IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-6, Issue-7, July 2018 E-ISSN: 2347-2693

A Study on Data Preprocessing Methods on Web Log Data in Web Usage Mining

R.Sandrilla[1 *], Dr. M. Savitha Devi[2]

[1] Department of Computer Science, Sacred Heart College (Autonomous) Tirupattur, India
[2] Department of Computer Science , Periyar University College of Arts and Science , Harur. India

Available online at: www.ijcseonline.org

Accepted: 06/Jul/2018, Published: 31/July/2018

Abstract— Web usage Mining is an extension of traditional data mining. As the tremendous amount of data is increasing, the prominence of internet is growing. This impact upholds the user's needs and the users are also increasing in enormous speed. Because of these realities the web data has been budding day to day. Therefore extracting the useful data from WWW has become the challenging one. Due to this fact the users are feeling disoriented. So it is necessary for the web usage miners to discover the new way of finding the desired information or the ease of accessing the web. As a result the web mining has become more popular and reached the peak in research field having in mind about mining the data and WWW as well. The aim of the proposed research is to survey on different Data preprocessing techniques carried out by most of researchers has been discussed, where this web log preparation is considered as the first step on web mining process to identify the user behavior. This phase is referred to be the most important process to ensure the quality of the log data. The log files are gathered and pre-processed by removing the unwanted or irrelevant information. A complete overview on data preprocessing may recommend better technique to find the user behavior and to improve the performance, and finally we concluded by providing a glimpse of various Web mining Applications.

Keywords— Web usage mining, Web server log, Data Preprocessing, User identification, Session Identification

I. INTRODUCTION

Web has become a very popular and interesting platform. Now a day's World Wide Web has a tremendous growth and with this abundant growth it has exceeded to reach all expectations. There are several millions and billions of important and interesting information that is available through internet and this huge volume is still rising up. In this rapid and increased growth of contents or varieties of information searching for the most important and impressive information has become an interesting task and this way of probing the needs is termed as web mining or mining the data. Web mining is more or less similar to traditional data mining, where the data are gathered and cleaned. Web mining is categorized into three aspect and they are Web Content, Web Structure Mining and Web Usage mining. Web content mining is nothing but mining is done based on the data that is available. The useful information is gathered and collected. Web structure Mining involves in structuring the data in the web through hyperlink level. Web Usage Mining is basically known as WUM which is mainly involved in finding or discovering the patterns from the information gathered through the usage of the client and the server or through some of the basic Web amenities. WUM is to analyze the exact patterns and the user behavior in the web by gathering the knowledge through the web log information. The major outcome of WUM is used in Web Personalization, Web Recommendations, Evaluating a Website or a Webpage, Business Usage, improving the performance of the system.

A. Web usage Mining

Web Usage Mining is one of the applications of the traditional data mining. The best Method to discover the interested usage information is done through WUM. The technique used in discovering the web user's pattern of browsing behavior helps to serve the user by providing the exact and relevant data in order to make the search easier and user friendly. In-fact the Content mining and the Structure Mining is done by utilizing the primary information available on the web whereas the WUM is done by utilizing the secondary data that is collected from the client server behavior which is called as the Log files.

II SOURCES OF WEB LOG DATA

In web usage mining the data is collected from various sources. Those web data collected from various sources are called Log files. Log files are nothing but the blogs. The log file used in web usage mining may comprise web data warehouse like: Web Server logs, Proxy server Logs, Browser Logs. An efficient web mining algorithm to mine

web log information proposed to manage the time and space complexity [1].

The Web Server Logs are the one which maintains the history of the web page requests. Web server Logs are maintained with a standardized format like gathering the information as the request given for the web pages, IP Address of the particular client, the time, date, Server port, HTTP request code, user agent, referrer log, and also includes the bytes served. [2] These attributes are gathered collectively and combined as a single unit of log files as Referrer log, Error log and Access log.

The proxy server logs is a mechanism which lies between the web servers and the client browsers .These are easily collected and gathered for the future requirements in order to maintain the load balancing mechanism. The risks of network traffic between the server and the client browsers are reduced with those logs. The web proxy logs are mainly used to find the access behavior or the usage pattern of any anonymous user.

Browser logs are collected from various client machines and browser. These logs collect the client side information which is used in reducing the bot and user session identification problems. The log files are stored in different formats like Internet Information service Log File format (IIS), Common Log File format (NCLF), Extended Log File format (ELF).These web log formats varies depending on the configuration of the web server.

W3C (World Wide Web Consortium) Extended log file format – This is considered as the default log file format by IIS. They use ASCII text format and UTC for the time format. This is termed as the only log file format where the properties can be customized by the user. There is an option to limit the size of the log files and collect the detailed information. The properties collected in the log files can be separated by using spaces.

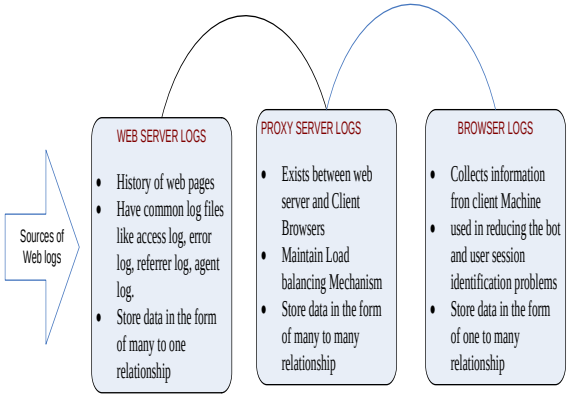

Figure 1. Log files Format

IIS (Microsoft Internet Information **Services) log file format** – Similar to the extended log file format this also uses the ASCII text format and uses fixed number of properties. IIS log file format is usually used when there is no need of the detailed information from the logs. The IIS logs gather more information than NCSA but less than extended W3C format. This log file is also known for comma separation file and uses the local time.

NCSA (National Centre for Supercomputing Applications) **log file format** – This is a limited

Information file format where logs only the basic information. It acts similar to IIS log file format where they use fixed number of properties. Here the time is recorded using the local time and the properties are separated by spaces. NCSA log file format never support the FTP sites. Since the entries are minor with this format, the loading space required for logging is relatively less compared to other formats.

In the following table 3 the Sample log file history with the different formats are listed for the reference

TABLE 1. SAMPLE WEB LOG FILE

Format	Sample Entries
W3C	#Software: Microsoft Internet Information Services 6.0 #Version: 1.0 #Date: 2009-06-11 05:12:03 #Fields: date time s-sitename s-ip cs-method cs-uri-stem cs-uri-query s-port cs-username c-ip cs(User-Agent) sc-status sc-sub status sc-win32-status 2009-06-11 05:12:02 W3SVC1893743816 192.168.1.109 GET / – 4677 – 192.168.1.109 Mozilla/4.0 (compatible;+MSIE+4.01;+Windows+NT;+MS+Search+5.0+Robot) 401 2 2148074254 2009-06-11 05:12:02 W3SVC1893743816 192.168.1.109 GET / – 4677 – 192.168.1.109 Mozilla/4.0+(compatible;+MSIE+4.01;+Windows+NT;+MS+Search+5.0+Robot) 401 2 2148074254
IIS	66.249.79.227 - - [02/Aug/2015:23:14:15 +0530] "GET /robots.txt HTTP/1.1" 200 59 "-" "Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)" 66.249.79.239 - - [02/Aug/2015:23:14:19 +0530] "GET /feed/ HTTP/1.1" 200 48433 "-" "Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)"46.36.67.110 - - [02/Aug/2015:23:21:53 +0530] "POST /xmlrpc.php HTTP/1.1" 200 403 "-" "-"41.222.181.175 - - [02/Aug/2015:23:53:06 +0530] "GET /wp-content/themes/sahifa/images/socialicons/rss.png HTTP/1.1" 200 1729 "http://shctptcs.org/"
NCSA	192.168.1.109 – - [08/Jun/2009:12:11:14 +0200] “GET / HTTP/1.0″ 401 1913 192.168.1.109 – NT+AUTHORITY\LOCAL+SERVICE [08/Jun/2009:12:11:14 +0200] “GET / HTTP/1.0″ 200 336 192.168.1.109 – - [08/Jun/2009:12:11:14 +0200] “POST /_vti_bin/sitedata.asmx HTTP/1.1″ 401 1889 192.168.1.109NT+AUTHORITY\ LOCAL+SERVICE [08/Jun/2009:12:11:14 +0200] “POST /_vti_bin/sitedata.asmx HTTP/1.1″ 200 1331

A. Attributes of Log Files

There exists various attributes for the log files. The following Table 2 Lists the few available attributes of the different Log File Formats. [3]

TABLE 2. DIFFERENT LOG FILE FORMAT

Attributes	**Description**	**IIS**	**NCSA**	**ELF**
Date	Displays the date when the activity took place	✓	✓	✓
Time	Displays the time when activity occurred	✓	✓	✓
Service Name	Displays the Internet service name running on the client server	✓	X	X
Server Name	Displays the name of the server on which log files entries are generated	✓	X	✓
Server IP Address	Display the IP address of the Server on which the Log Files are generated	✓	X	✓
Method(Type Request)	Displays the Type of the requested function. For example POST,GET	✓	✓	✓
URI Query	The query, if any, that the client was trying to perform. A universal Resource Identifier (URI) query is necessary only for dynamic pages.	✓	X	X
Server port	Displays the service port number that is configured to the currently working server	✓	X	X
User Name	Displays the name of the authenticated user accessing the server, The anonymous users are displayed with a hyphen	✓	✓	✓
Client IP Address	Displays the IP address of the Client who made the request.	✓	X	✓
Protocol Version	Displays the HTTP protocol version that the particular client used.	✓	✓	X
User Agent	Displays the type of the browser that the client used.	✓	X	X
Cookie	Displays if there any of the content of the cookie are sent or received.	✓	X	X
Referrer	Displays the particular site that the user last visited. This site provided a link to the current site.	✓	X	X
Host	Displays if there exists any host header name.	✓	✓	X
HTTP Status (Service code)	The HTTP status code. A value of 200 indicates that the request was fulfilled successfully	✓	✓	✓
Protocol Substatus	The sub status error code.	✓	X	X
Bytes Sent-Server	The number of bytes sent by the server.	✓	✓	✓
Bytes Received	The number of bytes received and processed by the server.	✓	X	X
Time Taken	The length of time that the action took, in milliseconds.	✓	X	✓

Client bytes sent	The number of bytes sent by the client	X	X	✓
Parameters	The parameters that are passed to a script.	X	X	✓

III DATA PREPROCESSING

The data preprocessing step is a most rigorous, time consuming and important step in integrating the data in web usage mining

a. *Data Cleaning:* Data cleaning, also called data cleansing or scrubbing, deals with detecting and removing errors and inconsistencies from data in order to improve the quality of data.[4] Data quality plays a very important role as the data quality ensures that the data is in well formed, accurate and can be trusted. The data cleaning approach is used to remediate data quality issues and deliver clean data. To ensure high data quality and to cleanse the incoming data tuple with the external data the authors [1] developed an efficient fuzzy match algorithm for online data cleaning. The data "cleaning" routine involves various tasks such as; data acquirement, filling up the missing data values, unifying date formats, conversion of nominal values to numeric data, identification of outliers and smoothening of noisy data, and rectifying inconsistent data[5]. Data Cleaning is a major step of KDD process in order to recognize any inconsistency and incompleteness in the dataset and to improve its quality [6].

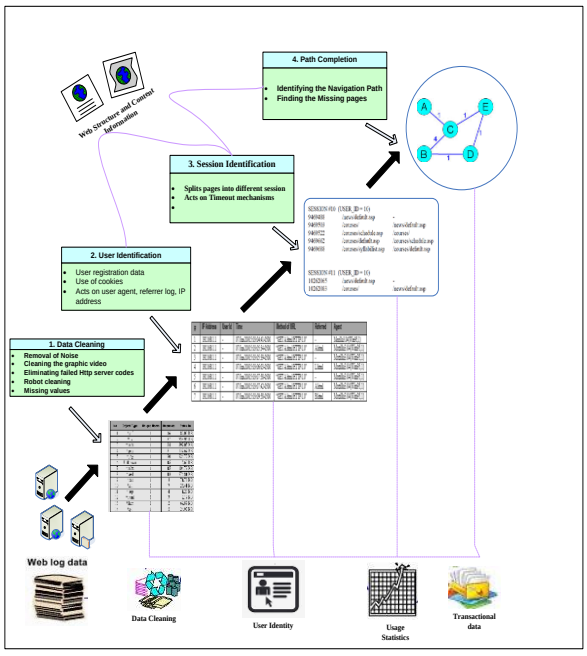

Figure 2. Phases of Data Preprocessing Technique

b. *User Identification:* The major task of User identification is to predict the particular user who accesses the website or the webpage. Here it is necessary to identify and distinguish the different users among the multiple views of the same user. This is considered to be a significant phase in data preprocessing technique. Various techniques and methods have been followed to identify the users. The one simplest and most relevant method in identifying the user is by tracking their IP address . Here if the Ip address is same and the user agent is dissimilar, and then they are assumed is a new user entry. Incase if the User id /user agent and the user IP address are the same then they cross verify with the referrer URL and site topology.

c. *Session Identification:* After the identification of the user, session identification is made. In session identification the timeout mechanism is mainly considered. Here the timeout technique indicates that if the user request for the page and the time exceeds the certain limit then it shows that that the user starts the new session. Session identification divides all the pages accessed by the user into different sessions.

d. *Path Completion:* Path completion is considered to be a significant phase in the steps of preprocessing. In some circumstances the previous phase session identification might miss identifying few important web pages due to the proxy servers and cached versions of pages which are viewed by user through "BACK" button. In those cases path completion step is carried in order to recognize the missing pages. Path completion refers to the inclusion of important page accesses that are missing in the access log due to browser and proxy server caching [7]

IV LITERATURE SURVEY

This section provides a few discussions on several Data preparation techniques.

Dafa-Alla, Mirghani. A. Eltahir and Anour F.A [8] labels the complete techniques of preprocessing that are used for removing the unclean data, data filtering, User identification, session identification and web session Clustering. The various web server logs from different sources were also described. They also proposed a new algorithm to support the preprocessing phase.

Tasawar *et al.,* [9] suggested a Data preprocessing method based on hierarchical clustering. The web session clustering is done to categorize the web users according to their navigation and similarity measure. In this paper the author proposed a framework for web session clustering in the data preprocessing level of WUM. The framework suggested converts the web log data to the numerical data and results a session vector in order to perform web session clustering. This proposed Hierarchical clustering will improve web session clustering to user identification session

Theint Theint Aye [10] proposed a technique for data preprocessing by extracting the fields and data cleaning. The major task worked on this paper is to clean the Web log Data by removing the unwanted information and inserting the cleaned data into the Database for further mining. The web Log files is a textual data that is not comma separated. So they provided a field extraction algorithm to separate the web log files from a single line of text. Since the web log files are collected and stored from different web servers each server might use its own separator's like comma, colon etc., The proposed Field extraction algorithm separate's the fields and stores in the relational database in the form of relational tables for ease of mining. They also provided data cleaning algorithm that is used to clean the unwanted and irrelevant data item in the explored file. The log file contains the record of .gif, .jpeg, and failed HTTP status codes. Inconsistency is deleted by clearing up of unwanted data. This algorithm keeps the log table as an input which is done by the Field Extraction Algorithm. The algorithm reads each record in the log table and removes the irrelevant data prefixed on to it.

Jia Li [4] in his paper explained that Data cleansing is the first stage of the data mining process. The data preprocessing work is done by low level merger work in order to combine multiple integrated referrer logs. This paper proposed a provision where the log entries are divided into logical clusters through one or multiple series of transaction identification modules to reduce the size of the original document. This identification module is considered as user identification to identify each specific user by users IP address. This is done to analyses the behavior of the user in the web.

Chandana S.Khatavkar, Prof. Mangesh Wanjari [11] defined that the conversion of usage content and structure information carried from various data sources are constructed for the necessary pattern recognition by removing the unrelated data is known as preprocessing methodology. In this paper they proposed a cluster optimization technique using ANT colony optimization to eliminate the irrelevant data. The proposed model uses Ant colony algorithm for clustering based on the user's session. Here the user with the similar access pattern fall on one and form the clusters. They also analyzed the user navigational pattern cluster with Fuzzy interface System.

Abdul Rahaman Wahab Sait, and Dr.T.Meyappan [12] suggested that data preprocess is the vital process in web mining. This paper focused on preprocessing of web log data and transforms into a numerical form to generate patterns. The transformation of a numerical format hides the user details and they also proposed that the transformation of numerical format will reduce the complexity of generation of the pattern.

Uma Maheshwari, Dr.P.Sumathi [13] experimented about the data preprocessing and clustering the web log. The data preprocessing treatment has been analyzed and they undergone various steps to remove the unrelated records. In this paper they mainly focused on data cleaning and session identification process which is used in appending the lost pages and construction of transactions in preprocessing stage. The records are cleaned effectively by removing the robot entries. MFR and RL algorithms were used along with time window concept to search the content pages and to be more effective and to give a reliable input for the further tasks.

Nithya P, Dr.P.Sumathi [14] proposed a novel preprocessing technique to remove the global and local noise and web robots. The research has presented an organized way to carry out cleansing process. They implemented two different techniques to identify the web robots that help in the accurate detection of user's interested patterns by providing the relevant web logs. Various steps were done such as eliminating of local and global noise,
Removal of graphics and videos. Removal of records with failed HTTP status codes and robot cleaning as a removal process in Data cleaning. This paper continues the research on web access log analysis to analyze the patterns of website usage features of user's behavior in the web.

Sheetal A. Raiyani, Shailendra Jain [15] presented a new technique for preprocessing a weblog data and for identifying a unique user and sessions from the data. They proposed a distinct user identification algorithm with the time complexity to find the unique users and they also present the extra cleaning steps such as removing maintenance pages, redundant pages and groups the sessions using the similar session length to clean the outliers of the data

Ramya C, Kavitha G [16] proposed a complete preprocessing methodology for discovering patterns in web mining process to improve the quality of data by reducing the quantity of data .In this paper they proposed a dynamic ART neural network clustering algorithm to group the users according to their behavioral pattern in the web. They also proposed a neat architecture to carry out the methodology. This

proposed methodology reduces the size of the web log files and produces the quality clusters.

V. Chitraa, Dr. A.S. Davamani [17] agreed on data cleansing process that comprises of deleting all the information that includes images and extensions in URL name. They also suggested that different IP address represents different users. If the two entries have the same IP address from different agents are considered to be different users. They considered the third step in preprocessing is session identification and they suggested two different methods which are time oriented and navigation oriented heuristics. In the final path completion step they considered three approaches i.e., reference length approach, maximal forward reference, Time window.

Surbhi Anand and Rinkle Rani Aggarwal [18] stated that Preprocessing enables to translate the unprocessed data which is composed from server log files into constructive data abstraction. This paper emphasis on web usage mining process and makes an exploration in the field of data cleaning. They proposed two different algorithms to separate the web log entries and to clean the data by removing the outliers. The web log entries are separated out by the process of separating different data field from single server entry which is identified as data field extraction. The second algorithm is to describe the data storage of extracted field from web log file using the previous algorithm.

Cooley et al. [19] have proposed techniques for data cleaning, user identification, session identification and transaction identification. In this paper the User Identification is done by identifying pages which are accessed in the website. Even though their heuristic techniques are well completed but some heuristics are not appropriate for complex web sites.

Castellano et al. [20] developed a tool LODAP (Log Data Pre-processor) which proceeds with stored log file as input and provides statistical analysis and user sessions as output. The tool developed was divided into three modules: Data cleaning module, Data structuration module and Data filtering module. In the first Data cleansing module multimedia files, status code, and robot's request from log files are removed. The second module is Data structuration module in which the users are identified by authentication data/IP address and the sessions are identified by time based techniques. The maximum elapsed value for session identification has set to 30 minutes and minimum to 2 seconds between two consecutive requests. Further the final module is the Data filtering. Here the most requested pages are clustered and least requested pages are plunged out based on threshold value. Except the path completion the authors have accomplished all the steps of data preprocessing. For the effective discovery more attributes of log files is included along with the IP address in the user identification.

G.T.Raju, Nandhini.N [21] proposed a comprehensive preprocessing methodology as a prerequisite and has taken four steps in r prefetching the application: Data Cleaning, Identification of users & Sessions, and finally the Data Formatting and Summarization. In this paper the authors have made an attempt to reduce the quantity of the WUD and thereby improve the quality of WUD for effective use in Prefetching application. They also proposed several heuristics for cleaning the WUD which is then aggregated and recorded in the relational data model. They also conducted several experiments on To validate the efficiency of the preprocessing methodology, several experiments were conducted and also proved that their methodology reduces the Web access log files to the initial size and offer richer logs that are structured for application in Prefetching.

Navin Kumar Tyagi, A.K. Solanki & Sanjay Tyagi [22] presented an algorithmic approach to Data preprocessing in Web Usage mining. The algorithm is proposed for data cleansing and data reduction by suggesting the heuristic based navigational behavior to separate robot sessions from actual sessions. They also insisted the importance of data preprocessing before applying Data mining techniques to discover the user access patterns from web logs which processes quality of Data.

Suneetha K.R, Dr. R.Krishnamoorthi [23] proposed an algorithm for Data cleaning, user identification, and session identification. The proposed algorithm is applied on each block of section in order to reduce the size of the actual data, to obtain the unique users and to find the navigational behavior of the user using the session key with start time and end time. They also presented a new approach to access the usage pattern of preprocessed data where the results of preprocessed web server logs were stored using snow flake schema of data warehouse to facilitate easy retrieval and analysis.

C.P. Sumathi, R. Padmaja Valli, T. Santhanam [24] stated that the data preprocessing is predominantly the significant stage in Web usage mining due to the characteristics of Web data and its association to other related data collected from multiple sources. They also stated that the preprocessing phase is the most time-consuming and computationally intensive step in Web usage mining, and the process is critical to the success of Pattern discovery and Pattern Analysis. In this paper the preprocessing process deals with the conversion of raw Web server logs into a formatted user session file in order to perform Web usage mining. They also focused on the Preprocessing of Web usage data from Web log files resulting in a user session file and Formatting of the user session file suitable for mining tasks.

Abdul Rahaman Wahab, Dr.T.Meyappan [25] proposed a preprocessing and transformation techniques to generate patterns from web log files. They mainly focused on processing the web log files to transform into a numerical form to generate a pattern. The Data transformation violates the privacy of the Users by hiding the user details. In their research the web log is divided according to the types of users to know the exact browsing pattern. They focused to give an insight to generate useful pattern for the improvement of the website.

Dharmendra Patel, Dr. Kalpesh Parikh, Atul Patel [2] focused on more complex part of data preprocessing that is sessionization. They covered the important aspect of sessionization stage which is used to identify the behavior of user and that information is very crucial for many applications. The author depicts the different strategies of sessionization. This paper deals with many software tools available in market for the generations of sessions form raw log data. They also dealt with the major problem and the solution related to the problem which is arising in the phase of sessionization.

Naga Lakshmi, Raja Sekhara Rao, Sai Satyanarayana Reddy [26] has provided the details of data preprocessing steps that are essential for performing Web Usage Analysis. They also presented different formats of web server log files and the preprocessing is done on those files for web usage analysis. This paper deals with cleaning the irrelevant or redundant information like image, video and sound files. They also concentrated on removing the HTTP errors, records created by spiders, crawlers and robots.

Mona S. Kamat, J. W. Bakal & Madhu Nashipudi [27] proposed many data preparation techniques to clean the data and to identify users and the sessions. The authors used an optimal algorithms to generate the user session sequences using data structures represented by two way hash table. Since the hash structure used in storing user session sequence, backward referencing is done for each search, instead of searching the entire pages. Hence they concluded that the backward referencing takes much lesser time and also identifies session with higher precision

Dr. Girish S. Katkar, Amit Dipchandji Kasliwal [23] proposed a predictive analysis for data mining using the web log files. In this paper the author used the third party application developed in JAVA and WEKA application for preprocessing representing the association rule. They also stated that the user
Log files are most important for the predictive analysis data to predict the future patterns.

B N ShankarGowda, Vibha Lakshmikantha, K R Venugopal, L M Patnaik [28] proposed an efficient methodology for analyzing user behavior. The methodology builds DWH for evaluating the user behavioral pattern and integrated those with the data mining framework. They suggested that a good data is obtained through preprocessing the web log files to enhance high performance, so the data set is prepared depending on the analysis and those datasets were transformed and aggregated to different level of abstraction and further obtained the better results in the reduction of the size of servers.

Vijay Kumar Padala1, Sayeed Yasin2, Durga Bhavani Alanka3 [29] concerned about web log mining. They introduced a novel method for data cleaning and session identification. The Authors focused on web logs and proposed an efficient algorithm for cleaning the web log file, and user identification. In order to reduce the log files and to provide the quality in available data preprocessing is done. In this paper the authors concentrated on cleaned the web log file by removing the noisy as well as unnecessary data and inserted the processed data onto the relational database. The cleaning is done by removing the log entry nodes that contains the file extension like jpg, gif which means remove request such as multimedia files, image, and page style file. Likewise the user identification is done by identifying the individual user by their IP address. The proposed algorithm provides information about total number of individual users, users IP address, browser used and user agent.

Shashi Sahu1, Leena Sahu [30] dealt with the process of detecting and correcting the irrelevant and incomplete data from the datasets. The authors provided a complete summary of log cleaner which filters out plenty of inconsistent data based on their URLs to provide the quality and efficiency of web log. They also discussed on various methodologies like Two-level clustering method, Noise Detector as an efficient technique, Community Detection Technique, Effective and scalable technique and EPLogCleaner filtering method available for web usage mining for cleansing the server log.

Xiaohua Hu [31] presented an algorithm DB-HReduction which is implemented for preprocessing step. This method discretizes or eliminates numeric attributes and generalizes or eliminates symbolic attributes very efficiently and effectively. The data reduction is done by merging identical tuple after substituting an attribute value by its higher value in a predefined concept hierarchy for symbolic attributes, or the discretization of continuous (or numeric) attributes, or the removal of insignificant or
Irrelevant numeric and symbolic attributes. The proposed algorithm greatly decreases the number of attributes and tuple of the data set and improves the accuracy and decreases the running time of the data mining algorithms in the later stage.

V. MOTIVATION AND APPLICATIONS OF WEB USAGE MINING.

There are numerous issues to challenge the preprocessing of web log files. Collection of the various request in a single block called Log files and analyzing them is the first major impact for preprocessing. This phase is considered to be the great challenge for analyzing the data and predicting the user's behavior. The preprocessing of web log data is done to influence the quality data. In future this quality data can be further transformed to discover patters using various mining algorithm. The mining algorithms help in predicting the patterns of user behavior and their interest in order to help the operational strategies of various enterprises. By understanding the navigational behavior of the user and discovering the patterns are effectively used to build various Web Usage Mining applications like:

- Improving the Web site.
- Modification of Web site design.
- Web Personalization.
- Web recommendations.
- Fraud Detections.
- Future prediction.
- E-Commerce.
- Website Evaluation.
- Improving the performance of the Web server.

VI. CONCLUSION

Data preprocessing is a significant and essential phase in web usage mining. It has been carried out with various tasks such as cleaning, user identification, session identification, path completion. In the phase of cleaning the data all the needless and flawed record is removed. Then user identification is done using Unique Internet address associated with agent mechanism. User Session identification is done either by using Time oriented or navigation oriented heuristic. So it was tried to provide a complete survey on preprocessing and mainly focused in each individual phase of preprocessing performed by various research communities. In this study we found that though the data preprocessing strategy comprises of various phases they all are interconnected with a similar approach in implementation and yet few areas are still in its infancy. Though various research issues has been done in session identification, user identification, it has to be much exploded in few areas such as visual Web Mining, where the preprocessing work has to be carried out in visual contents. This paper also has been motivated by providing a glimpse of knowledge on Web usage mining application. From the survey it is clear to enhance the preprocessing task and further continue to do a better and quality research in this field to make valuable resource for future.

REFERENCE

[1] R.Shanthil, Dr.S.P.Rajagopal, "*An Efficient Web Mining Algorithm to Mine Web Log Information*"

[2] Dharmendra Patel, Dr. Kalpesh Parikh, 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, Issue 1, Jan-Feb. 2012, pp. 327-330.

[3] http://www.surfray.com/blog/2009/08/11/iis-log-file-formats-overview/

[4] Jia Li (2013), "*Research of Analysis of User Behavior Based on Web Log*", International Conference on Computational and Information Sciences. 2013

[5] Jiawei Han et al, "*Data mining, concept and techniques*" .cs.sfu.ca, 2, Jan. 31, 2011. [Online]. Available: http://www.cs.sfu.ca.

[6] J. Han, M. Kamber and J. Pei, "*Data mining: concepts and techniques*", Morgan Kaufmann, (2006).

[7] C.P.Sumathi, R.PadmajaValli, Santhanam," An Overview Of Preprocessing of Web Log Files For Web Usage Mining" Journal of Theoretical and Applied Information Technology 31st December 2011. Vol. 34 No.2

[8] Dafa-Alla, Mirghani. A. Eltahir and Anour F.A(2013)," *Extracting Knowledge from Web Server Logs Using Web Usage Mining*", 2013 international conference on computing, electrical and electronic engineering (ICCEEE)

[9] Tasawar Hussain, dr.asghar and dr. Masood" *preprocessing techniques in web log mining*" 2010

[10] Theint Theint Aye "*Web Log Cleaning for Mining of Web Usage Patterns*", University of Computer Studies, Mandalay 2011.

[11] Chandana S. Khatavkar, Prof. Mangesh Wanjari, "A Hybrid approach for Clustering Weblog", International Journal of Advanced Research in Computer Science and Software Engineering. Volume 5, Issue 3, March 2015.

[12] Abdul Rahaman Wahab Sait and Dr. T.Meyappan "*Data preprocessing and Transformation techniques to generate Patterns from Web Logs*", International conference on Computer Science and Information Systems (ICSIS'2014) Oct 2014.

[13] B.Uma Maheshwari, 2P.Sumathi," *An Effective Method to Preprocess the Data in Web Usage Mining*", ARPN Journal of Science and Technology, Vol. 3, NO. 3, March 2013

[14] P.Nithya, 2 Dr. Sumathi," *A Survey on Web Usage Mining: Theory and Applications*", P Nithya et al, Int.J.Computer Technology & Applications, Vol 3 (4), 1625-1629.

[15] Sheetal A. Raiyani1, Shailendra Jain2, Ashwin G. Raiyani3." *Advanced Preprocessing using Distinct User Identification in web log usage data*", International Journal of Advanced Research in Computer and Communication Engineering Vol. 1, Issue 6, August 2012.

[16] Ramya C., Shreedhara K. S., and Kavitha G." *Preprocessing: A Prerequisite for Discovering Patterns in Web Usage Mining Process*", International Journal of Information and Electronics Engineering, Vol. 3, No. 2, March 2011.

[17] Chitraa, Dr. Antony Selvadoss Davamani," *An Efficient Path Completion Technique for web log mining*", IEEE International Conference on Computational Intelligence and Computing Research, 2010

[18] Surbhi Anand, Rinkle Rani Aggarwal," *An Efficient Algorithm for Data Cleaning of Log File using File*

Extensions", International Journal of Computer Applications (0975 – 888) Volume 48– No.8, June 2012.
[19] R. Cooley, B. Mobasher, J. Srivastava , "*Data Preparation for Mining World Wide Web Browsing Pattern*" in Journal of Knowledge and Data Engineering Workshop, IEEE, 1999Vol.1 Page(s): 5-32
[20] G. Castellano, A. M. Fanelli, M. A. Torsello," *Log Data Preparation for Mining Web Usage Patterns*" IADIS International Conference Applied Computing 2007
[21] G T Raju, Nandini N," *Preprocessing of Web Usage Data for Application in Prefetching to Reduce Web Latency*", International Journal of Electrical& Computer Sciences IJECS-IJENS Vol: 14 No: 04
[22] Navin Kumar Tyagi1, A.K. Solanki2& Sanjay Tyagi3," *An algorithmic approach to data preprocessing in Web Usage Mining*", International Journal of Information Technology and Knowledge Management July-December 2010, Volume 2, No. 2, pp. 279-283
[23] K. R. Suneetha, Dr. R. Krishnamoorthi," Identifying User Behavior by Analyzing Web Server Access Log File", IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.4, April 2009.
[24] Abdul Rahaman Wahab Sait, and Dr.T.Meyappan," Data Preprocessing and Transformation Technique to Generate Pattern from the Web Log", International conference on Computer Science and Information Systems (ICSIS'2014) Oct 17-18, 2014.
[25] Naga Lakshmi, Raja Sekhara Rao, Sai Satyanarayana Reddy,"An Overview of Preprocessing on Web Log Data for Web Usage Analysis" International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-2, Issue-4, March 2013.
[26] Mona S.Kamat, J.W.Bakal & Madhu Nashipudi," *Optimization of Web Preprocessing in Web Usage Mining*", Volume-2, Issue-6, 2013.
[27] Dr. Girish S. Katkar," *Use of Log Data for Predictive Analytics through Data Mining*", Current Trends in Technology and Science Volume: 3, Issue: 3 (Apr-May. 2014)
[28] B N ShankarGowda, Vibha Lakshmikanthab, K R Venugopal, L M Patnaikd," *A Framework for Preprocessing Web Log in the Data Warehouse Environment for Web User Behavior Analytics*" International Journal of Information Processing, 8(1), 40-52, 2014IK International Publishing House Pvt. Ltd., New Delhi, India.
[29] Vijay Kumar Padala1, Sayeed Yasin2, Durga Bhavani Alanka3," *A Novel Method for Data Cleaning and User-Session Identification for Web Mining*", International Journal of Modern Engineering Research (IJMER), Vol. 3, Issue. 5, Sep - Oct. 2013 pp-2816-2819.
[30] Shashi Sahu1,Leena Sahu2 "*A Survey on Frequent Web Page Mining with Improving Data Quality of Log Cleaner",* International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 4 Issue 3 , March 2015
[31] Xiaohua Hu," DB-HReduction: *A Data Preprocessing Algorithm for Data Mining Applications*", 0893-9659/03/$ - see front matter c° 2003 Elsevier Science Ltd. All rights reserved.

Authors Profile

Mrs. R.Sandrilla pursed Bachelor of Science from Auxilium College, Vellore, India in 2005 and Master of Computer Applications from Mount Carmel College, Bangalore in year 2008. She is currently pursuing Ph.D. and currently working as Assistant Professor in Department of Computer Sciences, Sacred Heart College, Tirupattur since 2010. Her main research work focuses on Web Mining, Big Data Analytics, Data Mining, IoT. She has 8 years of teaching experience and 3 years of Research Experience.

Mrs. Savitha Devi pursed her Ph.D. in Mother Theresa University, Kodaikanal,India and currently working as Assistant Professor and Head in Department of Computer Science, Periyar Constituent college of Arts and Science Dharmapuri. She has published more than 10 research papers in reputed international Journals. Her main research work focuses on Cryptography Algorithms, Network Security, Cloud Security and Privacy, she has 10 years of teaching experience and 6 years of Research Experience.

