques which are the application of web usage mining. [9]

Web data mining is the integration of data mining technology and web. In this paper authors analysing the 763 documents of web data mining published in 1986-2009 in Web of Science database. Lanjie Chen analyze on word frequency, detect key words that their term frequency changed and drawing the knowledge mapping of them by using CiteSpace: information visualization software. This work can draw three conclusions.

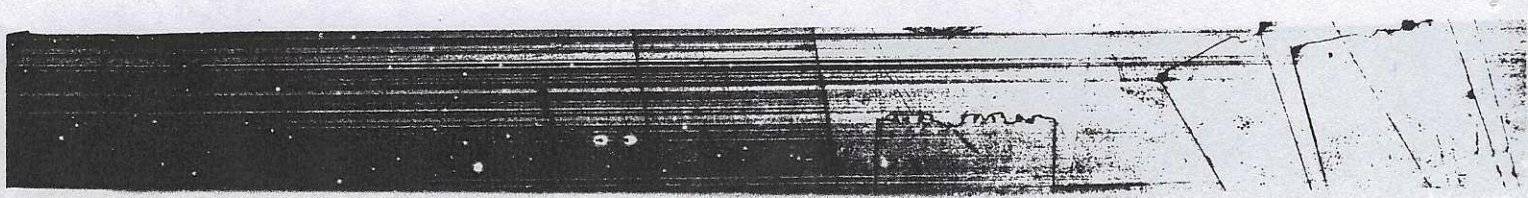
- Data mining, web mining, database, website, knowledge discovery, web usage mining etc. are the hot research topics in the field of international web data mining.
- Dillman, D.A., Agrawal R., Srikant, R.J. Srivastava, Ming-Syan Chen are the important literature authors in the field of WDM (Web Data Mining).
- In recent years, web service, originality value, design methodology approach, data sets, text mining, pattern mining, etc. becoming the research fronts in the field of web data mining. [10]

Authors focused on e-commerce websites how to use web mining technology for providing security on e-commerce website. The connection between web mining, security and e-commerce analyzed based on user behavior on web. This research works use different web mining algorithm and security algorithm which provide security on e-commerce website. Authors use page rank algorithm and trust rank algorithm is used for developing a web mining framework in e-commerce website. In existing web mining framework is based only web content mining. [11]

In this work authors proposed a new personalized recommendation method, which was applied to web log mining by mixing clustering and association mining techniques to improve web personalized recommendation. The work carried research on the most popular method of data pre-processing and applying data mining techniques. The algorithm in this research could satisfy the speed requirement but the page weight should be computed online and its algorithm efficiency to be improved in future. [12]

Authors divided web usage mining process into data preparation phase, pattern discovery and recommendation phase. He shows STPN which is a high level graphical model and used in modeling classifications activities. In this work Ching-Ming Chao presents how to implement a complete online data pre-processing process using STPN. These researches proposed the architecture of online web usage mining system and develop an online web usage mining system that provides web personalized online service. [13]

This research work based on web usage mining. This work is an application of data mining algorithm for different datasets to find user's access navigation pattern. The results of web usage mining have been used to improve website design and web server system performance. Here an improved Ward's method is proposed for web user clustering. Firstly the row web logs collected on web server are cleaned to obtain access sequence



database. Secondly the distance between two sequences is calculated and reduce distance matrix is built up. Thirdly the ward's method is applying on the matrix and obtains clusters. [14]

3. REVIEW FINDINGS AND ANALYSIS

The reviewed research literature has lead to summarized that the different research undertaken have accessed data from various data repositories available on web servers, web logs with regard to access data, usage data, content data, structure data etc. This data may include wide variety of data components of research interest with regard to user identification, session identification, transaction identification, session start, page load, page request, session end, security data etc. The various data sets are in different formats and are to be integrated or to be clustered for the purpose of web data mining after subjecting datasets through versatile process of pre-processing. The researchers have used field extraction algorithms, data preparation algorithm and data cleaning algorithms. The data sets are taken of to clustering techniques- k-means clustering, hierarchical clustering and fuzzy c-means techniques to establish remaining clusters. The researcher has also used various tools of data collection, data cleaning and pre-processing, data clustering, applying data mining and statistical analysis techniques to the integrated / cluster data to extract meaningful patterns. The tools are used as well as in wide use are generalized tool – Python, R, Pycogent, Carrot2, ELKI, KNIME, Orange, WEKA, RapidMiner, etc and specialized tools like KOINOTITIES. These research works outcome are applied to various utility application like personalized network education providing educational resources, self learning e-materials, ensuring of quality of learning through evaluation and application development skill used simulation based virtual learning environment. All these services to be made available to a learner after mining users interest pattens, access patterns, requirement patterns etc.

4. PROPOSED WORK

The proposed work is classified into three categories:

Table 2: Categories of Web Mining

Web Content Mining	Mining, retrieving and integration of information from web contents.
Web Structure Mining	The process of discovering information from the web.
Web Usage Mining	It is also known as web log mining.

The suggested work focused in the category of web usage mining of accurate user session data sets keeping in mind the predicted processing time overheads of monitoring and control by firewall and antivirus active tools of the system to extract more accurate patterns. The targeted work is expected to be useful in application areas of bioinformatics, cheminformatics, and medicalinformatics, web intelligence, intelligent adaptive e- learning etc.

5. ACKNOWLEDGMENT:

The Authors are thankful to DR. N.N.Jani, who have guided to explore the area of for the proposed work and also to Dr. C.D. SANKHAWARA, who encouraged and motivated. Sincere thanks to Mrs. Bhumi R. DOSHI, who extended help to review the paper for language correction and compilation.

6. CONCLUSION:

This work analyses the usage patterns of web sites in order to get an improved thoughtful of the users' comfort, demand, need and requirements and improve the customer satisfaction. Proposed works will suppose to use experimental research, meta-analysis and decision oriented research of datasets using data mining tools and techniques which improve the user access pattern and usability service.

Datasets will be taken from different sources like e- commerce website, and free sample log from weka or uci repository datasets.

First it will pre-process log file and then suppose to use data mining techniques using data mining tools and compare them with different techniques.

This proposed future work will suppose to use association mining, classification, clustering and frequent pattern mining as sample techniques.

This future research needs one of the task and method under the banner of tools like weka, KOINOTITIES, Google Analytics, Python, and AlterWind Log Analyzer Lite.

7. REFERENCES:

[1.] Mehdi Heydari, Raed Ali Helal, Khairil Imran Ghauth, A Graph-Based Web Usage Mining Method Considering Client Side Data, IEEE 2009 International Conference on Electrical Engineering and Informatics 5-7 August 2009, Selangor, Malaysia

[2.] Shahnaz Parvin Nina, Md. Mahamudur Rahaman, Md. Khairul Islam Bhuiyan, Khandakar Entenam Unayes Ahmed, Pattern Discovery of Web Usage Mining, 2009 International Conference on Computer Technology and Development, IEEE Computer society.

[3.] Theint Aye, Web log cleaning for mining of web usage patterns, IEEE 2011

[4.] Shuyan Bai, Qingtian Han, Qiming Liu, Xiaoyan Gao, Research of an algorithm based on web usage mining, IEEE 2009

[5.] Kavita Sharma, Gulshan Shrivastava, Vikash kumar, Web Mining: Today and Tomorrow, IEEE 2011.

[6.] Zhang Haiyang, The Research of web mining in E-commerce, IEEE 2011.

[7.] Dimitrios Pierrakos, Georgios Paliouras, Christos Papatheodorou, Constantine D. Spyropoulos, "Koinotities: A Web Usage Mining Tool for personalization."

[8.] FeiQiong Rong, The Application of Web Usage Mining in Personalized Network Education, IEEE 2011.

[9.] K. Suresh, R. Madana Mohana, A. Rama Mohan Reddy, A. Subrmanyam, Improved FCM Algorithm for Clustering on Web Usage Mining, IEEE 2011

[10.] Lanjie Chen, Li Wei, The Hot Research Topics and the Research Fronts in the Field of Web Data Mining (WDM) Based on Web of Science, The 5th international conference on computer science & education Hefei, China. August 24-27 2010, IEEE 2010

[11.] R. Manjusha, Dr. R. Ramachandran, Web Mining Framework for Security in E-commerce, ICRTIT 2011, MIT, Anna University, Chennai, IEEE 2011

[12.] Yi Dong, Huiying Zhang, Linnan Jiao, Research on Application of User Navigation Pattern Mining Recommendation, Proceeding of the 6th world congress on intelligent control and automation, June 21-23 2006, Dalian, China, IEEE 2006

[13.] Ching-Ming Chao, Shih-Yang Yaang, Po-Zung Chen, Chu-Hao Sun, "An Online Web Usage Mining System Using Stochastic Timed Petri Nets", Fourth international conference on Ubi-media computing, IEEE 2011

[14.] Peiqian Liu, Wei Li, Navigation Pattern Discovery on Web Site Based on the distance between Sequences, IEEE 2011.