International Journal of Engineering Science Invention (IJESI)
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org || PP. 26-33

Preprocessing Techniques in Web Log Mining to Group Users and Identify User Session

Vadivazhagan K[1], Dr. M. Karthikeyan[2]
[1]*Department of Computer & Information Science, Annamalai University, Annamalai Nagar-608002*
[2]*Department of Computer & Information Science, Annamalai University, Annamalai Nagar-608002*

Abstract: *In this current digital scenario, all information and communications are shared through web, and especially for educational institutions sharing of information has increased exponentially. The fast increase in the usage of web site hits demands to provide the exact and clear information to the users. To ease the accessibility of the web site can be improved by applying the Data mining techniques to extract the knowledge to personalise the web site. Web mining is classified into three types and they are Web Structure Mining, Web Content Mining and Web Usage Mining. Various preprocessing and clustering algorithms are used for implementing web based application with Structured Query Language and Programming Language. The main objective of this paper is to provide better preprocess method to parse the raw log files for user identification and session identification with the use of cleansed data. This paper discusses various preprocessing techniques for Web Usage Mining in the area of Web Log Mining for data preprocessing, pattern discovery, knowledge discovery and pattern analysis. Web based application is developed to provide better cleaning techniques and more appropriate cleansed data for further data mining process. Emphasis is given more on cleaning web log phase in preprocessing to provide the best method for web log data preprocessing.*

Keywords: *Web Mining, Web Usage Mining, Web Log Mining, Session Identification, User Identification.*

I. Introduction

Every single click made by the web visitor is recorded, because it is assumed that it can be a important source of any useful information. Big volume of data as a web log entries are stored in the web server. In each hit, a numbers of log entries will be stored as a web log file. The amount of web log entries is kept in web server as web log files. The users of these data expect more relevant information. The various data mining techniques have to be implemented in web based application to resolve this problem. The web based applications used to preprocess the web log to remove unwanted log entries to make the data relevant and make it ready for data mining. Initially the original server log file contains IP of client machine, time of the hits, bytes used to download pages, web documents, hyperlinks between documents, and etc.,

Web mining can be basically divided into three categories as follows, depending on the data to be mined and they are

A. Web Content Mining

Web Content Mining is the process in which the useful information is extracted from the content of web documents. Content data is the collection of facts a web page is designed to contain such as text, images, audios, videos, structured records such as lists and tables. Issues addressed in Web content mining include topic discovery and tracking, extracting association patterns, clustering of web documents and classification of web pages.

B. Web Structure Mining

Web structure mining is the process which is used to discover structure information from the web. The Structure of a typical web graph consists of web pages as nodes, and hyperlinks as edges connecting related pages. Hyperlinks are a structure that connects a location to a different location within a page or out of the page with different one. Documents Structure the content within a Web page can also be organized in a tree structured format, based on the various HTML and XML tags within the page. Mining efforts here have focused on automatically extracting document object model (DOM) structures out of documents.

C. Web Usage Mining

Web usage mining is the application of data mining techniques to discover interesting usage patterns from web usage data, in order to understand and better serve the needs of web-based applications (Srivastava,

Cooley, Deshpande, and Tan 2000). Usage data captures the identity or origin of web users along with their browsing behavior at a web site.

D. Web Log Mining

Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data in order to understand and better serve the needs of Web-based applications. Usage data captures the identity or origin of Web users along with their browsing behaviour at a Web site. In Web Usage Mining, the Web Log Mining plays a vital role for effective discovery of knowledge from the web log. Fig 1 shows the various mining process involved in web mining process.

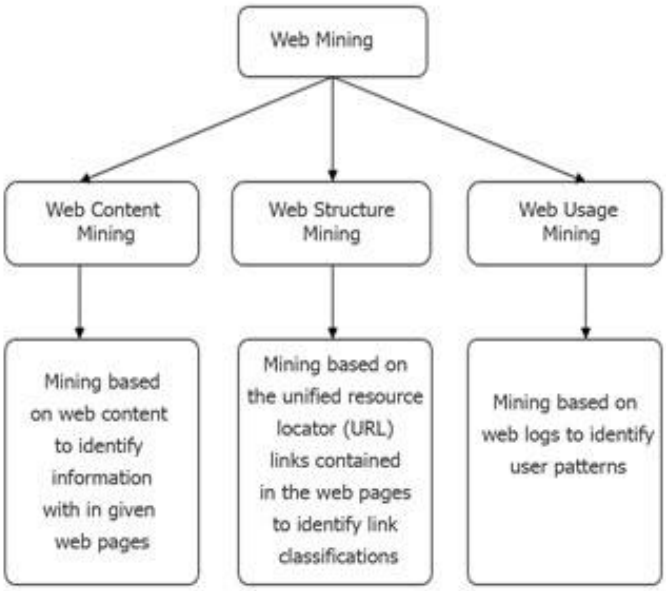


Figure 1: Web Mining Taxonomy

II. Literature Review

The literature review is focused to study and compare existing preprocessing techniques which can be used to adopt best method in the web log preprocessing. In the process of removing large amount of irrelevant log entries, the preprocessing of web log file becomes appropriate. Because the base log file cannot be directly used in Web usage Mining process.

Renuka Mahajan [1] discusses the detailed results of a case study of web data mining in a specific e-learning application. The main objective of this study is to conduct research on usability and effectiveness of the e-content by analyzing the web log. They have evaluated different features of e-content that can lead to better learning outcomes for the learners, by understanding their navigational behaviours, their interaction with system and their area of interest.

Prathibha Sharma in [3] performed comparative analysis between web based log formats pre-fetching using two main techniques, i.e. Apriori and FP Growth so that users' navigational behavior can be extracted easily and efficiently. To filter spam conventional strategies, such as dark white records (URL, IP, Address, Mailing information) is practically unimaginable. Use of content mining strategies to a web logs can raise proficiency of a filtration of spam.

Brijesh Bakariya [4] proposes a Rare Itemset Mining from Weblog Data (RIMWD) algorithm to extract rare itemset. Most of the algorithms follow a bottom-up strategy, but those strategies are only suitable for discovering frequent itemsets.

Mitali Srivastava [5] improved later phases of web usage mining like pattern discovery and pattern analysis several data preprocessing techniques such as Data Cleaning, User Identification, Session Identification, Path Completion etc. have been used. In this paper all these techniques are discussed in detail.

Michal Munk [6] focused on the reconstruction of activities of the web visitor. This advanced technique of data preprocessing is a time consuming one. In the article they tried to assess the impact of reconstruction of activities of a web visitor on the quantity and quality of the extracted rules which represent the web user' behavior patterns.

Bharat Chauhan [7] has designed a custom algorithm for the Clustering process. The main aim of this algorithm is to provide more efficient and accurate results as compared to the other Clustering algorithms. By analyzing the browsing behavior of various users, next web page predictions can be made which is an important aspect of most of the websites these days. But the prediction of future requests comes with its own set of issues which make it unreliable at times; there are some concerns with accuracy and efficiency.

Cooley et al. [11] proposed methods for data cleaning, user identification, session identification and transaction identification. Although their methods are good enough but some heuristics are not appropriate for complex web sites.

Robert et al. [18] introduced a new concept called integer programming for better session identification. This method generates session simultaneously and produced session better match to an empirical distribution.

III. Methodology

Data Preprocessing

The key steps involved in web usage mining process are web log data preprocessing, pattern discovery and pattern analysis. In these above processes, the data preprocessing is considered to be complex and time consuming due to vast amount of structured and unstructured nature of web log data stored in the web server. Also, preprocessing phase takes more time than other phases of web usage mining. It is important to apply necessary preprocessing techniques effectively to improve efficiency and scalability of data for meaningful web usage mining process. After downloading the web log from the web server, it has to be uploaded to the data base as a table. In this table, the date field will be formatted according to data base syntax for easy SQL access. Columns are separated and named meaningful to access records accordingly. The information stored in the web logs is generally heterogeneous and semi-structured. These log files also contain few entries which will not use for any analysis. So to make analysis in a better way it is important to remove irrelevant entries. This will reduce the volume of data by keeping only the relevant data for analysis. Preprocessing has number of challenges and issues which lead to a variety of algorithms and a number of heuristic techniques for each step of preprocessing. Various steps of preprocessing techniques are presented below:

Table 1 shows the attributes of Extended Common Log Format. It is the commonly used standard non-customised format log format which is suitable for http web sites.

Table 1: Log File Meaning

IP	IP address which has access to the users or the IP address the users' agent server
Date	The requested date
Time	The requested time with time zone
Request Path with Method	Request path of visitors, there are three common methods, GET, POST and HEAD
Status	The status code returned by the server
Bytes	Bytes used to download the requested page
Referral Path	Requested resource of visitors
User Agent	Additional information, including browser type, operating system

The Table 2 shows the downloaded web server log entries. This type of log includes user's *IP address/hostname, rfcname, log name, data with time zone, page access method, PATH, http version, server response code, byte received, referral path and user agent*.

Table 2: Downloaded Log Entry In A Database Table

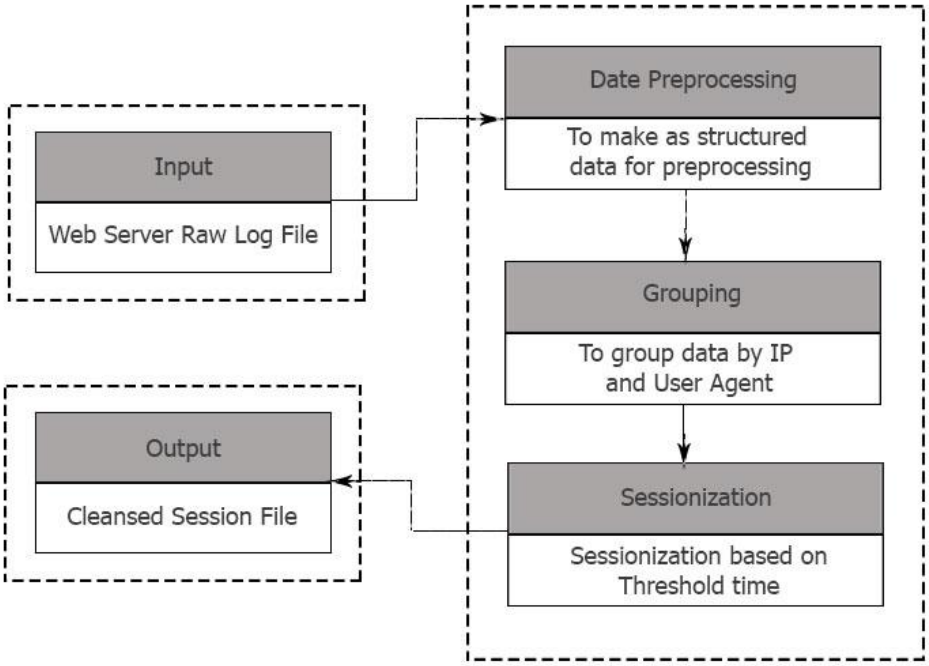
slno	ip_address	date_time	request_path	status_code	bytes	referral_path	user_agent
501	14.139.186.22	2017-03-24 11:38:18	GET /exam/assets/js/bootstrap.js HTTP/1.1	200	60681	http://www.annamalaiunive	"Mozilla/5.0 (Windows NT 6.1; Win64; x64) AppleWeb...
502	14.139.186.22	2017-03-24 11:38:18	GET /exam/assets/fonts/glyphicons-halflings-regula...	200	23320	http://www.annamalaiunive	"Mozilla/5.0 (Windows NT 6.1; Win64; x64) AppleWeb...
503	14.139.186.22	2017-03-24 11:38:18	GET /assets/staff_photos/03260.jpg HTTP/1.1	200	17782	http://www.annamalaiunive	"Mozilla/5.0 (Windows NT 6.1; Win64; x64) AppleWeb...
504	14.139.186.22	2017-03-24 11:38:19	GET /assets/staff_photos/04641.jpg HTTP/1.1	200	14651	http://www.annamalaiunive	"Mozilla/5.0 (Windows NT 6.1; Win64; x64) AppleWeb...
505	14.139.186.22	2017-03-24 11:38:18	GET /exam/images/headerk1.png HTTP/1.1	200	177539	http://www.annamalaiunive	"Mozilla/5.0 (Windows NT 6.1; Win64; x64) AppleWeb...

In this work it is proposed to work on web log file stored by web server. First, weblog file was collected and then preprocessing is implemented on weblog file. In preprocessing steps, irrelevant web log data were converted into formatted web log data. Formatted data obtained through preprocessing is used for further processing. The Preprocessing has three steps and they are Cleaning, User Identification and Session Identification. All log entries which has no meaning during mining process will be removed in cleaning process. Users are identified on the basis of IP address and equivalent User Agent entries which are available in the web log file. In User Session, identification process sessions are done by taking threshold value of time.

Log files contain information about User Name, IP Address, Time Stamp, Access Request, number of Bytes Transferred, Result Status, URL that Referred and User Agent. The log files are maintained by the web servers. Analysis of these log files, provides a clear idea about the user. This process gives a detailed discussion about these log files, their formats, their creation, access procedures, their uses, various algorithms used and the additional parameters that can be used in the log files which in turn will give way for effective mining.

The Fig 2 shows the steps involved in web log preprocessing techniques, such as user identification and session identification.

Figure 2: Various Steps Involved In Data Preprocessing



For pre processing following sample of web log is considered which was taken from Annamalai University Apache web server.

157.50.255.122 - - [02/Jan/2017:17:38:14 +0530] "GET /faculty_more.php?facultyid=11412&deptcode=0176 HTTP/1.1" 200 52208 "https://www.annamalaiuniversity.ac.in/faculty_dept.php?factcode=53" "Mozilla/5.0 (Linux; U; Android 4.2.2; en-US; P2S Build/JDQ39) AppleWebKit/534.30 (KHTML, like Gecko) Version/4.0 UCBrowser/10.4.1.565 U3/0.8.0 Mobile Safari/534.30"

Since HTTP is a stateless protocol, web log entries have been stored in web server as user activities in the web site navigations. In this web log entries contains all the files which is supporting files of the current web page. So those web log entries other than requested and referral URL entries contained rows to be deleted using preprocessing techniques. However, the access requests and referral log file records are not all what we want, so we must carry on the data preprocessing, this paper also puts forward a method of data preprocessing based on user interests.

B. Clustering Technique Used To Identify Users

A unique algorithm designed by us is used for the User Identification Clustering process. It intends to produce efficient and precise results. The main process of the algorithm is to group users that have similar navigation patterns with similar user agent .

Methods involved in Grouping users and Session Identification:

```
function GroupProcess(user, table) {
        SELECT ip_address, user_agent FROM tablename GROUP BY ip_address, user_agent
        for(i=0; i<num_rows; i++){
                UPDATE table SET `groupcount`=$i+1 WHERE ip_address=ip_address AND user_agent=user_agent
        }
        UPDATE log_property SET groups = num_rows WHERE logid=logid
                return Total Groups Found :num_rows1
}


function SessionsProcess(user, table, groupcount, threshold) {
        num_rows = 0
        cnt = 0
        do {
                SELECT slno, date_time FROM table_cleaned WHERE `groupcount`=groupcount AND
                                sessionscount=0 ORDER BY date_time
                cnt ++;
                if(num_rows != 0) UPDATE table_cleaned SET sessionscount=cnt WHERE groupcount=groupcount
                                AND date_time BETWEEN date_time AND ADDTIME(date_time)threshold AND clean=0
        } while(num_rows2 != 0)
        SELECT DISTINCT(sessionscount) FROM table_cleaned WHERE groupcount=groupcount AND clean=0
        return numrows
}
```

IV. Experimental Work And Discussion

Results Retrived By Removal Of Irrelevant Log Entries

Web base Preprocessing resulted in obtaining relevant data with lesser time consumption than other conventional methods. While removing irrelevant log entries, it counts the data and gets recorded in a separate table as properties for analysis. In fact, different virtual table is created for cleansed log entries instead of removing entries from data base table in order to preserve the data. Following is the sample SQL WHERE clause.

```
public function WhereTable() {
        return  WHERE
                        `status_code` IN('301', '302', '304', '404', '408') OR
                        `request_path` LIKE '%.css%' OR
                        `request_path` LIKE '%.jpg%' OR
                        `request_path` LIKE '%.gif%' OR
                        `request_path` LIKE '%.js%' OR
                        `request_path` LIKE '%.png%' OR
                        `request_path` LIKE '%.ico%' OR
                        `request_path` LIKE '%.woff%' OR
                        `request_path` LIKE '%.eot%' OR
                        `request_path` LIKE '%.ttf%';";
}
```

Statistical Analysis

For statistical analysis log data were collected from the web server of Annamalai University website for 5 users. Results obtained by implementation of web based application are shown below in figures and Table 3:

The following table shows the overall file formats and their percentage during the cleansing process. For this work, the preprocessing is done with the web site using the web logs and the total number of count for each type of file is calculated. The percentage of availability is also deduced.

Table 3: General Statistics Derived From Web Based Application

Description	Log Count
Start Time	2017-03-24 11:23:44
End Time	2017-03-24 11:53:21
Total Time Spent	00:29:37
Unique IP	1
Max Used IP	14.139.186.22
Total Logs	843
After Clean Logs	126
Cleaned Logs	717
Total Users	5
Total Sessions	8
Threshold Time	15 Minutes
Average Time/Sessions	0:3:42

Table 4 shows the total bytes over all and bytes used for after cleansing and before cleansing log entries.

Table 4: Details of Total Bytes used to downloaded

Protocol	Log Count	Bytes(KB)	Average(KB)
Total Bytes Used	843	38791.21	46.02
After Cleansing	126	17825.88	141.48
Filtered Bytes	717	20965.33	29.24

The Table 5 represents the total number for each file type found in the web log. Also, the relative percentage for each file type present in web log is determined by mathematical calculations. Among the tabulated file types, the data which does not influence in any way to identify a particular user are considered as noise and are removed from the web log. Fig 5 the horizontal axis represents the number of count and the vertical axis represents the file types. According to the results, it can be concluded that maximum number of data is irrelevant.

Table 5: Count and percentage of file types

Status Code	Count	Percentage %
.php	130	15.42
.pdf	4	0.47
.rar	0	0
.txt	0	0
.jpg	414	49.11
.gif	4	0.47
.png	64	7.59
.css	105	12.46
.js	66	7.83
.ico	31	3.68
.woff	25	2.97
.eot	0	0
.ttf	0	0
Total	843	100

Figure 3 shows the count of file types identified from web server log and in which to find the relevant log entries and to remove the irrelevant log entries.

Figure 3: Count of file types in log

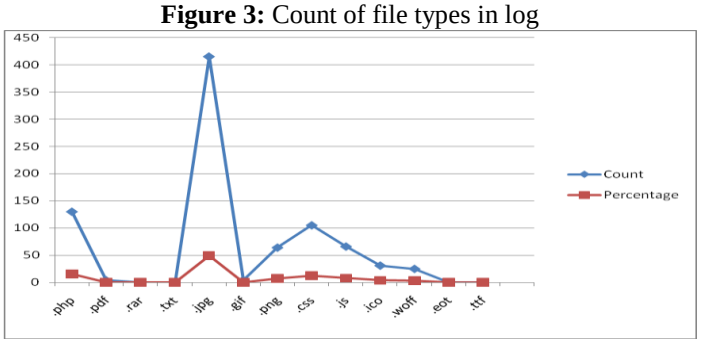


Table 4 shows the successful and unsuccessful count of the retrival of log entries. Among these log entries the unsuccesful logs does not possess any relationship with usage log mining considered as noise and are removed.

Table 4: Details of Status code from log entries found

Status Code	Count	Percentage %
Successfull		
200	620	73.55
206	2	0.24
Unsuccessfull		
301	5	0.59
302	1	0.12
304	147	17.44
404	68	8.07
Total Status Codes	843	100

Fig 4 shows the successful retrival log entries. The 73.79 % successful retrival was achieved.

Figure 4: For successful log entries found

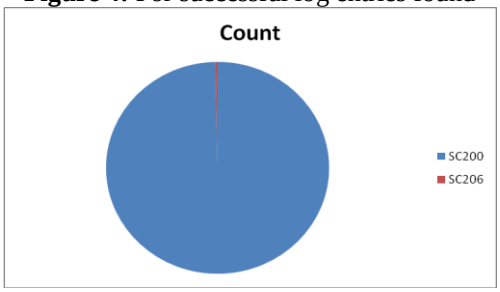
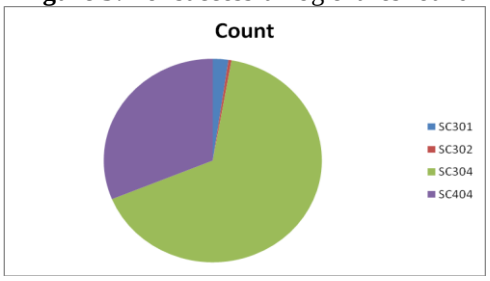


Fig 5 shows the unsuccessful retrival of log entries. The 26.21 % of unsuccessful retrival was achieved.

Figure 5: For successful log entries found



From the above stated analysis it is observed that for 5 users the total number of log entries found 843 instead of 126 relevant log entries and then all other irrelevant log entries were removed to get rid of unnecessary calculation for further phases of mining.

V. Conclusion

Web log mining is used to find behaviour of the users which is more important to provide better services to the web page visitors. Web usage mining is a field of study to analyse users and propose to generate useful patterns. Due to irrelevant vast data in web server log file, data preprocessing is considered to be an essential process in web usage mining. In this paper various preprocessing techniques were used, such as Data cleaning, User Group Identification and Session Identification with use of various threshold time. Web based application have been successfully implemented using different procedural ways. The proposed method yields an average process time of 3minutes and 42 seconds per session. Even though there is an exponential growth in web page hits and as well web log entries, web based application provides better preprocessed data for knowledge discovery. The work may be extended on ontology based user groups and user sessions in future.

References

[1]. 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*.

[2]. Vikas Tiwari, Dhruvesh Chudasama, Prof. Avinash Ingole, Model Survey on Web Usage Mining and Web Log Mining, *International Journal for Research in Applied Science & Engineering Technology (IJRASET), Volume 4 Issue XII*, December 2016 IC Value: 13.98 ISSN: 2321-9653.

[3]. Prathibha Sharma, Brahmdutt Bohra, Surendra Yadav, Comparative Analysis of Web-Mining Approaches for Efficient Mining of Server Log Formats, *2016 5th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions)*, Sep, 7-9, 2016.

[4]. Brijesh Bakariya, G. S. Thakur, Mining Rare Itemsets from Weblog Data, *Springer, Natl. Acad. Sci. Lett. (September-October 2016) 39(5):359-363*. DOI 10.1007/s40009-016-0465-x.

[5]. Mitali Srivastava, Rakhi Garg, P. K. Mishra., Preprocessing Techniques in Web Usage Mining: ASurvey, *International Journal of Computer Applications, (0975 – 8887) Volume 97 – No.18*, July 2014.

[6]. Michal Munk, Jozef Kapusta, Peter Svec, Data Preprocessing Evaluation for Web Log Mining: Reconstruction of Activities of a Web Visitor, *International Conference on Computational Science*, ICCS 2010.

[7]. Bharat Chauhan, Hemant Kumar, Mihul Singh, Piyush Kumar, Ms. Sakshi Hooda, An Improved Preprocessing and Clustering Using Web Log Data, *International Journal of Advanced Research in Computer and Communication Engineering ISO 3297:2007 Certified, Vol. 5, Issue 11*, November 2016.

[8]. P. Sukumar, L. Robert, S. Yuvaraj, Review on Modern Data Preprocessing Techniques in Web Usage Mining (WUM), *2016 International Conference on Computational Systems and Information Systems for Sustainable Solutions.* 978-1-5090-1022-6/16/$31.00 ©2016 IEEE.

[9]. J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan, Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, *SIGKDD Explorations, 1(2)*, 2000.

[10]. F. Masseglia, D. Tanasa and E.B. Trousse, *Web usage mining: Sequential pattern extraction with a very low support. Advanced Web Technologies and Applications, Lecture Notes in Computer Science, 2004, Vol. 3007, pp. 513–522,* ISBN 978-3-540-21371-0.

[11]. R. Cooley, B. Mobasher and J. Srivastava, Data Preparation for Mining World Wide Web Browsing Patterns, *Knowledge and Information System, 1999, Springer-Verlag, Vol. 1*, ISSN 0219-1377.

[12]. M. Baglioni, U. Ferrara, A. Romei, S. Ruggieri, and F. Turini. Preprocessing and Mining Web Log Data for Web Personalization, *AI*IA 2003, LNAI 2829, pp. 237–249, 2003. Springer-Verlag Berlin Heidelberg* 2003.

[13]. J. Vellingiri and S. Chenthur Pandian, A Novel Technique for Web Log mining with Better Data Cleaning and Transaction Identification, *Journal of Computer Science 7 (5)*: 683-689, 2011 ISSN 1549-3636 2011 Science Publications.

[14]. Surbhi Anand and Rinkle Rani Aggarwal, An Efficient Algorithm for Data Cleaning of Log File using File Extension, *International Journal of Computer Applications (0975 – 888) Volume 48– No.8*, June 2012.

[15]. Wasvand Chandrama, Prof. P.R.Devale, Prof. Ravindra Murumkar, Data Preprocessing Method of Web Usage Mining for Data Cleaning and Identifying User navigational Pattern, *IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 10*, December 2014.

[16]. Erhard Rahm Hong Hai Do, Data Cleaning: Problems and Current Approaches, University of Leipzig, Germany http://dbs.uni-leipzig.de

[17]. B. Prasanna Kumar Reddy and Duvvada Rajeswara Rao, Optimizing Web Log Data to Perceive User Behavior, *Asian Journal of Information Technology 15 (20):* 3899-3904, 2016 ISSN: 1682-3915, Medwell Journals, 2016.

[18]. R. F. Dell (2008), Web user session reconstruction using integer programming, *International Conference on Web Intelligence and Intelligent Agent Technology, IEEE/ACM/WIC, Vol. 1 Page(s): 385 – 388*, 2008.