A Survey Report On Current Research and Development of Data Processing In Web Usage Data Mining

Nandita Agrawal[1], Anand Jawdekar[2]
[1, 2] Department of CSE
[1, 2] SRCEM, Banmore, M.P., India

Abstract- *WUM is the part of web mining that identifies usages data from the web log server, in order to known and improved serve the requirements of the web applications. The WUM involves three types, Data Preprocessing, Pattern Discovery & Pattern Analysis. Pattern discovery phase contains many web data mining methods are apply to process the data so as to discover patterns. Once the pattern discovered. Analysis is done using various operations with unique session and unique user. This paper gives a brief overview of WUM and its phases.*

Keywords- Web mining, web usage, web log, data pre-processing, discovery phase and analysis phase.

I. INTRODUCTION

Website is a webpage collection. Several scholars have been carried out for the web improvement. Pattern analysis and discovery are the methods valuable for personalization of web. Web marketing becomes more popular in current for accurate result many visitor use the specific website Web mining uses novel way of web marketing. Web mining has the potential to discover the specific customers. Web log will not provide a clear idea about the website user. It shows details about visitors. Browsers, web servers and proxy server are the web mining keys. Cache and Cookies of web used to accessing website speed improve but it delays the procedure of visitor's identification. When data will take time to reach the web server at user side; it may be invalid while reaching the web server. Storage is the difficulty in the web mining procedure. Web developers hire another party to store their web log data in the cloud this is a possible in case of web log manipulating.

The World Wide Web continues to grow both in huge volume of traffic and the complexity and size of Web sites. It is difficult to identify the relevant information present in the web. Most of the contents in the web are unstructured in nature, but very little work deals with unstructured and heterogeneous information on the Web. Aims of web mining are to finding and extracting relevant information that is hidden in Web related data, in particular in text documents published on Web . Data Mining involves the concept of extraction meaningful and important information from the database. Web mining is an important part in data mining where we extract the interesting patterns from the contents. Generally three kinds of information are handled in web site namely Content, Structure, Log data. On bases of this knowledge the Web Mining consists of 3 processes, Content Mining, Web Structure Mining, Usage Mining [1] as shown in fig. content mining which is deals with the raw data that is available on the web. Structure mining is the second phase of web mining that is mainly deals with web structure sites[2].Usage mining is the last phase of web mining this involves mining that usage characteristics of the users of Web applications. It is in a less-structured format so that it need a lots of pre-processing And parsing earlier than the specific extraction of the desired information. This paper gives the survey of techniques of WUM. Data mining consist of several stages namely[3] Domain Understanding, Data selection, Data pre-processing and cleaning, Pattern discovery, Interpretation and Reporting.

Web Mining Can be generally classified into three different categories, in step with the kinds of data to be mined. The brief overview of the three important categories in below-

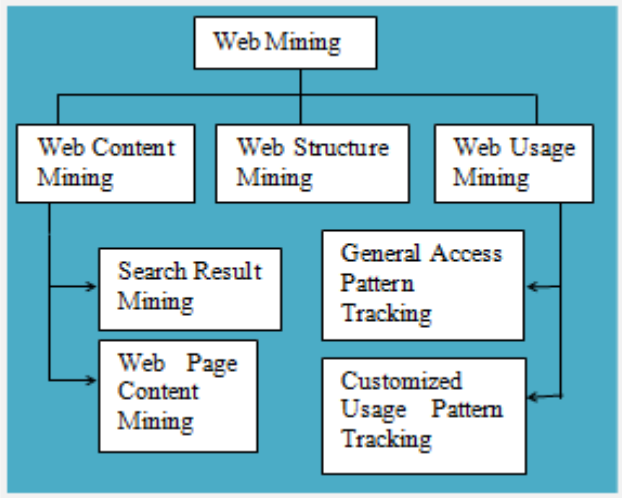

Fig. 1 Types of web mining.

A. Web Content Mining

It is the procedure of removing valuable knowledge from Web documents contents. Content data corresponds to the set of facts a Web page was aimed to users convey. It may contain of video, text, images, audio, or structured records for example tables and lists. Text mining to Web content has been most extensively researched. Issues text mining addressed are, extracting association patterns, topic discovery, web documents clustering and Web Pages classification. Research activities on this topic have drawn greatly on methods developed in other different disciplines for example IR (information retrieval) and NLP.

B. Web Structure Mining

In which, classical web graph Web pages contains like nodes, and also hyperlinks and edges connecting connected pages. It is the discovering structure information procedure from the Web. This can be further classified into two types.

Hyperlinks: It is a structural unit that connects a Web page location to various locations, either within the same Web page or on a various webpage. A hyperlink that connects to a same page part is known as Hyperlink Of Intra-record, and the hyperlink that connected two pages is known as an Inter-Document Hyperlink. These perform important work body on hyperlink analysis, of which Desikan et al. give an up-to-date survey.

Document Structure: In which, content are present in the format of tree-structured, based on numerous XML and HTML tags in page. Mining technique focused on automatically mining DOM structures out of documents (Wang and Liu 1998, Lim, Moh, and Ng 2000).

C. Web Usage Mining

WUM is the techniques of data mining to observe exciting usage patterns from web server log, within the order to comprehend and improved serve the Web-based application needs. Usage data capture the identity or Web users origin along with their performance of browsing at the web site.WUM itself can be categorized further depending on the type of useful data these are:

Web Server Data: in which, user collected logs through Web server. Classic data conclude Page reference, IP address and also time of access.

Application Server Data: Business utility servers for example web common sense story Server have important points to permit functions of E-commerce to be constructed on prime of them with the aid of little effort. A key function is the track potential numerous trade occasion types and log data them.

Application Level Data: Latest pursuit's forms may also be outlined in an application, and logging can be became on for them - generating histories of those in specific outlined events [4].

WUM is the data mining method application to the observe utilization data from the web server, with the intention to comprehend and improved serve the web-based Applications needs [5]. WUM carries out exciting and important knowledge for an assortment of people based on the various domains work. The WUM results can be used in system enhancement, personalization, business intelligence, site modification, usage characterization and so forth.

Usually, WUM consists of three different procedures: data pre-preprocessing, discovery of patterns and analysis of patterns. The patterns discovery data sources, the outcomes' data preprocessing influences quality the patterns discovery results. It improved pattern resource but any discover top quality data but also better the algorithm of WUM. So, data preprocessing is mainly significant for the whole WUM procedures and the key of the WUM's quality [6]. However, all kind of data set used in data preprocessing varies not only in the location terms of the data source, but also the data available type, the population segment from which the data are collected, and it's implementation method.

II. EXISTING TECHNIQUE

A. Data Preprocessing in WUM

In which, the input data comes from various session file that provides an exact account of who Web site accessed, who send the request, in what way, also how long all page was seen and in what order, and also User session is the page accesses set that happen at the time of a single Web site visit. However, reasons because we will discuss in the following, the knowledge contained in a raw data, it does not reliable for a user, session file before data preprocessing. Usually, data preprocessing incorporate of direction completion, , cleaning of data and identification of user & session.

a) Data Cleaning

It is to get rid of the redundant and irrelevant log entries for mining system. There are three distinct types of

redundant or irrelevant information needed to the smooth: accessorial resources set in file of the HTML, robots'& error request.

Accessorial Resources: when your HTTP protocol has been connectionless. A consumer's request to view a specific page often outcomes in various log entries since scripts and graphics are down-loaded in the HTML file addition. Since the basic intent of WUM is to get a user's behavior picture, it does not create sense to contain file requests that the user did not request explicitly. Items elimination deemed unrelated can be reasonably accomplished through checking the URL name suffix.

Robots' requests: Robots (also referred to as spiders) are instruments of software that scan a website online to mine its content material. Spiders mechanically comply with the entire hyperlinks from web log page. Search engine is use spider for graph the every pages from online web site and modified their search indexes [7]. To do away with robots' request, we are able to seen for each host which have requested in, "robots.Txt".

Error's request: These are main for mining procedure. They may be able to be eliminating via checking the request fame. Comparable to, if the status is 404, it is present that requested useful resource which have no existence. These data entry in log will also be get rid of then.

b) User Identification

User identification is importantly .complicated through local caches existence, firewalls of corporate, and proxy servers. The WUM approaches that user cooperation rely are the easiest ways to the deal with this issue. However, it's challenging because of privacy and security. In our experiment, we use many steps to user identify:

1) All user have unique Ip Address;
2) For many logs, if the IP address of user's are the same, but agent log present a modification in browser operating system and software, An IP address with represents a various user ;
3) Making use of the access log along referrer logs and site topology to assemble browsing method for all user. If a web page sends a request that isn't the directly reachable via a hyperlink from any of the pages users visited by way of, there may be yet another one-of-a-kind consumer with the equal IP.

c) User Session Identification

A person session method a delimited collection of the person clicks throughout a number of web servers. The session identification aim is to divide the page accesses of all user into periods of person. At gift, the techniques to establish person session conclude timeout mechanism [8] and likewise maximal forward reference [9] . The following rules that is use in our project:

1) In a session of user, if the refer page is null, there is a novel session;
2) If there is a novel user, there is a novel ;
3) If time b/w the page requests then it exceeds a numerous limit(30 or 25.5mintes), it is supposed that the user is starting a novel .

d) Path Completion

It is the last phase or data-preprocessing. In this, access the important page records that are missing in access log over the browser and proxy server and the local cache. The path completion task is to complete with these missing page references.

Techniques similar to those used for identification of user can be used for completion path. If a page send a request which is not directly connected to the last page a requested of user firstly referrer log check to present page request where it come from. If page is in user's present request, then the assumption is that consumer Backtracked with the "again" button presented on finest browsers and calling up cached varieties pages unless a novel page used to be requested. If the referrer log is just not clear, website topology can be used to the similar outcomes. If more than one web pages in the customer's historical past include a link to requested page, it's assumed that page close to the earlier requested page is the source of the novel request.

The path completion is complete on either dynamic pages or static pages. It is easy to work on static pages while for Web sites designed applying the CMS concept don't contain a unique page name for all page, pages instead contain id through which the pages content can be retrieved. So performing way completion becomes complex and difficult. In the preprocessing stage, after identify sessions, while attempting for way completion, we build the page name through page reading page name and id from the xml files for example site map and RSS feed, and at the time of completion path, we add missing pages which is missing in the session, Eliminate pages which is duplicate in the consecutive access within a provide session and Map the pages name with the page number which present in this session. Apart from this, we add the idea of the event as per academic calendar in context with the University atmosphere. Adding events permit

to mine the web logs based on temporal idea, as the accesses to web through the users are not same each time. Based on event lots of patterns can be discovered and also analyzed.

B. Pattern Discovery in web usage mining

The Pattern Discovery stage purpose is to create knowledge full data stored patterns after data reforming and cleaning. Pattern discovery is performed only after data cleaning and after the user transactions identification and also sessions from access logs. The pre-processed data analysis is most valuable to all the organizations performing various businesses over the web [10].

This is the Web mining key component. Pattern discovery converges the algorithms and methods from various research areas, for example recognition, statistics, data mining, and pattern machine learning. According to the methods adopted in this area, I will introduce this procedure in the separate subsections as follows.

a) Statistical Analysis

Statistical methods are the most powerful tools in mining knowledge about Web site visitors. Through analyzing the statistical knowledge includes in the periodic report of Web system, the mined report can be potentially valuable for improving the system performance, system security enhancing, and facilitation the site modification venture, and supplying aid for the marketing decisions [25].

b) Association Rules

This rule associates one items to another items in case of visit the URL which is requested by the user. Association mining systems can be used to detect unordered correlations between found items in a transactions database [24].

c) Clustering

Analysis of clustering is a method to group together data items (pages) and users with the similar characteristics. user information clustering or pages can facilitate the execution and development of future marketing approaches [24].

d) Sequential Pattern

This method follows in session, such that a items collection follows the occurrence of another in a time-ordered collection of episodes or sessions. It's most important for the Web marketer to predict the future trend, which help to place advertisements aimed at various user groups. Sequential patterns also conclude few different kind of temporal analysis for example trend analysis, detection of change point, or similarity analysis [25].

C. Pattern Analysis in web usage mining

Analysis of Pattern is a last phase of the complete WUM. The aim of this procedure is to eliminate the irrelative patterns or rules and to remove the patterns or interesting rules from the pattern discovery procedure output. The Web mining algorithm's output is often not in the form appropriate for direct human consumption, and thus necessity to be transform to a format can be assimilate simply. This can be complete with the help of few analysis methodologies and also tools. There are two common methods For the patter evaluation. One is to use knowledge mechanism of question for example SQL, even as another different is to assemble multi-dimensional information dice before achieve operations of OLAP [15]. Each of these approaches assumes the output of the earlier stage has been structured. There are more methods coming out in present years, for example visualization etc.

This is also a fertilized area of research. Although there are quite a some commercial analysis applications presented and numerous more are free on the Web, most of them are dislike by users, considered too slow, inflexible, challenging to maintain and limited in the functionality. To develop the efficient, flexible, and powerful tools, lots of work needs to be done for both developer and researcher.

III. LITERATURE SURVEY

Nasraoui and Krishnapuram et al. [16] discovered documents of user session and formulated collection on basis of same characteristics applying fuzzy algorithms. According to the research a page contain more than one cluster. After the usage data distinct matrix preprocessing of preprocessed information is formed. This is used through fuzzy logic algorithms in order to the cluster of user session.

Mobasher et. al. [17] , most advanced method, "Web Personalize". It is the most powerful structure mining of web log files to the eliminate the valuable knowledge for the recommendations purpose based on browsing similarities of present user to few user. After cleaning and collective of the data (making several abstractions of collected data), mining methods for Example organization rule, clustering, sequential sample discovery, and classification are applied to be able to detect important usage patterns.

Probably the largest contribution of Berendt [18] in the WUM area is STRATDYN add-on module. It defines the various between navigational patterns and then it's develops the site semantics in the visualization of the outcomes. In this method, web pages are grouped together on the concept hierarchies' basis. He focused on "interval based the coarsening" method for usage data at abstraction various stages. For this purpose he used Common and coarsened stratograms for visualization of the outcomes. The main drawback of this technique is that the web analytic service is not up to the mark.

Magdalini Eirinaki et. al. [19] focused on WUM. This procedure relies on the statistical application of data mining approaches to the web log data, resulting in a useful pattern set that indicate the user's navigation behavior. In this method numerous algorithms of data mining are applied . All above mostly focused on web usage mining and user profiling for personalized recommendations applying search queries and their method is to provide common personalized atmosphere for all web users' type.

The referrer-based technique and time-oriented techniques are collective to accomplish identification session file [20]. Web access log enters the all records set in the web log and saved according to time sequence. A User Session Set is found from Web Log Set through following rules for example various users are distinguished with several IP address. If the addresses of IP are same, the several browsers or OS indicate several users and when the IP addresses are same, the several browsers and OS are same, the referrer knowledge is taken from account. The URL is checked and a novel user session is perceived if the URL is not found, or there is a substantial interim (over 10 seconds [21]) between the access time of this record and the earlier one if the current URL field is empty. If the sessions identified in the previous phase include more than one visit through the same user at quite a lot of time, the time-oriented heuristics is then used to divide the various visits into various user sessions. In this method, drawback was that the complexity is high while this increased the accuracy.

Baoyao Zhou [22]. An access session is generated as a URL pair and the requested time in a requests sequence with a timestamp. The URL duration is estimated as the alteration of solicitation time of successor section and present session. If last URL does not have any successor .So the at the time of is estimated as the average duration of the present entry. The end time of the session is the start time and duration. This algorithm is suitable for more number of URL's in a session. The time set by author is 30 minutes per session. In this method, drawback was that the time is very much in a session.

Another technique applying Integer Programming was proposed through Robert F.Dell [23]. The benefit of this technique is construction of each sessions simultaneously. He suggests that all web log is considered as a registers from the same IP location and operators and additionally connected are gathered to shape a session.

Another calculation proposed by Junjie Chen and Wei Liu in which session recognizable proof and information cleaning is combined [24] In this deleting the content foreign to algorithms of mining gathered from web logs. Session record is searched if no session exists, a novel session is established. In the event that the present session closes or surpasses the preset time limit, the example will end it and founds a novel one. Graph mining approaches, builds precise sessions and the time taken is additionally nearly less.

G. Arumugam and S. Suguna [25] to create accurate way sequences through applying two different way hashed structure based access history collection to frame a complete way with optimal time. There are two challenges in this pursuit as in reverse reference expends additional time in unused pages likewise and pages which are straightforwardly alluded from other distinctive server's prompts wrong session recognizable proof. To succeed these issue's authors provide various calculation. In this Session Identification calculation information structures for instance Array List to tell to Web Log data and User Access List, a Hash table to server pages, in which two-way hashed structure are used. Two ways hashed structure is used to store user accessed page sequence. Two different hash tables secondary and primary hash tables are utilized as a part of which essential is utilized to store sessions and pointers to auxiliary table which is containing a full path navigation. To solve the time utilization just visited pages are put away in access history rundown and unused is not considered. Utilizing a solitary pursuit as a part of history list, the page arrangements are straightforwardly found. At the point when pages are alluded from different servers straightforwardly begin from the page and not from root. On the off chance that the page is not accessible in present sessions, begin another session and we can derive this is not a retrogressive reference but rather the page is skimmed in another server. This technique produces correct path than maximal forward reference length strategies.

IV. PROBLEM STATEMENT

A The work till now is done on static web pages, that is why the problem arises in dynamic web pages.

B Another problem is the differentiation of the grouping done in the pattern analysis phase i.e. we have not differentiated the users according to the type of users.

C The users have not been distinguished based on various factors till now which leads to a consistent growth only. Thus, a new approach can be suggested on this.

V. PROPOSED IDEA

Web mining is the web marketing tool to the provide customers to website. Web log is the source for the WUM to produce the visitor's pattern. Pattern evaluation and generation is the discovery web mining stage. Data preprocess and transformations are the significant web mining stage. This paper is explaining the whole idea of the web mining and the web usage mining. Data pre-processing is the main and first web usage mining stage. Data transformation is used to protect the user data from malicious activity. We have discussed various data preprocessing method on the basis of various factors like data, input, output, memory required, working concept, complexity etc.

We will try to solve the problems discussed above in an innovative way by applying less complex approach.

VI. EXPECTED RESULTS

A The proposed idea will try to reduce the time complexity for the whole process.

B The pattern analysis phase will be different which will give a new approach for analyzing the data on various factors.

C Dynamic web pages will be taken so that the process can be applied on the web log of dynamic websites.

REFERENCES

[1] Chidansh Amitkumar Bhatt · Mohan S. Kankanhalli, "Multimedia Data Mining: State Of The Art And Challenges" Published Online: 16 November 2010© Springer Science+Business Media, LLC 2010.

[2] Rajni Pamnani, Pramila Chawan 1 Qingtian Han, Xiaoyan Gao, "Web Usage Mining: A Research Area In Web Mining".

[3] Wenguo Wu, "Study On Web Mining Algorithm Based On Usage Mining", Computer- Aided Industrial Design And Conceptual Design, 2008. CAID/CD 2008. 9th International Conference On 22-25 Nov.2008.

[4] Web Usage mining for a Better Web-Based Learning Environment Osmar R. Department of Computing Science University of Alberta Edmonton, Alberta, Canada email: zaianecs.ualberta.ca.

[5] Shahabi C., Kashani F.B.. A Framework for Efficient and Anonymous Web Usage Mining Based onClient-Side Tracking. Proc. WEBKDD 2001: Mining Web Log Data across All Customer Touch Points, LNCS 2356, Springer-Verlag, 2002: 113-144.

[6] Pirolli P., Pitkow J., Rao R.. Silk from a sow's ear: Extracting usable structures from the Web. In:Proc. 1996 Conference on Human Factors in Computing Systems (CHI-96), Vancouver, BritishColumbia, Canada, 1996.

[7] Trousse B. ,Tanasa D.. Advanced data preprocessing for intersites Web usage mining. Intelligent Systems, IEEE,2004(19): 59 – 65.

[8] Pitkow J.,Catledge L. . Characterizing browsing behaviors on the World Wide Web, Computer Networks and ISDN Systems ,1995,27(6):1065-1073.

[9] Park J.S., Chen M.S., Yu P.S.. Data mining for path traversal patterns in a web environment. InProceedings of the 16th International Conference on Distributed Computing Systems, 1996:385-392.

[10] Mirjana Maric, Zita Bosnjak, Sasa Bosnjak, "The role of web usage mining in web application evaluation", Management Information Systems, Vol. 05(01):31-36, 2010.

[11] Bamshad Mobasher, Robert Cooley, Jaideep Srivastava, -Creating Adaptive Web Sites through Usage-Based Clustering of URLs‖. 1999.

[12] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-Ning Tan, ―Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data‖. 1999.

[13] Mike Perkowitz, Oren Etzioni, ―Adaptive Web Sites: Automatically Synthesizing Web Pages‖.1998.

[14] Tak Woon Yan, Matthew Jacobsen, Hector Garcia-Molina, Umeshwar Dayal, ―From User Access Patterns to Dynamic Hypertext Linking‖. 1996.

[15] Magdalini Eirinaki and Michalis Vazirgiannis, "Web mining for web personalization", ACM Transactions on Internet Technology, 03(01):1-27, February 2003.

[16] Nasraoui O., Frigui H., Krishnapuram R., and Joshi A, "Extracting web user profiles using relational competitive fuzzy clustering", IJAI Knowledge Discovery, 09(04):8-14, April 2000.

[17] Mobasher B., Cooley R., and Srivastava J, "Automatic personalization based on web usage mining", ACM Communication,43(08):142-151, August 2000.

[18] Berendt B, "Understanding web usage at different levels of abstraction: Coarsening and visualizing sequences", ACMSIGKDD Knowledge discovery & Data mining, 04(07):104-108, August 2001.

[19] Jose M. Domenech1 and Javier Lorenzo, "A Tool for Web Usage Mining , " , 8th International Conference on Intelligent Data Engineering and Automated Learning, 2007.

[20] Jia Hu and Ning Zhong "Clickstream Log Acquisition with Web Farming,,", Proceedings of the International Conference on Web Intelligence, IEEE,2005.

[21] Baoyao Zhou, Siu Cheung Hui and Alvis C.M.Fong,"An Effective Approach for Periodic Web Personalization,", Proceedings of the IEEE/ACM International Conference on Web Intelligence. IEEE,2006.

[22] Robert F.Dell ,Pablo E.Roman, and Juan D.Velasquez, "Web User Session Reconstruction Using Integer Programming," , IEEE/ACM International Conference on Web Intelligence and Intelligent Agent,2008.

[23] Arumugam G. and Suguna S,"Optimal Algorithms for Generation of User Session Sequences Using Server Side Web User Logs, ",ESRGroups, France , 2009.

[24] R. Cooley, B. Mobasher, and J. Srivastava. Data preparation for mining world wide Web browsing patterns. Knowledge and Information Systems, 1(1), 1999

[25] R. Cooley. Web Usage Mining: Discovery and Application of Interesting Patterns from Web data. PhD thesis, Dept. of Computer Science, University of Minnesota, May 2000.