

A Comprehensive Survey on Data Preprocessing Methods in Web Usage Mining

Sujith Jayaprakash

*Sr. Lecturer
BlueCrest College
Accra, Ghana*

Balamurugan E.

*Assoc. Professor
BlueCrest College
Accra, Ghana*

Abstract—Web usage mining is the application of data mining technique which is used to extract information about user's interest from web server log files. Web usage mining is widely used by companies to analyze the customer's interest and predict future of their business. It is used in various fields like E-Business, E-Commerce, E-learning, etc., Web usage mining entails of three phases :- Data Preprocessing , Pattern Discovery and Pattern analysis. Data Preprocessing is one of the essential and a preliminary step in web mining to enforce quality in the input data. The raw data from web server log file is preprocessed to eliminate the noisy, vague and redundant data for efficient mining. It involves different phases namely Field Extraction and Data cleaning, User Identification, Session Identification, Path completion and Transaction Identification. In this paper, we have discussed about various researches carried out in Data Cleaning and the various attributes considered in the process of cleaning.

Keywords—Web usage mining, Data Preprocessing , Session Identification, Path Completion, User Identification

I. INTRODUCTION

Proliferation of Internet and the popularity of mobile devices and computer have paved way for the increase in e-commerce application. The growth of e-commerce has exploded all over the world. With the information overload on the web, it is highly desirable to mine data and extract patterns and information relevant to the user, making the task of browsing much easier. Therefore, there has been much interest in the field of web mining, which is essentially mining databases on the web, mining usage patterns by analyzing the web logs, or mining web links to extract structure [1]. Web mining is a technique to discover the useful information from hyperlinks, page content and usage log. Web mining is an old wine in a new bottle.

Web mining = Databases + Information Retrieval + Artificial Intelligence

Web mining is divided in to the following categories,

1. Web Content Mining
2. Web Structure Mining
3. Web Usage Mining

Web content mining is the scanning and mining of text, pictures and graphs of a web page. This mining is used to gather, categorize, organize and provide the best possible solution in the internet to the user's request. Search engine like google, yahoo uses this technique to get results through text mining. Companies are also using this technique to increase their visibility and traffic through search engines.

Web structure mining is a tool to identify the relationship between the web pages linked by information. Web structure mining is used in the business sector to link information of its own website to enable an efficient navigation.

Web usage mining allows collecting the web access information for the web page. It helps the businesses to understand customer's behavior by accessing web server log files. The raw information stored in the server log file is preprocessed to identify the usage patterns.

Preprocessing is the first step in Web usage mining. It is an essential activity which will help to improve the quality of the data and successively the mining results. Different attributes are used in the data cleaning and in this paper we have analyzed the research works carried out in Data Cleaning.

This paper is organized as follows, In section II, related research works of data preprocessing in web usage mining is discussed. In Section III, we will discuss the web log file and its structure. In Section IV, we will discuss about various researches in data preprocessing phases. In Section V, We will discuss about pattern analysis and pattern discovery, following that we will discuss on the conclusion and future research works.

II. RELATED RESEARCH WORKS

Preprocessing of the raw data in the server log file is an important step in web usage mining to ensure the quality of information used in mining.

Wasavand et al. identified user's navigation pattern by data cleaning on web log file. They also used classification algorithms to identify users interested website [3].

Manisha conducted data cleaning and distinct user identification technique which enhance the preprocessing steps of web log usage data. Using user identification they found out the distinct user based on their attended session time. This will help in personalizing the websites.[4]

Cooley et al. Presented methods for user identification, session identification, page view identification, path completion, and episode identification [5]. They proposed some heuristics to deal with the difficulties during data preprocessing. Pankaj M. Meshram, Prof. Gauri A.

Chaudhary used clustering techniques to complete the path and improve the websites performance. [6]

T. Vijaya Kumar, H. S. Guruprasad, Bharath Kumar K. M., Irfan Baig, and Kiran Babu S introduced a new idea of incorporating available website knowledge for better session construction which would eventually lead to better patterns during pattern discovery. By using concept based approach they captured the actual intuition of the user which is sole purpose of any mining process. By identifying user's navigation between concepts, they have generated user profiles which will be useful for administrator to predict user behavior for a particular group of users. Recommendation models based only on usage information are inherently incomplete because they neglect domain knowledge. Field Extraction [7]

Addanki Ramya, Konda Sreenu, P Ratna Kumar used multi-layered network architecture with a back propagation learning mechanism to discover and analyze useful information from the available web log data. The discovered data is used to predict the users behavior E-Commerce applications. [8]

Bagilon et al. Aimed at extracting models of the navigational behavior of website users. After the data cleaning process they carried out two experiments: the first one tries to predict the sex of a user based on the visited web pages, and the second one tries to predict whether a user might be interested in visiting a section of the site.[9]

Nawal Sae11 , Abdelaziz Marzak2 and Hicham Behja3 developed a data preprocessing method which applied to moodle logs to understand the students learning process according to the different levels of contents accessed. After the data cleaning process, they applied clustering technology to analyze the group of learners and association rules mining to find more relationship between different parts of content. [10]

Priyanka and Roshni have proposed a customized application specific methodology for preprocessing the web log data and a modified frequent pattern tree for discovery of pattern efficiently. They proposed a method of customized web log preprocessing, rather than traditional approach, which may reduce the size of raw web log file. Improved Frequent Pattern Tree structure is used for iterative approach with support count restriction to reduce execution time and memory rather than traditional Frequent Pattern Tree. [11]

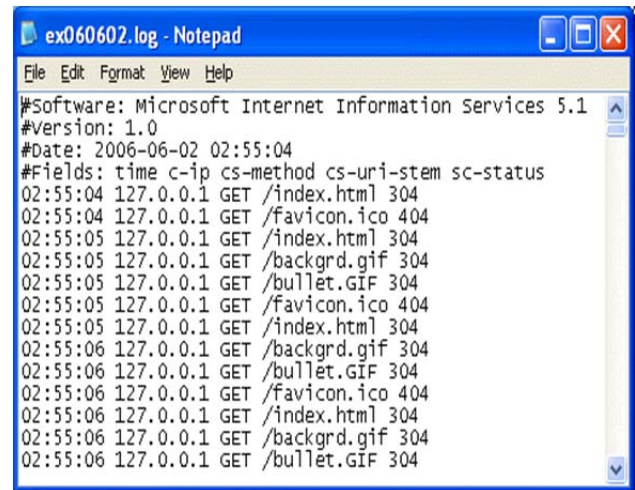
Priyanka and Nishta proposed a methodology for data cleaning and ip address identification stages of preprocessing.

They used different data mining algorithms for data cleaning and ip address identification.

From the related research works, it's clearly evident that preprocessing is an inexorable process in the web usage mining to understand the customer behavior.

III. WEB SERVER LOG FILE

Web server log files are created automatically in the web server. Log file will contain information like page request history, IP address, Time stamp, HTTP code, and user agent etc., Server log files are not accessible to the internet users and are accessible only to the Web administrators or Web masters.



```

ex060602.log - Notepad
File Edit Format View Help
#Software: Microsoft Internet Information Services 5.1
#Version: 1.0
#Date: 2006-06-02 02:55:04
#Fields: time c-ip cs-method cs-uri-stem sc-status
02:55:04 127.0.0.1 GET /index.html 304
02:55:04 127.0.0.1 GET /favicon.ico 404
02:55:05 127.0.0.1 GET /index.html 304
02:55:05 127.0.0.1 GET /backgrd.gif 304
02:55:05 127.0.0.1 GET /bullet.GIF 304
02:55:05 127.0.0.1 GET /favicon.ico 404
02:55:05 127.0.0.1 GET /index.html 304
02:55:06 127.0.0.1 GET /backgrd.gif 304
02:55:06 127.0.0.1 GET /bullet.GIF 304
02:55:06 127.0.0.1 GET /favicon.ico 404
02:55:06 127.0.0.1 GET /index.html 304
02:55:06 127.0.0.1 GET /backgrd.gif 304
02:55:06 127.0.0.1 GET /bullet.GIF 304

```

Fig (a) Sample Server Log File entry

First few lines of the log file in the Fig (a) show the Web server name and the version details. Following are the details about the fields displayed in the log file.

Date / Time: Date and Time of the request.

C-IP: Displays the ip address of the client/ browser.

CS-Method: Request method used for this request.

Common methods are GET, POST or HEAD.

CS-URI-STREAM: Address of the page requested.

SC-STATUS: Response given by the server for the request made.

Log file data are noisy, redundant and vague. Hence, preprocessing is required to eliminate the unwanted data. Interpretation of data from the log file can be done manually or using a software. According to Bertot, McClure, Moen, and Rubin [12], Web server automatically updates four types of usage log file:

Access log file: It stores access information of the user regarding date/time, IP address and user action.

Agent log file: This will provide information about the browser.

Error log file: Error Logs contain information on specific events such as "file not found," "document contains no data," or configuration errors; the time, user domain name, and the page on which a user received the error is recorded, providing a server administrator with information on problematic and erroneous links on the server.

Referrer log file: Referrer logs provide information on what Web pages, from both the site itself and other sites, contain links to documents stored on the server.

IV. PREPROCESSING PHASES

Preprocessing is a data mining technique used to process the raw data into an understandable format. Preprocessing transforms the data into a format that can be used for efficient mining and produce accurate results. In real world, data are generally incomplete, noisy or inconsistent. In Web usage mining, preprocessing is carried out to eradicate noisy data in the web server log file. Following Fig (b) will depict the various phases of preprocessing in web usage mining.

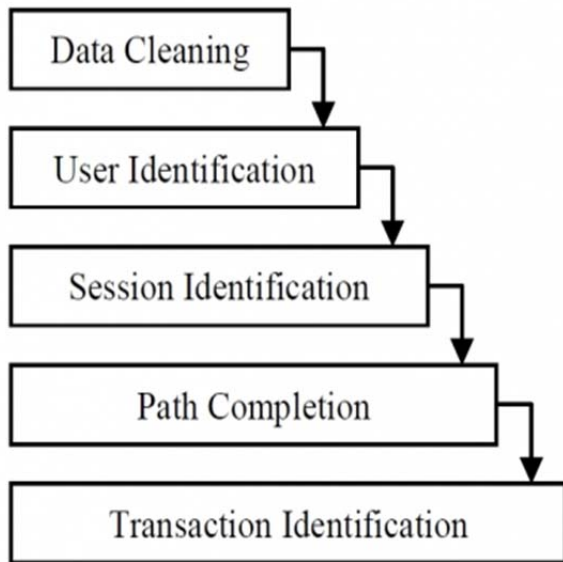


Fig (b) Preprocessing Phases

S.No	Date	Time	IP Address	Method	URI	Status
1	Monday, February 06, 2006	2:55:04	127.0.0.1	GET	/index.html	304-Not Modified
2	Monday, February 06, 2006	2:55:04	127.0.0.1	GET	/favicon.ico	404-File not found
3	Monday, February 06, 2006	2:55:05	127.0.0.1	GET	/index.html	304-Not Modified
4	Monday, February 06, 2006	2:55:05	127.0.0.1	GET	/backgrd.gif	304-Not Modified
5	Monday, February 06, 2006	2:55:05	127.0.0.1	GET	/bullet.gif	304-Not Modified
6	Monday, February 06, 2006	2:55:05	127.0.0.1	GET	/favicon.ico	404-File not found
7	Monday, February 06, 2006	2:55:05	127.0.0.1	GET	/index.html	304-Not Modified
8	Monday, February 06, 2006	2:55:06	127.0.0.1	GET	/backgrd.gif	304-Not Modified
9	Monday, February 06, 2006	2:55:06	127.0.0.1	GET	/bullet.gif	304-Not Modified
10	Monday, February 06, 2006	2:55:06	127.0.0.1	GET	/favicon.ico	404-File not found
11	Monday, February 06, 2006	2:55:06	127.0.0.1	GET	/index.html	304-Not Modified
12	Monday, February 06, 2006	2:55:06	127.0.0.1	GET	/backgrd.gif	304-Not Modified
13	Monday, February 06, 2006	2:55:06	127.0.0.1	GET	/bullet.gif	304-Not Modified

Fig (c) Sample log file extracted data

A. Field Extraction and Data Cleaning

Web server log file contains hulking raw data and it is important to extract the fields from the file to remove the inconsistent data. Usually log file data are separated using

(,) or (“”). Field extraction is the first phase in Preprocessing where the data will be extracted from the different fields. This can also be achieved using Excel or any third party software which will excerpt fields and place it in a tabular column. A sample extracted data is show below in fig (c).

B. User Identification

User identification is about identifying the identity of the user and the web pages accessed. In these days, corporate companies use a proxy server to hide the IP address from hackers or intruders. Hence, the requests coming through a proxy server will have the same IP address. To overcome these problems we can make a user to register in the server and through that his details can be captured. But, not every user registers to access a web page. Another solution for this can be the use of cookies [23].

C. Session Identification

The aim of session identification is to divide the page accesses of each user at a time into individual sessions [24]. Session is a time period in which the user is active in a website. Within a particular time period a user can access different pages for different times. By finding the total number of pages accessed and the time spent by the user on each page will help us in identifying an effective user. A user may have a single or multiple sessions during a time period. Once a user has been identified, the click stream of each user is portioned in to logical clusters. The method of portioning into sessions is called Sessionization or Session Reconstruction.

D. Path Completion

Due to local cache, the cache agent, post technique and the back button the URL paths might be incomplete. Hence to finding the missing pages, Path completion is necessary. Path completion is done based on the URL and the referrer URL’s in a user session. If a page request made is not directly linked to the last page requested, the recent history of session is searched and if the page is available previously as referrer URL, the related URL of the previous entry is added in path. [25]

E. Transaction Identification

The goal of Transaction identification is to create meaningful clusters of references for each user. The identification of transactions varies from case to case, depending on the web usage mining.

V. EXPERIMENT

Data Cleaning plays a vital role in the Preprocessing and various attributes are considered to remove the inaccurate and inconsistent records from the web log file.

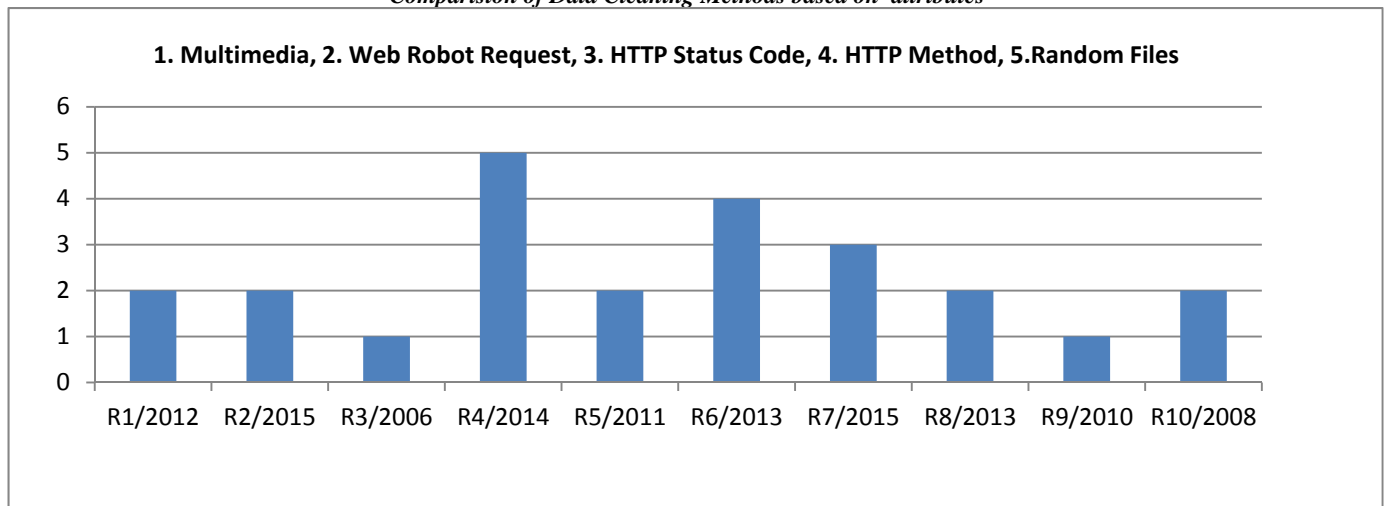
Following matrix table will show the various research works done in the field of data cleaning with different attributes

Research Work	Author / Year	Attributes
Field extraction is done in Java Programming language by considering the data as String. It is assumed that the space character is used as separator. All the characters are read and using the String Tokenizer class the data fields are broken into tokens and save in an array. Data cleaning is done using various algorithms which retain the data where its status code is 200 and the method is GET and file type is except from gif, jpeg and css [13]	Surbhi, A. & Rinkle, R / 2011	HTTP Status Code HTTP Method
An algorithm is proposed for the extraction of data sorted based on the analysis of time duration. Also, an algorithm is proposed in data cleaning to remove almost all irrelevant files, irrelevant HTTP methods and wrong HTTP status codes. After experiment it is analyzed that raw log data reduces to almost 80% which shows the importance of initial phases of data preprocessing.[14]	Mitali, S. & Rakhi, G. & Mishra, P. K./ 2015	HTTP Status Code HTTP Method
A learning based algorithm is designed based on the differences between the retrieval target pages and the ordinary pages and based on these features data cleaning has been done.[15]	Yiquin, L. & Zhang, M. & Liyun, R. & Shaoping, M./2008	HTTP Method
A web log cleaning algorithm with rules and conditions is presented for intrusion detection by taking 5 major attributes in to consideration namely Multimedia, Web Robots Request, HTTP Status code, HTTP Method, Other files.[16]	Yew Chuan Ong & Zuraini Ismai. / 2014	Multimedia Web Robot Request HTTP Status Code HTTP Method Random Files
Proposed Data Cleaning algorithm to eliminate irrelevant or unnecessary items in the analyzed data. This algorithm will eliminate the unwanted records by checking the HTTP Method, Status code and the suffix of URL Link for Multimedia images.[17]	Aye, TT / 2011	HTTP Status Code HTTP Method
Proposed an algorithm which moves the unsuccessful request, Not equal to Get Method, Request for Multimedia Objects, Request from web robots are moved to an anomaly table as a part of data cleaning process.[18]	Kamat, M. & Bakal, J. & Nashipudi, M./2013	Multimedia Web Robot Request HTTP Status Code HTTP Method
Data Cleaning is achieved by removing Global and Local Noise, Multimedia Images, HTTP Status code, Request from web robots.[19]	Shahu, M.S & Leena/2015	Multimedia Web Robot Request HTTP Status Code
Proposed an algorithm to remove the log file data which checks the url contains gif, jpeg, jpg and css. AND different error like HTTP 404.[20]	Langhnoja, S. G, Barot, M. & Mehta, D./2013	Multimedia HTTP Status Code
Data Cleaning is done using an algorithm which filters out and removes the links containing gif, jpg and css[21]	Tyagi, N.K & Solanki, A. K./2010	Multimedia
HTTP error codes and multimedia files are removed using an algorithm.[22]	Sharma, A.	Multimedia HTTP Status Code

In the above matrix, various attributes are considered for the cleaning of data. From the research papers analyzed, it's obvious that the increase in the attributes has resulted in better optimization of results. The complexity of the raw data in the log file can be reduced by considering the above

mentioned parameters. Research works done in the recent times reveals that by considering the below mentioned attributes, the data can be cleaned with 80% accuracy.

Comparison of Data Cleaning Methods based on attributes



VI. CONCLUSION AND FUTURE RESEARCH

Due to irrelevant data in log file, preprocessing is considered as an essential step in web usage mining. In this paper we have discussed about various data cleaning algorithms. High dimensionality and large volume of data results in high computational complexity of mining process. So there is need to compress data without losing essential information regarding user's behavior. Most of the algorithms use very minimal attributes like HTTP Status code and Multimedia to remove the inconsistent data from the log file. There are very minimal researches done in the process of eliminating request from web robots or other unwanted files like Java Scripts. Future research works can consider including more attributes to minimize the irrelevant data. Hence this paper concludes that by removing the Multimedia, Request from web robots, HTTP Stats Codes, HTTP Methods and Random files a quality data can be retrieved from the log file.

REFERENCES

1. Thuraisingham, B. (2003). *Web Data Mining and Applications in Business Intelligence and Counter-Terrorism*, New Delhi: CRC Press
2. Sudheer Reddy, K. & Kantha Reddy, M. & Sitramulu, V. An effective data preprocessing method for Web Usage Mining. *Information Communication and Embedded Systems (ICICES)*, 2013.
3. Wasavand, C. & Devale, P.R & Ravindra, M. "Data Preprocessing Method of Web Usage Mining for Data Cleaning and Identifying User navigational Pattern". *IJSET - International Journal of Innovative Science, Engineering & Technology*, Vol. 1 Issue 10, 2014.
4. Manisha V. "A Step up in Data Cleaning and User identification of Preprocessing on Web Usage data". *International Journal of Advanced Research in Computer Engineering and Technology IJARCET*, 2014
5. Cooley R, Mobasher B, Srivastava J., "Web Mining: Information and Pattern Discovery on the World Wide Web". In *International Conference on Tools With Artificial Intelligence*, pages 558-567, IEEE, 1997.
6. Pankaj, M. & Gauri, A. "Mining of Web Logs Using Preprocessing and Clustering". *International Journal of Innovative Research in Computer and Communication Engineering*, Vol. 2, Issue 12, December 2014
7. Vijay Kumar, T. & Guruprasad, H.S. & Bharath Kumar, K.M. & Irfan, B. & Kiran, B. "A New Web Usage Mining Approach for Website Recommendations Using Concept Hierarchy and Website Graph". *International Journal of Computer and Electrical Engineering*, Vol. 6, No. 1, February 2014
8. Ramya, A. & Sreenu, K. & Ratna, K. "Preprocessing and Unsupervised Approach For Web Usage Mining", *International Journal of Social Networking and Virtual Communities*. Vol. 1, Issue 2.
9. Bagilon, M. & Ferrara, U. & Romei, A. & Ruggieri, S. & Turini, F. "Preprocessing and Mining Web Log Data for Web Personalisation". *Advances in Artificial Intelligence Lecture Notes in Computer Science Volume 2829*, 2003, pp 237-249
10. Sael, N. & Marzak, A. & Behja, H. "Web Usage Mining data preprocessing and multi-level analysis on Moodle", *Computer Systems and Applications (AICCSA)*, 2013 *ACS International Conference*
11. Priyanka, D. & Roshni D, "Pattern Detection With Improved Preprocessing in Web Log", *An International Research Journal of Computer Science and Technology*.
12. Charles, R. & Bertot, J. (2002) "Evaluating Networked Information Services", ASIST Members.
13. Surbhi, A. & Rinkle, R. "An Efficient Algorithm for Data Cleaning of Log File using File Extensions", *International Journal of Computer Applications*, Vol. 48, Issue 8.
14. Mitali, S. & Rakhi, G. & Mishra, P. K. "Analysis of Data Extraction and Data Cleaning in Web Usage Mining", *Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015)*.
15. Yiquin, L. & Zhang, M. & Liyun, R. & Shaoping, M. "Data Cleansing for Web Information Retrieval using Query Independent Features". *Journal of the American Society for Information Science and Technology*, Vol. 58, Issue 12, Pages 1884-1898.
16. Yew Chuan Ong & Zuraini Ismai. "Enhanced Web Log Cleaning Algorithm for Web Intrusion Detection". *Recent Advances in Information and Communication Technology Advances in Intelligent Systems and Computing Volume 265*, 2014, pp 315-324
17. Aye, TT. "Web log cleaning for mining of web usage patterns". *Computer Research and Development (ICCRD)*, 2011.
18. Kamat, M. & Bakal, J. & Nashipudi, M. "Improved Data Preparation Technique in Web Usage Mining", *International Journal of Computer Networks and Communications Security*, 2013. Pp 284-291
19. Shahu, M.S & Leena "A Survey on Frequent Web Page Mining with Improving Data Quality of Log Cleaner", *International Journal of Advanced Research in Computer Engineering & Technology*, 2015
20. Langhnoja, S. G, Barot, M. & Mehta, D. "Web Usage Mining Using Association Rule Mining on Clustered Data for Pattern Discovery", *International Journal of Data Mining Techniques and Application*, 2013.
21. Tyagi, N.K & Solanki, A. K. "An Algorithmic approach to Data Preprocessing in Web Mining", *International Journal of Information Technology and Knowledge Management*, 2010, Volume 2, No. 2, pp. 279-283
22. Sharma, A. "Web Usage Mining: Data Preprocessing, Pattern Discovery and Pattern Analysis on the RIT Web Data".
23. Sudheer, K. Nagarjuna, A. Reddy, K & Sitramula, V. "An effective data preprocessing method for Web Usage Mining", *Information Communication and Embedded Systems (ICICES)*, 2013, pp 7-10.
24. Bari, P.H. Chawan, P.M "Web Usage Mining", *Journal of Engineering, Computers & Applied Sciences*"
25. Chitra, V. Thanamani, A.S. "A Novel Technique for Sessions Identification in Web Usage Mining Preprocessing", *International Journal of Computer Applications*, 2011.