ANALYSIS OF WEBSERVER LOG FOR THE EFFECTIVE WEBSITE

[1]D. Uma Maheswari,
Part-Time Ph.D- Research Scholar,
Karpagam University,Coimbatore,
Tamilnadu, India.
(uma.smv@gmail.com)

[2]Dr. A. Marimuthu
Associate Professor,
Department of Computer Science,
Government Arts College, Coimbatore.
Tamilnadu, India.
(mmuthu2005@yahoo.com)

Abstract---World Wide Web is rapidly increasing with huge amount of user's interactions. Web usage mining (WUM) is the process of extracting navigation behavior of user information in the Web-based environments. The user's access on the webpage information is stored in the web server logs. This paper discusses the preprocessing phase in web usage mining. The server log files must be preprocessing before applying some data mining techniques to predict pattern for pattern analysis. The main work of preprocessing is to eliminate noisy and unrelated data. This paper mainly concentrates about the first preprocessing task i.e. field extraction .The field extraction algorithm is used to divide the single line of the web log file. The data cleaning algorithm is used to prune such a noisy data which is not important for the purpose of analysis, thus improve the quality and efficiency of the data.

Keywords- World Wide Web; Web Mining; Preprocessing;

I. INTRODUCTION

The World Wide Web is considered as a huge library. The number of websites and its usage by the users are increasing rapidly. The World Wide Web consists of documents, images and some resources and interconnected by links and it has referenced with uniform Resource Identifiers. It identifies documents, files, service provider, and servers' services. The main access protocol of WWW is HyperText Transfer Protocol (HTTP) is used to communicate hundreds of protocols on the Internet. Through web services can communicate with different applications and share some information and services. A web service has more opportunity to connect with partners to exposing more services through that business for increase the revenue. To increase the web services some browser software such as Opera, Apple's safari, Google Chrome, Mozilla Firefox, Netscape Navigator, Mosaic, and Internet Explorer 9 is used to navigate from one page to another by hyperlinks rooted in the pages. These pages may contain some combination of Graphics such as 2D, 3D and animated graphics, single line of the web log file audio, video, text, etc. These pages will run automatically while the user interacts on the pages .By using the keywords can get some relevant information's by using some search engines like Google, Yahoo!, Bing, etc. Through the web, the person can share some ideas to audience on online from this can reduce the expenses and time. On web many cost free services are also supporting and build the web page application, build the web site, and blog. Web mining is the application of data mining. Web mining can be defined as to extract the knowledge from the web data including web documents, logs of websites; etc. The web mining is divided into three c categories i.e. Web content mining, Web structure mining and Web usage mining. Web usage mining is the part of web mining. Web usage mining is divided into three phases according to the kinds of data mined i.e. Data collection, preprocessing, pattern discovery and pattern analysis. The data are collected from three

main sources. They are web servers, Proxy servers, and web clients. This paper focuses on Preprocessing of data from logs.

II. RELATED WORK

Web usage mining is one of the important areas of many researchers in one the novel approach introduced the combining of web server logs and web contents for classifying user navigation pattern and predicting users future requests[5].There exist various researches in this area and we discuss few of them here. Arvind K.sharma et al., [1] described an effectiveness of website using web mining tool. The analysis is done by web expert tool of the website, this can find the information about the users, how people viewed that page, what they have downloaded, what search keywords they have used to get the website .From the obtained result can find out the browsing needs of the website users.

Yogis H K et al.,[2] has proposed two algorithms for web usage mining. The first method is used to separate the log set. The second method is used for prune noisy data for speed up the extraction time of the users need in the website.
Didit.D et al [3] the organizations keep their attention of their user needs in their website. An online navigation behavior grows each day extracting information is a difficult issue. A WUM is designed to operate on web server logs, here two tier architecture is proposed for capturing user's information about the user visited page.Practiaclly implemented this architecture with algorithm for accuracy. C.P.Sumathi et al., [4] discussed an overview of the various steps involved in preprocessing stages. The web data is suitable for the pattern discovery and analysis from click stream the information will be stored in web servers. This log files can be preprocessed for future work.

Norhaiza Ya Abdullah presented the detail approach of preprocessing step i.e is used to clean the web server logs. Web query log contain information such as the client's IP address, time and date of request made, the resources requested, status of request HTTP method used and the type of web browser and operating system. Web query logs from an online newspaper .The web query logs undergo preprocessing stage. Clickstream information is cleaned and partitioned into a set of user interactions which will represent the behavior. The web query logs will undergo necessary task in preprocessing which data are cleaning.
Marathe Dagadu Mitharam presented the preprocessing of web usage mining for extracting useful data from server log files to discover patterns in clickstream .The result obtained were satisfactory and contained valuable data about the log files. C.P.Sumathi discussed an overview of the various steps involved in preprocessing stages. The web data is suitable for the pattern discovery and analysis from click stream the information will be stored in web servers. This log files can be preprocessed for future work. Vellingiri.J.S the access data usually stored in the web server log files, web usage mining is used for exposing the usage patterns with sequential pattern mining, cluster mining, and association rule mining. These techniques are suitable for building adaptive web site.
Didit.D the organizations keep their attention of their user needs in their website. An online navigation behavior grows each day extracting information is a difficult issue. A WUM is designed to operate on web server logs, here two tier architecture is proposed for capturing user's information about the user visited page. Practiaclly implemented this architecture with algorithm for accuracy.

III. WEB USAGE MINING

The web server log files record the periodic attribute and resource attribute activity. The server log files are normal text files and stores the activity on the server. Internet is the most popular way for communication, retrieving and disseminating data. Day by day the number of users and their usage is increasing rapidly. When finding data from the website shows billons of information to see the desired information from the web is challenging task. The owner of the website have to help the user needs to provide the information to the users in the personalization mechanism. Thus the users fulfill their needs from the website and increase the profit from their online. The automated tools are used for helping the user's desired information by means of search, extract, and filter.

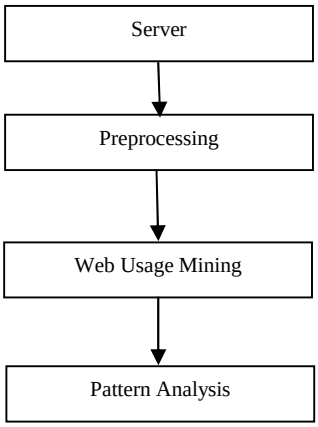
Web mining is the broad research area to address these issues, the growth of the web categorized into three i.e. content mining, structure mining, usage mining. The Content mining is used some techniques to find its web documents. The structure mining used in-links and out-links of web pages analyses. The usage mining defined the navigational behavior of the users. The usage mining techniques for the usage of the data based on the association rules, sequential patterns, classification and clustering. When the

content changes new pages inserted but not included in web log. For improving the personalization process, address these problems and detailed view about the semantic web technologies.

IV. PREPROCESSING AND EXATRACTION FEATURE

In many cases, preprocessing must need to be done before using web mining algorithm. Several tasks are involved in preprocessing. The tasks are cleaning, Pageview identification, user identification, sessionization, and path completion and transaction identification. Data cleaning technique is used to removes the extraneous references that are not important for analysis. Web log records the website URLs but not the web content requested by the users, it is very difficult to predict the user's interest, activities and behaviors. Data Preprocessing is an important task for converting the usage, content and structure information contained in primary data sources. In web usage mining the server log files include web access log, referrer log, and agent log. Sample from the population has collected and applied some implementation task to extract the browsing behavior of the consumer. The information recorded in proxy server logs in various formats such as Common log format and extended log format. Some clustering techniques have been used to group the similar characteristics of the user's behavior. This methodology is applied in large organization.

Fig 1.Web Usage Mining Process



V. EXPERIMENTAL RESULTS

In this experiment we have analyzed web server logs. We used 751MB data after preprocessing we got 3.5 MB of data. Data cleaning is the first step to remove all gif, jpeg, css etc., some images are hidden into folder, the log file examined all the hidden folders.

Number of Log files before preprocessing	Number of Log files after preprocessing
401810	44015

TABLE: 1 Result of Log files

After the images are removed the next step is to filter the status code.

Series	Status
100	Start again
200	Success
300	Redirect
400	Non Success
500	Server Error

Fig.2 HTTP Status Code

After the IP address of each user is identified, the users are further divide into different user agents. From the graph can find the user identification based browser. The role of Web Usage Mining is to analyze the log files of the website. It represents the information about the users, browsers, Os and errors used by the visitors.

Different browser of each IP page requested for user identification based on IP address. There are many IP address if the page is accessed from different browser, it shows that they are different users.

VI. CONCLUSION

In this paper, we have presented detailed of preprocessing phase, which is used to clean the log files. By using vb.net we have defined the rules to clean the log files. Through this experiment we have cleaned the irrelevant data. The primary function of

the server is record the web log information for characterization, evaluation, website development and reporting. Moreover there are some issues to resolved session identification. The future work involves session identification due to the fact that the client request most recent pages from the server.

REFERENCES

[1] Arvind K.Sharma and P.C.Gupta," Analysis of Web Server Log files to Increase the Effectiveness of the website Using Web Mining Tool," IJACMS, Vol.4, Issue.1, PP.572-579, 2013.

[2] Yogish HK and G.T Raju," Pre-Processing of Web Logs for Mining World Wide Web Browsing Patterns," IJRRASE, Vol.3, No.1, PP.28-33, Mar.2013.

[3] Dixit,D. and Gadge.J," Automatic Recommendation for Online Users Using Web Usage Mining,"IJMIT, 2,PP.33-42, 2010.

[4] C.P.Sumathi,R.Padmaja Valli and T.Santhanam," An Overview Of Preprocessing of Web Log Files for Web Usage Mining,"JATIT, Vol.34, No.1, Dec.2011.

[5]Liu,H.,et al.," Combined mining of web server logs and web contents for classifying user navigation patterns and predicting user's future requests", Data and Knowledge Engineering,2007,Vol61,Issue 2,pp.304-330.

[6 Yogish HK and G.T Raju,"Pre-Processing of Web Logs for Mining World Wide Web Browsing Patterns," IJRRASE, Vol.3, No.1, PP.28-33, Mar.2013.

[7] Raju G.T and Satyanarayana.P.S."Knowledge Discovery from Web Usage Data:Complete Preprocessing Methodology",IJCSNS,Vol.8,No.1,2008.

[8] Theint Aye aju,"Web Log Cleaning for Mining of Web Usage Patterns",ICCRD ,Vol.2,PP.490-494,2011.

[9] Shaily Langhnoja, Mehul Barot, nd Darshak Mehta," Pre-Preocessing:Procedure on Web log file for web usage mining",IJETAE,Vol.2,Issues.12,2012.

[10].Vijayashri Losarwar and Madhuri Joshi ,"Data Preprocessing in Web Usage Mining",International Conference on Artificial Intelligence and Embedded Systems (ICAIES'2012), 2012 .

[11].Dharmendra Patel, Kalpesh Parikh and Atul Patel, "Sessionization –A Vital Stage in Data Preprocessing of Web Usage Mining-A Survey", International Journal of Engineering Research and Applications (IJERA), Vol. 2, Pp. 327-330,2012.

[12] Norhaiza Ya Abdullah, Husan Sarirah Husin,Herny Ramadhani and Shanmuga Vivekanada Nadarajan,"Pre-Processing of Query Logs in Web Usage Mining," IEMS,Vol.11, No.1,PP.82-86,Mar.2012.

[13] Marathe Dagadu Mitharam,"Preprocessing in Web Usage Mining," IJSER, Vol.3, Issues.2, Feb.2012.

[14] C.P.Sumathi,R.Padmaja Valli and T.Santhanam,"An Overview Of Preprocessing of Web Log Files for Web Usage Mining,"JATIT, Vol.34, No.1, Dec.2011.

[15]Vellingiri.J.S. and PAndian.C,"A Survery On Web Usage Mining,"GJCST, 1, PP.4343-4350, 2011.

[16]Dixit,D. and Gadge.J,"Automatic Recommendation for Online Users Using Web Usage Mining,"IJMIT, 2,PP.33-42, 2010.