International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 04, April-2015

A Review Analysis of Preprocessing Techniques in Web usage Mining

Mr. Jitendra B. Upadhyay
Assistant Professor
Shrimad Rajchandra Institute of Management and Computer Application, UTU
Bardoli, India

Dr. S. V. Patel
Professor
Department of Computer Science,
Veer Narmad South Gujarat University
Surat, India

***Abstract*— The Internet web has become popular tool to assist human for their information needs from web server. Due to increasing number of users for web access day by day, there is a need to analyze behavior of such user, in order to monitor and improve performance and throughput of website. Web usage mining is one of the data mining applications which deal with web log files and extract useful information from web. There are different phases are for web usage mining: Data preprocessing, discover pattern and pattern analysis. Among them data preprocessing is the most crucial phase of web usage mining because without good quality of data it is difficult to identify pattern of users behavior. This paper provides reviews of different data preprocessing methods like data collection, data cleaning, User identification, session identification and path completion which will be useful for the community to select one or combination of available techniques in order to carry out efficient preprocessing in order to obtain reliable data mining outcome.**

***Keywords*— Preprocessing, Web Usage Mining, Web Logs**

I. INTRODUCTION

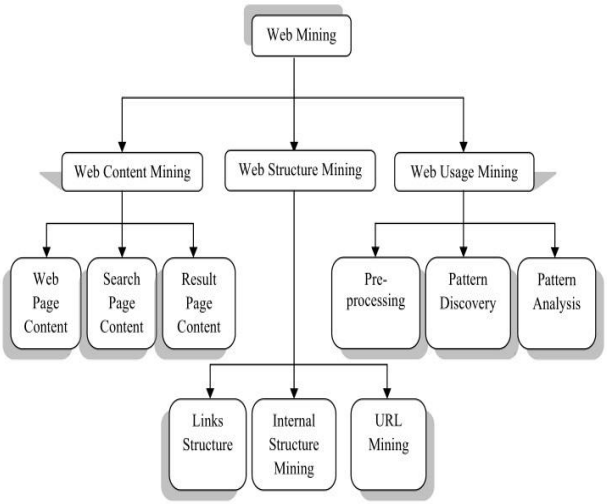
Nowadays, World Wide Web is main source to find any type of world entity information. The Web continues to grow on high speed rate as information doorway. As there is huge amount of structured, semi-structured, unstructured, distributed, static and dynamic data available on web pages of web, it is difficult task to access relevant information with speed. Web mining has emerged as most effective techniques to overcome above problem [5]. Web mining is the application of data mining techniques to discover interesting and potentially useful patterns and implicit information from the Web. Web Mining applies data mining, the artificial intelligence and so on to the web data and extracts the user's using pattern [8]. Based on different mining objects of web sites there are three main categories for knowledge discovery: Web content mining, Web structure mining and Web usage mining.

A. *Web content mining* is process of extracting knowledge from the scanning and mining text, pictures, graphs, etc of web pages to indentify relevant contents.

B. *Web structure mining* is the process of inferring knowledge to identify relationship between web pages in the web using structuring of code, links between references and organization of World Wide Web.

C. *Web Usage Mining*

Web Usage mining is process to extract useful and interesting patterns from log files which show the usage behavior of the users. Web Usage Mining is application of data mining techniques to identify user's usage or behavior pattern from web data which can be used to personalize the Web or to enhance the quality of commerce services or to improve the web structure and web server performance [10]. Web usage mining consists of major three tasks: Data Preprocessing, Pattern Discovery and Pattern Analysis



[Figure [1]: Types of Web Mining]

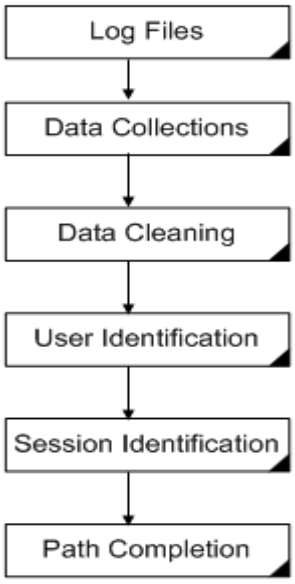
As doing mining process need data collection which are available through log files. [To mine the web data, log files are used.] Log file is location where recoded entry of requests occurred by user on website. Basically log files are located from[1] three different sources like web server, proxy server and browser. Log files contain each record of any user visit or any single click on web site so log files having huge amount of data. Due to large amount of data in web log file, it may have irrelevant data, so original log file can't be used directly for pattern discover. By data fusion, data cleaning, user identification, session identification and path completion the information in the web log can be used as transaction file for discover pattern. Pattern discovery is process to extract user action pattern by using different data mining algorithms or statistics, machine learning or pattern recognized[18], After the preprocessing and pattern discovery, the obtained usage patterns are analyzed to filter uninteresting information and extract the useful information.

Among these phases data preprocessing is the most crucial phase of web usage mining because without good quality of data it is difficult to identify pattern of users behavior. This paper provides reviews of different data preprocessing methods like data collection, data cleaning, User identification, session identification and path completion which will be useful for the community to select one or combination of available techniques in order to carry out efficient preprocessing in order to obtain reliable data mining outcome.

The rest of the paper is organized as follows. Section 2 describes the data preprocessing including it techniques. Section 3 details the literature reviews of related work. Section 4 compares various preprocessing techniques used by researchers and conclusion is given in section 5.

II. DATA PREPROCESSING

Three main stages of web usage mining are Data Preprocessing, Pattern discovery and Pattern analysis. Among them data preprocessing is essential stage which can done by data collection, data cleaning, user identification, session identification, path completion, transaction identification and formatting [5]. Because of log file contains bulk of data, preprocessing of log file reformats the entries of a log file into a form that can be used directly by subsequent steps of the log analyzer [7].



[Figure [2] : Data Preprocessing Techniques]

A. *Data Collection*

Generally, huge number of records is inserted in web log files at server which may be created different log files day wise or month wise. So in beginning of the data preprocessing records of all log files are gathered into one log file [17]. Figure 3 is sample log file data.

B. *Web Logs*

When any user request a particular page on website, an entry is entered into one file of server is called log file. It is consider as reliable source to predict user's behavior. There are mainly three data source for log file: Web Server, Proxy server and Client browser.

I) Web Server is most common source of data. There are four types of server logs [3]:
1. Access log file
2. Error log file
3. Agent log file
4. Referred log file

Access log file contains all the information that provides to the client by the server. Error log file contains a list of any server error. Agent log file provide the information about user's browser, operating system and version of browser. Referred log file is used to allow websites and web servers to identify where people are visiting them from, for promotional or security purpose. Amongst its first two file are very common and useful to fetch the required information of user behaviors. The server log does not contain cached page visited record. The cached pages are called from local storage of browser or proxy server. Different web server having different format of log files like Common log format, IIS standard/ extended log file, Combined/Extended common log format, Log markup language, etc. The following figure [3] shows server log file sample.

```
#Software: Microsoft Internet Information Services 7.0
#Version: 1.0
#Date: 2014-09-01 00:00:15
#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-substatus sc-win32-status
2014-09-01 00:00:15 W3SVC479003068 103.241.244.48 HEAD / - 80
- 54.226.13.24 - 200 0 0
2014-09-01 00:00:49 W3SVC479003068 103.241.244.48 HEAD / - 80
- 46.137.229.25 - 200 0 0
2014-09-01 00:01:17 W3SVC479003068 103.241.244.48 HEAD / - 80
- 54.226.13.24 - 200 0 0
2014-09-01 00:01:34 W3SVC479003068 103.241.244.48 GET
/images2011/QDRDA1_small.JPG - 80 - 5.255.253.195
Mozilla/5.0+(compatible;+YandexImages/3.0;++http://yandex.com/
bots) 304 0 0
2014-09-01 00:01:44 W3SVC479003068 103.241.244.48 GET
/images2012/ERW1.JPG - 80 - 66.249.64.224 Googlebot-Image/1.0
200 0 0
2014-09-01 00:02:17 W3SVC479003068 103.241.244.48 HEAD / - 80
- 54.226.13.24 - 200 0 0
2014-09-01 00:02:47 W3SVC479003068 103.241.244.48 GET
/robots.txt - 80 - 207.46.13.88
```

[Figure [3] : Server Log file Sample from srimca.edu.in website]

These server log files contains various fields like date, time, remote server name and IP address, mode of request, URL of request done by clients, server port number, client IP address, user agents and different status.

II) Proxy Server Log: A proxy server is a server which acts as an intermediate between the client and web servers. Therefore if the server gets a request of the client through the proxy server. So the client's request entries are maintained in separate log file at proxy server.
[Figure of Proxy Server]

III) Client/Browser Log: This type of file can be made to reside in the client's browser itself. Client side logs are useful to handle web page caching, session reconstruction problems. HTTP cookies could also be used for this purpose.

C. *Data Cleaning*

The data cleaning is process that removing irrelevant information/fields/records that are not required for mining. Figure 1 show steps for data cleaning. To analysis of huge amount of data from file which may be irrelevant is a cumbersome activity. So cleaning is necessary at initial stage. If a user requests a specific images pages like .gif, .JPEG, etc. are downloaded which are not useful for further analysis are eliminated. If there is unsuccessful http

request, need to be removing from log files. Automated programs like web robots, spiders and crawlers are also to be removed from log files.

D. User Identification

User identification means indentifying individual users by observing their IP address. There are following rules to identify unique users: 1) If there is new IP address then there is a new client; 2) If the IP Address is same but the operating system or browsing software are different, a reasonable assumption is that each different agent type for an IP address represents a different client; 3) While the IP address, operating system and browsers are all the same , the client can be a new one by identifying whether the requesting page can be reached by the pages accessed before, according to the topology of the site.

E. Session Identification

This is time duration of user's spent in web page. This can be done by nothing down the user id those who have visited the web page and had traversed through the links of the web page.

F. Path Completion

After completion of session identification need to verify for any missing web pages of traveling by user. Path completion is process to the addition of important page accesses that are missing in the access log due to browser and proxy server caching [13]. After the path completion process, the user session file results in paths consisting of a collection of page references including repeat page accesses made by a user.

III. LITERATURE REVIEW

Sanjay Bapu Thakare et al.[2] described the effective and complete preprocessing of access stream before actual mining process can be implemented. They suggested Improvised merging algorithm for data preprocessing method for margining process. They did implementation of field's extraction in core Java. They also suggested improvised TransLog algorithm for convert log file in database. They did not discuss about user identification and more processes of web usage mining. For implementation they used IIS log file format.

Amit Dipchandji Kasliwal et al. [9] proposed a web usage mining methods using well known tool RapidMiner for predicting access behavior. They were taking a log file from KDD repository and performed web usage mining techniques. Using RapidMiner tool they did data preprocessing methods and obtaining ARFF file and applying association rule to ARFF file using MATLAB identify specific result of web pages visits.

Navin Tyagi et al. [21] surveyed about the data preprocessing activities like data cleaning, data reduction and related algorithms. They presented algorithms for data cleaning and data reduction based on CERN (Common Log Format) log file format. About further procedures of preprocessing did not mention.

K. R . Suneetha et al. [23] focused on data preprocessing techniques including web log structure, data cleaning and user identification. They used data from NASA web server log files. To improve efficiency of log files they remove unnecessary data from web log files and analyze process including generating various reports. They also did analysis on most system error found during visit of website. They did not apply any data mining algorithms for pattern discovery.

Jaideep Srivastava et al. [22] provides a detailed taxonomy of the work in area of web mining, including research area as well as commercial offerings. They provide up-to-date survey of the existing work is also provided for Web Usage Mining Research project and products. They provides idea of web usage mining applications likes personalization, system improvements, site modification, Business Intelligence and user characterization. In added in work they provides overview of WebSHIFT system which is designed to perform Web usage Mining from server logs in the extended NSCA format.

V. Chitraa et al. [6] proposed a new technique for identifying sessions for extraction of user patterns. Their experimental results show that the proposed Session Identification technique is an effective one to construct sessions accurately. In their proposed method a matrix is constructed from which sessions are identified using MATLAB tool. They proposed session construction algorithm based on browsing time. They mainly focus on preprocessing web data using data cleaning, user identification and session identification.

Sheetal A. Raiyani et. al. [25] Introduced proposed Technique DUI (Distinct User Identification) based on IP address, Agent, Referred pages on desired session time. They discussed all methods of preprocessing including web log format. They used R K University's library web log of 8 months from Jan1, 2012 to Aug 3, 2012 and implemented preprocessing techniques. In result they got different distinct user's number on their month wise data.

Vellingiri J. et al.[26] focuses on providing techniques for better data cleaning and transaction identification from the web log. They used data preprocessing methods including data cleaning by remove unnecessary data, robot cleaning; user identification using reference length, where reference length is the time taken by the user to view a particular page; session identification, path completion and transaction identification using reference length. They focused on two algorithms one is Maximal Forward References (MFR) and Reference Length (RL). Using these two algorithm author helps to determine only the relevant logs that the user is interested in.

G. T Raju et al.[27] proposed a complete preprocessing methodology that allows the analyst to transform any collection of web server log files into structured collection of tables in relational database model. They compared preprocessing techniques which used by researchers including data source, data cleaning and data formatting and structuring. They showed experiment results likes reducing size of log file for preprocessing, day wise unique visitors, user session identifications.

P. Nithya et al.[28] proposed a novel pre-processing technique by removing local and global noise and web robots. They implemented data cleaning phase will helps in determining only the relevant logs that the user is interested in. Anonymous Microsoft Web Dataset and MSNBC.com. They did not mention other preprocessing methods like user identification, session identification and path completion.

Vellingiri J. et al.[7]provided three phases of web usage mining for user navigation discovery including preprocessing phase, Identify user's behavior and classifications of user behaviors. In the preprocessing phase, the data cleaning process includes removal of graphics, video, status code and robots cleaning. In the second phase, design a set of clusters using Weighted Fuzzy-Possibilistic C-Means (WFPCM), which consists of "similar" data items based on the user behavior and navigation patterns for the use of pattern discovery. In the third phase, classification of the user behavior is carried out for the purpose of analyzing the user behavior using Adaptive Neuro-Fuzzy Inference System with Subtractive Algorithm (ANFIS-SA).

Ashwin Riyani et al.[8]focused on a complete preprocessing style having data cleaning, user and session Identification activities to improve the quality of data. They introduced proposed technique (algorithm) DUI (Distinct User Identification) based on IP address, Agent and Session time, Referred pages on desired session time. They did not use pattern discovery and pattern analysis methods.

Renáta Iváncsy et al.[10]provided novel approach that uses a complex cookie-based method to identify web users. They developed an implementation called Web Activity Tracking (WAT) system that aims at a more precise distinction of web users based on log data. Using WAT they presented different useful result.

Arvind Dangi et al.[11] proposed a new method for web data preprocessing in which it has three phases. In the first phase some websites are selected and by different locations access these website & by applying the (java) tools & methods then find out the IP address of that websites, session usage time & navigations, in the final phase combine them as framework which may be helps to investigate the web user usage behavior.

S. Prince Mary et al.[12] describes the preprocessing methods and steps involved in retrieving the required information effectively. They included data collection, data cleaning by local and global noise removal & graphics records, HTTP failure status etc, User and session identification with path completion.

Shaily Langhnoja et al.[17] gives detailed description of how pre-processing is done on web log file and after that it is sent to next stages of web usage mining. They Used algorithms of data cleaning, user and session identification on web log files and showed results.

Zidrina Pabarskaite [19] presented two new techniques for enhancing web log mining. First is novel framework for performing advanced web log data cleaning and second is data mining is visualization.

C. E. Dinuca [24] propose a new method for identifying sessions based on average time of visiting web pages based on the use of fixed values cause errors in identifying sessions. They implemented in Java programming language by using NetBeans IDE and used two algorithms to identify sessions including 30 minutes to indicate end of session and average time spent on page by users. They showed result and conclude that complexity of classify algorithm is not modified by new approach.

R. Suguna et al.[29] discusses the basics of web log preprocessing, existing preprocessing techniques, the proposed User Interest Level based preprocessing (UILP) algorithm and performance of the proposed (UILP) algorithm with existing algorithms to identify user interest level. They considered session and frequency values as the key for identifying user interest level.

IV. SUMMARY AND ANALYSIS OF LITERATURE REVIEW

Authors	Preprocessing Techniques					Discovery Of Pattern	Analysis of Pattern	Remarks
	Data Fusion	Data Cleaning	User Identification	Session Identification	Path Completion			
Sanjay Bapu Thakare et al.[2]	Yes [Improvised Merging Algorithm]	Yes	NA	NA	NA	NA	NA	- Implement Field extraction using JAVA Source code and Convert log file into database using algorithm named "*Improvised TransLog Algorithm"*
Amit Dipchanji Kasliwal et al. [9]	NA	Yes	Only Discussion	30 min expiration time for user session	NA	Association Rule	NA	- Use MatLab7.0 for Data Cleaning and filtering - Implement association rule using RapidMiner tool - Used log file from KDD repository Con: To model the frequent user visits the website had accessed during the specific period.
Navin Tyagi et al. [21]	NA	Yes [Algorithm]	NA	NA	NA	NA	NA	- Give data cleaning and data reduction algorithms based on CERN (common log file) format. - Other preprocessing techniques missing
K.R. Suneetha et al. [23]	NA	Yes	Yes	NA	NA	NA	NA	- Implemented data leaning and user identification on NASA web server

								log file and display result for improving website performance
Jaideep Srivastava et al. [22]	Yes	Yes	Yes	Yes	Yes	Yes	Yes	- Used WebSIFT system to perform WUM from server log in the extended NSCA formats(include referred and agent fields)
V. Chitraa et al. [6]	Yes	Yes [Including robot cleaning]	Yes [using Reference Length]	Yes [using Reference Length]	Yes [using Reference Length]	NA	NA	- Provide result of data cleaning and path completion after experiment.
Sheetal A. Raiyani et al. [25]	NA	Yes	Yes [Distinct User Identification]					- Mentioned standard user identification with their problem and proposed "Distinct User Identification" algorithm
Vellingiri J. et al.[26]	NA	Yes [Including robot cleaning]	Yes [using Reference Length]	Yes [using Reference Length]	Yes [using Maximal Forwarded Reference]	NA	NA	- Computing reference length and maximal reference forwarded algorithm to fulfill path completion technique.
G. T Raju et al.[27]	Yes [Algorithm with pseudo code]	Yes	Yes	Yes [Pseudo code]	Yes	NA	NA	- Implemented all pseudo code on NASA and academy website log file.
P. Nithya et al.[28]	Yes	Yes [Including local and global noise removing, Robot cleaning]						- Implemented on Microsoft Web Dataset and MSNBC.com website log files.
Vellingiri J. et al.[7]	Yes	Yes [Including local and global noise removing, Robot cleaning]	NA	NA	NA	Yes [Weighted Fuzzy Possibilistic C-Means Algorithm]	Yes [ANFIS with Subtractive Algorithm]	- Implemented Data cleaning in preprocessing including robot cleaning and apply Weighted Fuzzy-Possibilistic C-Means Algorithm for pattern discovery, Adaptive Neuro-Fuzzy Inf erence System with Subtractive Algorithm for pattern analysis. - Included result of pattern discovery
Ashwin Riyani et al.[8]	NA	Yes	Yes [Distinct User Identification Algorithm]	NA	NA	NA	NA	- Introduced proposed Technique DUI (Distinct User Identification) based on IP address, Agent and Session time, Referred pages on desired session time.
Renáta	Yes	Yes	Yes	NA	NA	NA	NA	- Develop Web

Iváncsy et al.[10]			[Cookie-based user identification]					Activity Tracking (WAT) system that finds distinction of web users based on log data. - More focuses on cookie based analysis. - Implemented and show results
Arvind Dangi et al.[11]	NA	Yes	Yes [Using JAVA coding]	Yes [Using JAVA coding]	NA	NA	NA	- Provide a framework to identify user and session with each step of data preprocessing
S. Prince Mary et al.[12]	Yes	Yes [Common Algorithm]	Yes [Based on IP and Agent]	Yes [Time Oriented steps]	Yes [Using Graph Model]	NA	NA	- Apply algorithms of data preprocessing on web logs and showed result of user and session identification.
Shaily Langhnoja et al.[17]	Yes	Yes [Common Algorithm]	Yes [Common algorithm]	Yes [Common algorithm]	NA	NA	NA	- Used algorithms of data cleaning, user and session identification on web log files and showed results
Zidrina Pabarskaite [19]	NA	Yes	Yes	NA	NA	NA	NA	- Introduced Novel advanced cleaning improves web log mining results. Improved filtering removes pages with no links from other pages
C. E. Dinuca [24]	NA	Yes	NA	Yes [Modified basic algorithm using mean value of sessions]	NA	NA	NA	- Implemented in Java programming language by using NetBeans IDE, two algorithms to identify sessions and showed results.
R. Suguna et al.[29]	Yes	Yes [UILP algorithm]	Yes [UILP algorithm]	Yes [UILP algorithm]	Yes [UILP algorithm]	NA	NA	- Provided User Interest Level Processing algorithm to improve quality of preprocessing

[Table [1]: Summary of Literature Review of Preprocessing Techniques]

Analysis and Issues: On basis of review,, it can be seen that there are many common data preprocessing techniques applied in various types of log files. Some authors used common techniques like remove graphical records, HTTP failure status record etc. But three or four authors included robot cleaning in data cleaning preprocessing which helps while log files are collected from proxy server or having proxy server between server and client. To identify user, researchers used different methods based on relevant data with predicting IP address relationship, based on cookie, based on reference length, etc. Researcher implemented different algorithms for identifying session based on 30 minutes expiration time or average time spent by users, etc.

Further, it is found that there is no algorithm focusing on multiple clients accessing website through the same IP address. It can be considered an area of further research.

V. CONCLUSION

The web server having huge amount of data with different fields which are not reflected directly about user visiting pages during users interaction with website. Due to availability of irrelevant data of user's visiting activity in web logs there is not surly getting user and session identification properly. So web log files need to be preprocessing before mining processes like to discover knowledge of user behavior pattern. Data preprocessing is required phase in web usage mining process. There are different heuristics algorithms or methods used to remove irrelevant data from web logs files and identify user and session. This paper has focused on applied various data preprocessing methods with their advantages and disadvantages. After completion of data preprocessing, applying data mining algorithms like clustering, associations, classifications etc. for web usage mining with respect to identifying web user's behavior.

REFERENCES:

(1) L.K.Joshila Grace, V. Maheshwari, Dhinaharan Nagamalai," Analysis of Web Logs and Web User in Web Mining", International Journal of Network Security & Its Applications (IJNSA), Vol.3, No.1, January 2011
(2) Sanjay Bapu Thakare, Prof. Sangram Z Gawali, "A Effective and Complete Preprocessing for Web Usage Mining", International Journal on Computer Science and Engineering, Vol. 02, No. 03, pp. 848-851, 2010,
(3) Neha Jagannath, Sidharth Samat, "A Survey on Cleaning of Web Pages Before Web Mining", International Journal of Innovations & Advancement in Computer Science, ISSN 2347-8616, Vol. 3, Issue 8, October 2014
(4) Vijayashri Losarwar, Dr. Madhuri Joshi, "Data Preprocessing in Web Usage Mining", International Conference on Artificial Intelligence and Embedded Systems, July 15-16, 2012, Singapore
(5) Mitali Srivastava, Rakhi Garg, P.K.Mishra, "Preprocessing Techniques in Web Usage Mining: A Survey", International Journal of Computer Applications (0975-8887), Vol. 97, No. 18, July 2014
(6) V.Chitraa, Dr. Antony Selvadoss Thanamani, "Web Log Data Cleaning for Enhancing Mining Process", International Journal of Communication and Computer Technologies", Vol. 01, No. 11, Issue 03, December 2012.
(7) Vellingiri J., S. Kaliraj, S. Satheeshkumar and T. Parhtiban, " A Novel Approach for User Navigation Pattern Discovery and Analysis for Web Usage Mining", Journal of Computer Science 11(2); 372-382, 2015
(8) Ashwin G. Raiyani, Sheetal S. Pandya, "Discovering user identification mining techniques for preprocessed web log data", Journal of Information, Knowledge and Research in Computer Engineering, ISSN: 0975-6760, Vol. 2, Issue. 2, Pages 477-482, OCT-2013
(9) Amit Dipchandji Kasliwal, Dr. Girish S. Katkar, " Web Usage mining for predicting User Access Behavior", International Journal of Computer Science and Information Technology, Vol. 6 (1), 2015, 201-204
(10) Renata I., Sandor J., "Analysis of Web User Identification Methods", World Academy of Science, Engineering and Technology, 2007
(11) Arvindkumar Dangi, Sunita Sangwan, " A new approach for user identification in web usage mining preprocessing", IOSR Journal of Computer Engineering, e-ISSN: 2278-0661, p-ISSN: 2287-8727, Vol. 11, Issue. 3, (May-June2013), Pages 57-61
(12) S. Princy Mary, E. Baburaj, "An efficient approach performs pre-processing", Indian Journal of Computer Science and Engineering, ISSN: 0976-5166, Vol. 4, No.5, Oct-Nov-2013
(13) C.P.Sumathi, R. Padmaja Valli, T. Santhanam, "An overview of preprocessing of web log files for web usage mining", Journal of Theoretical and Applied Information Technology, ISSN: 1992-8645, e-ISSN: 1817-3195, Vol. 34, No.1, December 2011
(14) Dr. Sanjay Dhawan, Mamta Lathwal, "Study of preprocessing Methods in Web Server Logs", International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128X, Vol. 3, Issue. 5, May - 2013
(15) C. Oostuizen, J Wesson, C Cilliers, "Visual Web Mining of Organizational Web Sites", Proceeding of the Information Visualization
(16) A. Jameela, P.Revathy, "Comparisons of Decision and Random tree algorithms on A web log data for finding frequent patterns", International Journal of Research in Engineering and Technology, e-ISSN: 2319-1163, p-ISSN:2321-7308, Vol. 3, Issue. 7, May-2014
(17) Shaily Langhnoja, Mehul Barot, Darshak Mehta, " Pre-processing: Procedure on Web Log File for Web Usage Mining", International Journal of Emerging Technology and Advanced Engineering, ISSN: 2250-2459, ISO 9001:2008 Certified Journal, Vol. 2, Issue. 12, December 2012
(18) Naga Lakshmi, Raja Sekhara Rao, Sai Satyanarayan reddy, " An overview of Preprocessing on Web Log Data for Web Usage Analysis", International Journal of Innovative Technology and Exploring Engineering, ISSN: 2278-3075, Vol. 2, Issue. 4, March - 2013
(19) Pabarskaite Z (2002), "Implementing advanced cleaning and end-user interpretability technologies in web log mining", in 24th International Conference on Information Technology Interfaces (ITI), Vol. 1 Page(s): 109-113.
(20) D. Tanasa, B. Trousse (2004), "Advanced Data Preprocessing for Intersites Web Usage Mining" in IEEE Intelligent Systems, Vol. 19 Issues. 2 Page(s): 59-65.
(21) Navin Kumar Tyagi, A.K.Solanki, Sanjay Tyagi, " An Algorithmic approach to data preprocessing in Web Usage Mining", International Journal of Information Technology and Knowledge Management, Vol. 2, No. 2, pp. 279-283, Dec-2010
(22) Jaideep Srivastava, Robert Cooley, Mukund Despande, Pang-Ning Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", SIGKDD Explorations, Vol. 1, Issue 2 Jan 2000
(23) K.R. Suneetha , Dr. R. Krishnamoorthi, "Indentifying User Behavior by analyzing Web Server Access Log File", International Journal of Computer Science and Network Security, Vol. 9, No. 4, April 2009
(24) C.E. Dinuca, D. Ciobanu, " Improving the session identification using the mean time", International Journal of Mathematical Models and Methods in Applied Sciences, Vol. 6, Issue 2, 2012
(25) Sheetal A. Raiyani, Shailendra Jain, Ashwin G. Raiyani, " 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
(26) Vellingiri J. 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, ISSN: 1549-3636, 2011
(27) G T Raju and P S Satyanarayana, "Knowledge Discovery from Web Usage Data: Complete Preprocessing Methodology", International Journal of Computer Science and Network Security, Vol. 8, No. 1, January 2008
(28) P. Nithya and Dr. P. Sumathi, "Novel Pre-Processing Technique for Web Log Mining by Removing Global Noise, Cookies and Web Robots", International Journal of Computer Applications, Vol. 53, No.17, September-2012