Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.017

IJCSMC, Vol. 6, Issue. 5, May 2017, pg.148 – 153

REVIEW PAPER ON TO PERFORM PREDICATIONS USING DATA MINING & SESSION IDENTIFICATION

Mr. Pratik K.Hutke[1], Mr. A.P.Rudey[2]
[1]Student, ME(CSE), Dr.Sau.Kamaltai Gawai Institute of Engineering & Technology, Darapur
[2]Professor, Dept.of CSE, Dr.Sau.Kamaltai Gawai Institute of Engineering & Technology, Darapur
pratik.hutke@gmail.com

Abstract— It is the method to extract the user sessions from the given log files. Initially, each user is identified according to his/her IP address specified in the log file and corresponding user sessions are extracted. Two types of logs i.e., server-side logs and client-side logs are commonly used for web usage and usability analysis. Server-side logs can be automatically generated by web servers, with each entry corresponding to a user request. Client-side logs can capture accurate, comprehensive usage data for usability analysis.

Keywords—"Web log mining, User identification, Session identification"

I. INTRODUCTION

A web may allow users to interact and collaborate with each other in a social media dialogue as creators of user-generated context in a virtual community. So, World Wide Web becomes more popular and user friendly for transferring information. Therefore, people are more interested in analyzing log files which can offer more useful insight into web site usage. Web mining is one of the technique of data mining to extract useful information based on users' needs, under web mining, web usage mining is one of the application of data mining technology to extract information from weblog to analyze the user access to websites by. Web mining is the use of data mining technique to automatically discover and extract information from web documents and services.

Log files are files that list the actions that have been occurred. These log files reside in the web server. Computers that deliver the web pages are called as web servers. The Web server stores all of the files necessary to display the Web pages on the users computer. All the individual web pages combines together to form the completeness of a Web site.

User Name: This identifies who had visited the website. The identification of the user mostly would be the IP address that is assigned by the Internet Service provider (ISP). This may be a temporary address that has been assigned. Therefore here the unique identification of the user

is lagging. In some websites the user identification is made by getting the user profile and allows them to access the web site by using a user name and password. In this kind of access the user is being identified uniquely so that the revisit of the user can also be identified.

Visiting Path: The path taken by the user while visiting the web site. This may be by using the URL directly or by clicking on a link or through search engine.

Path Traversed: This identifies the path taken by the user with in the web site using the various links.

Time stamp: The time spent by the user in each web page while surfing through the web site. This is identified as the session.

Page last visited: The page that was visited by the user before he or she leaves the web site.

Success Rate: The success rate of the web site can be determined by the number of downloads made and the number copying activity under gone by the user. If any purchase of things or software made, this would also add up the success rate.

User Agent: This is nothing but the browser from where the user sends the request to the web server. It's just a string describing the type and version of browser software being used.

II. LITERATURE REVIEW

In Web Usage Mining, data can be collected in server logs, browser logs, proxy logs, or obtained from an organization's database. These data collections differ in terms of the location of the data source, the kinds of data available, the segment of population from which the data was collected, and methods of implementation[2].The log file that resides in the web server notes the activity of the client who accesses the web server for a web site through the browser. The contents of the file will be the same as it is discussed in the previous topic. In the server which collects the personal information of the user must have a secured transfer.

Web mining employs the technique of data mining into the documents on the World Wide Web. The overall process of web mining includes extraction of information from the World Wide Web through the conventional practices of the data mining and putting the same into the website features.

Learning the users expectation is a very tedious process[3]. A single word may have different views by different user. If the users area of interest is identified then we can have an efficient mining process. How is this done. If questions are posed to the user it would be a tiring process for a user to answer the question each time he makes a search. Therefore the users interest can be analysed by the first attempt made to open a page[6]. Then the next step done by the miner is to mine the web once again and provide the list of result meant only for the users area of interest[8]. This may in turn minimize the list of options and make the searching process even more effective. This can be done along with analysis of the log files to have utility as one of the factor.

Information is frequently gathered and automatically stored into access logs through Web server. Web usage mining process is similar to data mining process. The difference is in data collection phase. The data are collected from databases for data mining whereas it is collected from web log files in web usage mining.

In conventional data mining techniques information pre-process includes data cleaning, integration, transformation and reduction. But web mining pre-processing categorize into Content pre-processing, Structure pre-processing, Usage pre -processing[5]. Once the data is collected from log files, a three - step process is performed in web usage mining namely data preparation, pattern discovery and pattern analysis.

There are three types of log files which are as follows:
1 Web Server Logs
2 Proxy Server Logs
3 Browser Logs

1 Web Server Logs :

History of web page requests is maintained as a log file. Web servers are the costly and the most common data source. They collect large volume of information in their log files. These logs contain name, IP, date, and time of the request, the request line exactly came from the client, etc. These data can be bound together as a single text file, or divided into different logs, like access log, referrer log, or error log. However, user specific information is not stored in the server logs [15].

2 Proxy Server Logs :

It acts as an intervening level of catching lies between client browser and web servers. Proxy caching is used to decrease the loading time of a web page as well as the reduce network traffic at the server and client side. The actual HTTP request from multiple clients to multiple web servers are tracked by the proxy server [9]. The proxy server log is used as a data source for browsing behaviour characterization of a group of unauthorized users sharing a common proxy server.

3 Browser Logs :

On client side using JavaScript or Java applets the browsing history is collected. To implement client side data collection, user cooperation is needed. Here pre-processing discussed using Web Server Logs [11]. Web server logs are used in the web page recommendation to improve the E-Commerce usability.

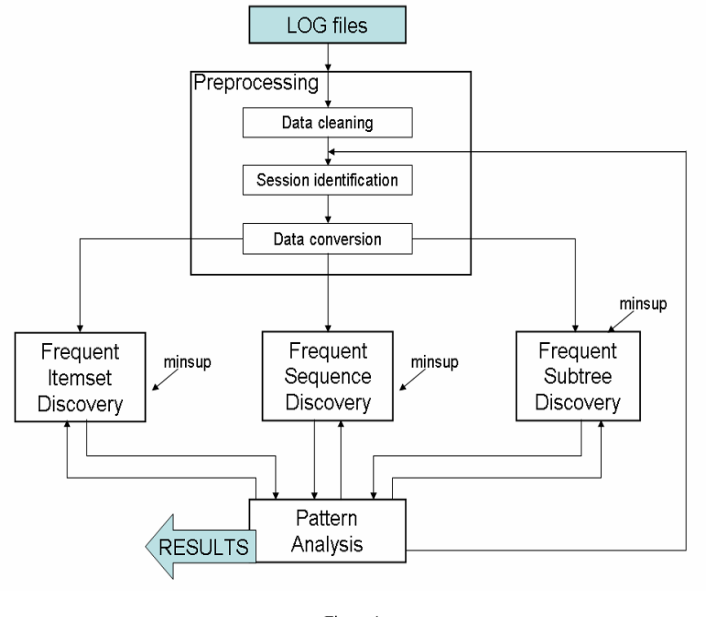
III.PROPOSED WORK AND OBJECTIVE

The proposed method consists of several phases such as file integration or merging, pre-processing, pattern discovery and pattern analysis. This paper focuses only on pre-processing phase that deals with three major issues such as data cleaning, user identification and session identification.

In general most of the users have tendency to open several pages simultaneously and in between, use some non browsing applications such as Ms-word, Excel etc for their own personal work, in such cases data recorded in server log only shows the requested time of the web pages and cannot help us to find out which web page and for how long has been really browsed on client machine. The calculated browsing time comparison shows that the time is reduced by considering the actual scenario of web page usage, which gives realistic browsing time of the user behavior at the web page.

Web Usage Mining and its algorithms have a bigger scope as far as research is concerned. Web mining and its application area is still in its infancy and requires more research. Besides Web content and Web Link, the Web Usage Mining is one of the most important areas of web mining research.

These application areas have got more research interest. The kind of data we can recently have from Web Log is not adequate. So, this research area is also highly promising. Web Mining and specifically Web Usage Mining can give rise to different application areas, which will really be beneficial for Web Users, society, and obviously for the governments. Log files are the best source to know user behavior. But the raw log files contains unnecessary details like image access, failed entries etc., which will affect the accuracy of pattern discovery and analysis. So preprocessing stage is an important work in mining to make efficient pattern analysis. To get accurate mining results user's session details are to be known. The research in future could be targeted to create more efficient session reconstructions through graphs and mining the sessions using graph mining as quality sessions gives more accurate patterns for analysis of users**.**



Data Cleaning:

It is also called as spider or not, it is a software tool that scans a website periodically to extract the content. All the hyperlinks from a web page are automatically followed by WR. The uninterested session from the log file is removed automatically when WR is removed.

User Identification:

Each different user accessing the website is identified in the user identification process. The aim of this process is to retrieve every user's access characteristics, then make user clustering and provide recommendation service for the users. Different users are identified by different ip addresses.

Session Identification:

A sequence of pages viewed by a user during one visit is known as the Session. The session is recorded in the log file. In pre-processing it is necessary to find session of each user. It defines the number of times the user has accessed a web page. It takes all the page reference of a given user in a log and divides them into user sessions. These sessions can be used as an input data vector in classification, clustering, prediction and other tasks.

The Records of- graphics, video and the format information:

In every record of URI field, JPEG, GIF, CSS filename extension is found, these extensions are going to be eliminated from the web log file. The files with these extensions are the documents embedded in the web page. So it is not necessary to include these files in identifying the user interested web pages. This process support to identify user interested sessions.

Failed HTTP- Status Code:

This cleaning process will reduce the evaluation time for finding the user's interested sessions. In this process the status field of every record in the web access log is checked and the status code over 299 or below 200 are removed.

Robots- Cleaning:

It is also called as spider or not, it is a software tool that scans a website periodically to extract the content. All the hyperlinks from a web page are automatically followed by WR. The uninterested session from the log file is removed automatically when WR is removed.

WEB CONTENT MINING

Web content mining data may be structured or unstructured/semi structured even though much of web is unstructured. It is the process of retrieving the information from the web into more structured forms and indexing the information to retrieve quickly or finding valuable information from web content or web documents. Web content mining includes the web documents which may consist of text, html, multimedia documents i.e., images, audio, video and sound etc. The search result mining contains the web search results. It may be a structure documents or unstructured documents.

Web content mining used many algorithms and tools such as Genetic algorithm, Cluster Hierarchy Construction Algorithm (CHCA), Correlation algorithm. Web Info Extractor (WIE), Mozenda, screen-scrapper, ontology based tools; webcontent extractor and automation anywhere are content mining tools. Cloud users require to extract the information from the cloud provided by web servers can make use of the web mining. For instance, Web communities can be maintained the information such as facebook. That is the users of same field of interest can be grouped and they can communicate through the network. This can be analyzed for the customers and enable provision to the customers based on their recommendations [18].

It has approaches; they are Database Approach. Figure 2 gives the web content mining approaches.

Database Approach

Database approach consists of databases which contain attributes, tables and schema with defined domains. It focused on techniques for organizing the semi structured data on the web into more collections of resources, and using standard database querying mechanism and data mining techniques to analyze it, for example multilevel database and web querying system [5]. Web content mining has the other approaches to mine the data. These are unstructured text data mining, structure mining, and semi-structure text mining and multimedia data mining.

IV. CONCLUSIONS

Web usage mining is indeed one of the emerging areas of research and important sub-domain of data mining and its techniques. In order to take full advantage of web usage mining and its all techniques, it is important to carry out preprocessing stage efficiently and effectively. This paper tries to deliver areas of preprocessing, including data cleansing, session identification, user identification. Once the preprocessing stage is well-performed, we can apply data mining techniques like clustering, association, classification etc for applications of web usage mining such as business intelligence, e-commerce, e-learning, personalization, etc. Web log mining is one of the recent areas of research in Data mining. Web Usage Mining becomes an important aspect in today's era because the quantity of data is continuously increasing. We deal with the web server logs which maintain the history of page requests

Web log file analysis began with the purpose to offer to Web site administrators a way to ensure adequate bandwidth and server capacity to their organization. By analyzing

these logs, it is possible to discover various kinds of knowledge, which can be applied behavior analysis of users.

Our proposed system is used to analyze the user sessions from which information regarding the problems occurred to the users and usage of the website can be obtained within particular intervals of time. This is used to configure the server and adjust the Web site which is highly useful for administrators.

ACKNOWLEDGEMENT

We take this opportunity to express our gratitude and indebtness to our guide Assistant Prof. PROF.A.P. Rudey Computer Science and Engineering department ,who is a constant source of guidance and inspiration in preparing this work.

We express our sincere gratitude towards our H.O.D. Prof. V.P. Nikam whose constant help and encouragement helped us to complete our Paper. We are grateful to Dr .S. M. Kherde Principal for his encouragement and support. We are also thankful to all the staff members of Computer Science and Engineering department ,whose suggestions helped us to complete the Paper work and those who have directly and indirectly helped for completion of the Paper.

REFERENCES

[1] G.Neelima and Dr.Sireesha Rodda "Predicting User Behavior Through Sessions Using The Web Log Mining" International Conference on Advances in Human Machine Interaction (HMI-2016),March 03-05,2016.

[2] Ruili Geng, and Jeff Tian "Improving Web Navigation Usability by Comparing Actual and Anticipated Usage"IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS, VOL. 45, NO. 1, FEBRUARY 2015.

[3] G. Neelima and Sireesha Rodda, "An Overview on Web Usage Mining", Springer International Publishing Switzerland December 2015.

[4] Gan Teck Wei, Shirly Kho, Wahidah Husain, Zurinahni Zainol " A Study of Customer Behaviour Through Web Mining"Volume 2, Issue 1 available at www.scitecresearch.com/journals/index.php/jisct/index, February, 2015.

[5] Thi Thanh Sang Nguyen, Hai Yan Lu, and Jie Lu, "Web-Page Recommendation Based on Web Usage and Domain Knowledge", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 10, OCTOBER 2014.

[6] Zhen Liao, Yang Song, Yalou Huang, Li-wei He, and Qi He "Task Trail: An Effective Segmentation of User Search Behavior", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 12, DECEMBER 2014.

[7] S.Vijayalakshmi V.Mohan, S.Suresh Raja, (2009) "Mining Constraint-based Multidimensional Frequent Sequential Pattern in Web Logs," European Journal of Scientific Research., Vol.36, pp 480-490.

[8] D.Vasumathi, and A.Govardan,(June 2009) "BC-WASPT : Web Acess Sequential Pattern Tree Mining," IJCSNS International Journal of Computer Science and Network Security., Vol.9, pp. 569–571.

[9] Jatin D Parmar, and Sanjay Garg,(June, 2007)"Modified web access pattern (mWAP) approach for sequential pattern mining," INFOCOMP Journal of Computer Science., pp. 46-54.

[10] Renuka Mahajan & J. S. Sodhi & Vishal Mahajan ,"Usage patterns discovery from a web log in an Indian e-learning site: A case study", Springer Science+Business Media New York 2014.

[11] Pani, S.K., Panigrahy, L.: Web Usage Mining: A Survey on Pattern Extraction from Web Logs. International Journal of Instrumentation, Control & Automation (IJICA) 1(1) (2011)

