International Journal of Computer Applications in Engineering Sciences
[VOL III, ISSUE I, MARCH 2013] [ISSN: 2231-4946]

Hybrid Model for Preprocessing and Clustering of Web Server Log

T. Subha Mastan Rao[1], Thinley Lhendup[2], Thinley Wangdi[3], Sujata Pradhan[4]

[1]*KL University, AP, India*

[1]mastan@kluniversity.in
[2]charms.lhendup@gmail.com
[3]thinley11@gmail.com
[4]sweet-suji@hotmail.com

***Abstract*- With increased rate in the usage of the World Wide Web (www) is growing both in its complexity and the volume of traffic of web site, it has become very important to analyze this web traffic and the usage of the web site by the users. Web usage mining is a main research area in web mining focused on learning about web users and their interaction with web sites. The information like server log, browser cookies and other relative information can be used to find user's access models automatically and quickly from the web log data, such as most frequent access parts, least recent access page group and user cluster[1][2]. In this paper we are analyzing the system by implementing the major recommendation approaches and preprocessing of log files stored in web server by applying clustering algorithm. Based on that, we are extracting valuable information along with the behavior of interested users and web designer to enhance better foundation for decision making of an organization and better service to the customer and web users**

***Keyword*-: Clustering approach, Knowledge set, Preprocessing and, Server log file, Web usage mining**

I. INTRODUCTION

The web usage mining is a data mining technique used to discover the patterns of the web and its behavior by extracting and integrating useful information like server logs, user history and other relevant information through user interaction with the web[2][3][4]. Since web has become reservoir of information of all web user's throughout the world for every association, offices and e-commerce. For enhancement of the quality of web system and user's interaction with web sites, we are analyzing the system and the patterns of web through server logs and preprocessing it using some approach. As data mining is main research area to expertise the web and it's system, many research groups had came out with various processing technique and classification algorithm, but due to alarming growth in world wide web (WWW), it has fail to meet the demand of rapid growing volume of traffic of web and its complexity.

This paper focuses on preprocessing and clustering of web server logs, and provides information on importance of doing such research and its advantages, so that we can achieve the conversion of browser's into buyers.

The rest of paper provides information as follows: Section 2, we explain about our approach, its process along with our proposed work. In section 3, we present an overview of preprocessing, particularly an algorithm used. In section 4, we describe clustering, its features and algorithm used. In section 5, describes about web server log and present our experimental result. Finally, in section 6, we present our conclusion.

II. APPROACH AND RELATED WORKS

A. Preprocessing of Log Files Stored in Web Server.

Unlike existing system, we do not interact with users directly rather we have developed our own algorithm to extract required and relevant information from the server log files from various sources.

B. Applying Clustering Algorithm.

In our approach, the clustering of the Web users is studied based on the browsing activities and their accessing behavior. We then cluster the users into various classes based on their similarity in performed activities.

C. Analyzing the System to Provide Knowledge set.

We first find out possible approaches to how and from where we can gather the required information to be mined. So that by using those approaches we can categorize and provide better recommendation to the users.

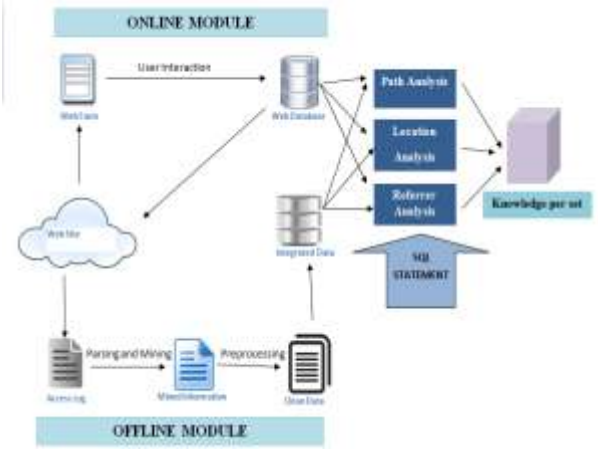

Fig 1: Architecture

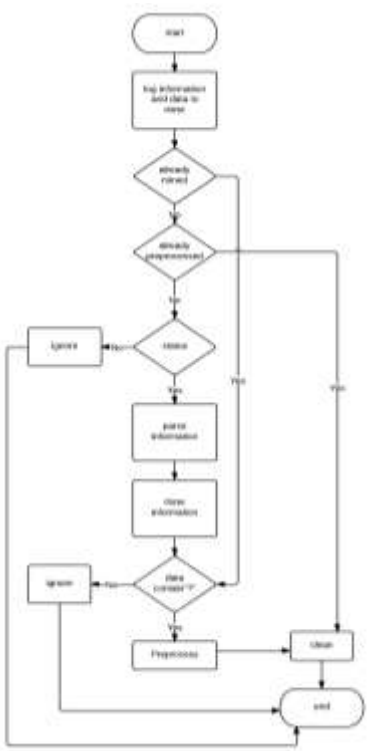

Figure 2: Approach

III. Web Log Files

Web log files are files that contain information about website users and are created by web servers automatically. Log file contain information about like user name, IP address, date, time, byte transfer, access request.

Web log file is a simple plain text file which record information about each web users and display log files data in different formats.[6][10]

I. W3C extended log file format.
II. IIS log format.
III. NCSA common log format.

```
127.0.0.1 - - [24/Feb/2013:19:51:56 +0530] "GET /xampp/ HTTP/1.1" 302 237 "-" "Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.22 (KHTML, like Gecko) Chrome/25.0.1364.97 Safari/537.22"
12.6.9.1 - - [24/Feb/2013:19:51:56 +0530] "GET /xampp/splash.php HTTP/1.1" 200 1325 "-" "Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.22 (KHTML, like Gecko) Chrome/25.0.1364.97 Safari/537.22"
10.45.10.4 - - [24/Feb/2013:19:51:56 +0530] "GET /xampp/xampp.css HTTP/1.1" 200 4178 "http://localhost/xampp/splash.php" "Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.22 (KHTML, like Gecko) Chrome/25.0.1364.97 Safari/537.22"
```

Fig 3: Web Server Log

IV. Mining and Preprocessing.

Mining of data is done in order to enhance the quality of the data before applying for any preprocessing algorithm. The goal of the preprocessing is to transform the raw data into a set of user profiles and to make it suitable web log data for various data mining technique.

A. User Identification

Identification of different users by analyzing their IP addresses, and categorize as per the web site and pages they accessed. If IP address is same but browser version is different then it is considered as different user.

B. Session Identification

The session means time duration spent in each web by any users. The goal of session identification is to divide the page accesses of each user into individual session.

By analyzing this information, web usage mining system can determine temporal relationships among data items and web sites.[7]

C. Path completion

The task of path completion is to fill the missing page reference that is not recorded in the access log when user browses local cache and proxy servers [5]. Similar to user identification, if the page is requested that is not directly linked to the previous page accessed by the same users, the referrer log can be checked to see what page the request came from. If the page is in the user's recent click history, it is assumed that the user browsed back using cached session of the pages.

After the path completion is done, it is then assumed that the page of previously requested page is the source of the new page requested.

D. Status Analysis

Every other identification is associated with its browsing status. Status of the user interaction with the given path is denoted with a three digit number.

The number denoted by 200 series tells us that the viewing was successful. The viewing which went through redirections is denoted with 300 series and those with client error with 400 series. Thus, only data with 200 series are considered useful.

E. Data cleansing

Data cleansing are done mainly to remove unwanted items stored in the log file containing extension like jpg, gif, png and other irrelevant data not used during analysis time.

Data cleansing being site specific, involves numerous extraneous information that may not be important during the time of analysis[15]. The problem of inconsistencies and irregularity of data in multiple data sources in ambiguity, duplication, non-integrity and some others, the data cleaning accordingly performs the cleaning operations for the error data. Thus data cleaning is considered as an important task in data mining.

Step1: Set Variable array for mining.
Step2: Set Pattern to match all log files.
Step3: Open file to read.
Step4: While (!EOF)
If (Pattern match Record)
Set ParseData;
If (ParseData[“path”]<>’ ‘ &&
ParseData[‘status’]<299)
MineLine();
Else
Ignore;
Else
Ignore;
Repeat:
Step5: read mined file
Step6: for (!EOF)
If(path have”.jpg,.gif,.png”)
Ignore;
If (path contain “?”)
Save data();
Repeat:

Approach 1: Mining and Preprocessing Approach

V. CLUSTERING

Due to rapid development of World Wide Web (WWW), Web is now became repository and information provider to every web user such as online-shopping, user feedback, technical support, etc.[8] Thus, for better and customize service for the customer, we are clustering web into different classes based on their common properties to analyze the characteristics of the web users throughout the World Wide Web.

Approach 2 have been used for the following clustering techniques to find out the best mean of the group set.

5.1 Path Analysis

Path Analysis is a clustering approach to find out the cluster with respect to path of the log data. The count of the repeated path information helps us to find out which path has been frequently visited. However, if you want to find out the least recently visited path, the clustering is done with respect to date visited and not the count of path repetition.

5.2. Location Analysis

Location analysis can be done with respect to the IP Address of the client recorded in the log information. IP to Country can be mapped in various ways. Although there are many APIs provided by various domains in the internet, few organizations also provide huge IP to Country mapping CVS Database.

5.3. Referrer Analysis

Referrer is the path of the page which was referred just before the client referred to the current website.

This is a very important piece of information with respect to knowing from which site the current website is referred from.

Step1: Find the count of reference to every group
Step2: Randomly assign one count as initial.
Step3: Find the farthest count from the initial.
Step4: Find the centroid of the above two.
Step5: Assign the centriod as the initial.
Step6: Repeat step 3 and step 4.

OUTPUT: Best Mean Count Set.

Approach 2: Clustering Approach

Figure 3: Path Analysis Screen Shot.

Figure 4: Location Analysis Screen Shot.

Figure 5: Referrer Analysis Screen Shot.

VI. CONCLUSION

Web usage mining essentially has many advantages which makes technology and web system more attractive to corporation and organization throughout the world by enabling the web users to grape the useful information and knowledge as per their requirement.

In this paper, for fulfillment of our objective, to convert "browsers into buyers", we are focusing on two main key point in web usage mining i.e. pre-processing and clustering for better services to the every web users.

REFERENCES

[1] Sheetal A.Raiyani, and Shailendra Jain, "Efficient Preprocessing Technique using Web Log Mining," International Journal of Advancements in Research and Technology,vol 1, issue 6,Nov-2012.

[2] C.P.Sumanthi, R.Padmaja Valli, and T.Santhanam, "An Overview of Preprocessing of Web Log Files for Web Usage Mining," Journal of Theoretical and Applied Information Technology, Vol-34, No-2, Dec-31st.

[3] Priyanka Patil and Ujwala Patil, "Preprocessing of Web Server Log File for Web Mining," National Conference on Emerging Trends in Computer Technology (NCETCT-2012), April 21, 2012.

[4] Mr.Ravindra gupta, and Prateek Gupta, "Fast Preprocessing of Web Usage Mining with Customized Web Log Preprocessing and Modified Frequent Pattern Tree,' International Journal of Computer Science and Communication Network, Vol 1(3), 177-279.

[5] Ms.Dipa Dixit and Ms.M Kiranthika, "Preprocessing of Web Logs," International Journal on Computer Science and Engineering, Vol-02, No-07, 2010, 2447-2452.

[6] Ravindra Gupta and Prateek Gupta, "Application Specific Web Log Prepreocessing." Int.J.Computer Technology and Application, Vol-3(1), 160-162.

[7] Shaily Langhaoja, Mehal Borot, and Darshak Mehta, "Preprocessing:Procedure on Web Log file for Web Usage mining," International Journal for Emerging Technology and advanced enfineering, Vol-2, issue 12, Dec-2012.

[8] K,Ranvichandra Rao, "Data Mining and Clustering Technologies," DRTC Workshop on Semantic Web 8th-10th December, 2003, DRTC, Bangalore.

[9] Grabriel Fiol-Roig, Margaret Miro-Julia, and Eduardo Herraiz,"Data Mining Technologies for Web Page Classifcation," Ctra.de Valldemossa km.7, 5, 07122 Palma de Mallarca, Spain.

[10] Surbhi Anand, and Rinkle Rani Aggarwal, "An efficient Algortihm for Data Cleaning of Log file Using File Extensions," International Journal of Computer Application (0975-88), Vol-48, No-8, Jan 2012.

[11] Siarash Emtiyaz, and Mohammad Reza Keyvanpour, "Adaptive Classification of Web Mining Methods and Challenges of Customer Relationship Management Domain," International Journal of Scientific and Engineering Research, Vol-2, Issue-5, May-2011.

[12] Marathe Dagadu Mtharam, "Preprocessing in Web Usage Mining," International Journal of Scientific and Engineering Research, Vol-3, Issue-2, Feb-2012.

[13] Risto Vaarandi, "A Data Clustering Algorithm for Mining Pattern From Event Log," Proceeding of the 2003, IEEE Workshop on IP Operations and Management.

[14] Siu-Tong Au, Rong Duan, Siamak G.Hesar, and Wei Jaing, "A Framework of Irrigularity Enlightment for Data Preprocessing in Data mining," Ann Oper Res, DOI 10.1007/s10479-008, 0495-z.

[15] T.Ravathi, M.Mohana Rao, Ch.S.Sasanka, K.Jayanth Kumar, and B.Uday Kiran, "An Enhanced pre-preocessing Research Framework for Web Log Data," International Journal of Advanced Research in Computer Science and Software Engineering, vol 2, issue 3, March-2012.