ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 4, Issue 2, March 2015

Data Preprocessing: A Milestone of Web Usage Mining

Pooja Kherwa, Jyotsna Nigam

Abstract:-.Internet is today full of structured or unstructured information. and this information is directly or indirectly influencing society or peoples. Because today internet is part our daily life activity. But using this abundant and ambiguous in most efficient manner in useful decision making is still a big challenge. During our web surfing either it is online shopping or blogging or using tweets and chatting everything is recorded. Web servers and log files are used to collect all the activity information of a user accesses to a web server. Log data is usually noisy and ambiguous. Web usage mining, also known as web log mining, is the application of data mining techniques on large web log databases to discover knowledge about user behavior pattern and web site usage statistics that can be used for various website design tasks. WUM consists of three phases:-Preprocessing, Pattern discovery and Pattern analysis. It is a fact that normal log files data is very huge, noisy, unclear and confusing with lots of redundancy. It is very important to preprocess the log data for efficient web usage mining process. Preprocessing results also influences the later phases of web usage mining. This makes the preprocessing of server log files a significant step in web usage mining. This study includes analysis, comparison and contrast of the available preprocessing techniques. How we can be more focused and guided at preprocessing level. So in this paper, we have given a complete preprocessing analysis by reviewing the existing work done in the preprocessing stage.

Keywords: **Data cleaning, Preprocessing, Path Completion, Session Identification.**

I. INTRODUCTION

Today, everybody wants to search anything on the internet, within a fraction of a second by just a single click. There is a requirement of designing websites according to the user's needs. So analyzing user's behavior (recorded in web server log files) is an important part of web site design. Web users usually suffer from the information overload due to the significant increase and the rapidly expanding growth in the amount of information available on the web [5].

Web data mining [3] is the application of data mining techniques on web data. Web mining is divided into-

- Web content mining
- Web structure mining
- Web usage mining.

1) Web content mining deals with the useful discoveries of web contents and services.
2) Web structure mining aims to understand the structure of hyperlinks within the web itself.
3) Web usage mining mines the log data stored in the web server

Web usage mining (WUM) [23] [25] can be defined as the application of data mining techniques to weblog data in order to discover user access pattern. Web usage mining has various application areas such as user behavior prediction, site-reorganization and web personalization.

Web usage Mining comprises of three phases:

- Preprocessing
- Pattern discovery
- Pattern Analysis

The data stored in the log files don't present an accurate picture of the user's accesses to the web server [23]. Data preprocessing is the process to convert the raw data available in log files into the database tables for making it suitable for applying the data mining algorithm.

Hence preprocessing of web log data is most essential and a pre-requisite phase before it can be used for the pattern discovery task. Due to large amount of irrelevant entries in the web log file, the original log files cannot be directly used in the web usage mining process. Therefore the preprocessing of web log file becomes significant and important. The research on data preprocessing of Web Usage Mining is a field in focus nowadays. This paper attempts to present the process of data preprocessing in Web Usage Mining.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 4, Issue 2, March 2015

So, in this paper we have explained the complete roadmap for the preprocessing stage of web usage mining for determining web access patterns. For this, we have reviewed the work done in the preprocessing stage by various researchers. In 2, we have given an overview of data preprocessing and related work in this area. In 3, we define web log files. In 4, the complete roadmap for preprocessing of web log files and detailed analysis of work done by various researchers is presented.

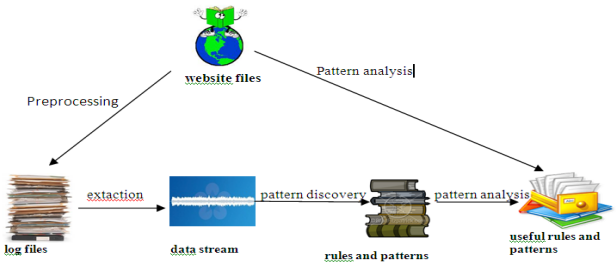

Fig 1: Web Usage Mining

II. RELATED WORK OF DATA PREPROCESSING IN WEB USAGE MINING

Web usage mining only concentrates on pattern discovery and pattern analysis. Data preprocessing is considered not as important as it is required to be in Web Usage Mining. Preprocessing of web log is most important for effective and efficient web mining.

Robert Cooley, Bamshad Mobasher and Jaidep Srivastava presented methods for user identification, session identification, page view identification, path completion, and episode identification [13]. They proposed some heuristics to deal with the difficulties during data preprocessing. Bettina Berendt and her colleagues compared time-based and referrer-based heuristics for visit reconstruction [1]. They found that a heuristic's appropriateness depends on the Web site's design and on the visits' length. The preprocessing work includes identifying users, server sessions, and inferring cached page references through the recorded information in web server logs.

Ali Bayir Murat and Ismail H.Torosly[14] have purposed a strategy by combining time and navigation oriented heuristics for Reactive Session reconstruction. In this, they find increase in accuracy of the reconstructed session. In this, they have also implemented an agent simulator, which models the behavior of web users and generates web user navigation as well as the log data kept by the web server.

Zhang Huiying, Liang Wei [34] gave an intelligent algorithm for data preprocessing in web usage mining. Nasraou presented a web usage mining framework for mining evolving user profiles in dynamic web sites, provides a whole framework and findings in mining web usage navigation from web log files of a genuine web site.

In the paper titled "An enhanced preprocessing research framework for web log data using learning algorithm.", Dr.V.Vallikumari**,** [6] presented a learning algorithm to separate human user and search engine accesses intelligently, in less time. The work ensures the goodness of spirit by using popular measures like entropy and Gini index.

Tanasa, D. and Trousse, B. [30] focused on web server logs from several web sites, generally belonging to the same organization. The authors provide a solution to join all the log files and reconstitute the visit. The authors added a new step called data summarization which allows the analyst to select only the information of interest based on storing data generalization.

FangYuan, Li-Juan Wang, Ge Yu,[10] mainly focused on analyzing visiting information from logged data in order to extract usage pattern, which can be classified into three categories: similar user group, relevant page group and frequency accessing paths. These usage patterns can be used to improve web server system performance and enhance the quality of service to the end users.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 4, Issue 2, March 2015

Marquardt C.G., Becker. K. Ruiz, D.D.[21], discusses the impact of developing the pre-processing phase according to concepts, problems and goals specific to web based learning environment. They focus on reduction of preprocessing efforts by automation of a number of tasks, the easier alignment of mining goals with the results that can be obtained and the possibility of involvement of domain related people with less technical skills w.r.t mining.

III. WEB SERVER LOGS

A. Web Server Logs

A Web server is a computer that delivers (serves up) Web pages. Every Web server has an IP address and possibly a domain name. For example, if you enter a URL in your browser, this sends a request to the server. The server then fetches the page and sends it to your browser. Web server data are created from the relationship between a person(s) interacting with a Web site and the Web server. A Web server log, containing Web server data, is created as a result of the http process that is run on Web servers [4].

All server activity (success, errors, and lack of response) is logged into a server log file (HTTP-ANALYZE). According to Bertot, McClure, Moen, and Rubin, Web servers produce and update dynamically four types of usage log files:

- Access log,
- Agent log,
- Error log
- Referrer log.

Access Logs: Access log provide the bulk of the Web server data, including the date, time, users IP address, and user action (e.g., whether or not the user downloaded a document or image). The following is some of the information that can be obtained from an access log:

The IP (Internet Protocol) address of the computer making the request for a document
The time stamp (user access date and time)
The user's request (e.g., html document or image requested, or data posted)

Agent Logs: Agent Logs supply data on the browser, browser version, and operating system of the accessing user.
Error Logs :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. Other kinds of data written to the error log include stopped transmissions; information on user-interrupted transfers are recorded (e.g., a user might click the browser Stop button which would produce a stopped transmission error message;
Referrer Logs: Referrer logs provide information on what Web pages, from both the site itself and other sites, contain links to documents stored on the server. The log provides information such as the URLs of sites and pages on sites that referred visitors to a particular page. For example, users may often arrive at a particular Web site through a search engine, and the referring search engine along with the keywords used in the originating query, can be obtained from the Referrer log.

B. Types of File Formats

Web server logs are stored in Common Log file Format (CLF) or Extended Log file Format. Common Log file Format includes date (date, time, and time zone of a request), client IP (remote host IP and/or DNS entry), user name (remote log name of a user), bytes transferred, server name, request (URI query), and status (http status code returned). Extended Logile Format includes bytes sent and received, server (name, IP address, and port), request (URI query and stem), requested service name, time taken for transaction to complete, version of transfer protocol used, user agent (the program making the request, such as a Netscape browser or a search engine "spider"), cookie ID, and referrer.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 4, Issue 2, March 2015

IV. DATA PREPROCESSING

Data preprocessing is done to improve data quality and to increase mining accuracy. The raw log data is preprocessed to get reliable session for efficient mining. The data preprocessing is often the most time consuming and most calculative step in the web usage mining process. The process may involve preprocessing the original data, integrating data into a form suitable for input into specific data mining operation. This process is known as data preparation [17]. Actually the input for the WUM process is a preprocessed file that gives exact information of who accessed the web site, what pages were requested and in what order, and for how long time each page was viewed. The purpose of data preprocessing is to improve data quality and increase mining accuracy.

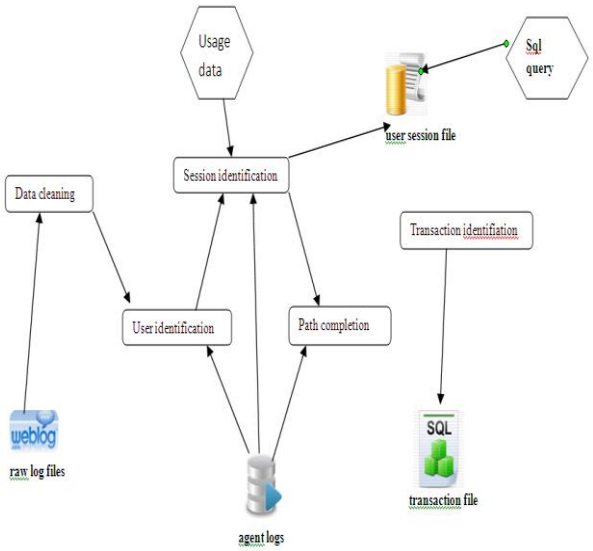

Fig 2.Phases of Data Preprocessing in Web Usage Mining

A. Field Extraction

Field extraction is the first phase of data preprocessing. The web log format entry contains various fields which need to be separate out for processing. The process of separating field from the single line of the log file is known as field extraction. The server used different characters which work as separators. The most widely used separator character is comma (,) and space (“ “) character.

Table 1.Analysis Matrix for Field Extraction

Authors	Work Done
Ciesielski and Lalani[8]	Idea is to get as much as information possible about the user-IP.
Anand Sharma[28]	1) First 3-last2 visited pages. 2) List of directories.

B. Data Cleaning

The data cleaning module is intended to clean Web log data by deleting irrelevant and useless records in order to retain only usage data that can be effectively used to recognize user’s navigational behavior. Since Web log files record all user interactions, they represent a huge and noisy source of data, often comprising a high number of unnecessary records.

The process of data cleaning is removal of outliers or irrelevant data. Data Cleaning enables to filter out useless data which reduce the log file size to use less storage space and to facilitate upcoming tasks. Analyzing the huge amounts of records in server logs is a complex and frustrating activity. So initial cleaning is necessary. If a user

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 4, Issue 2, March 2015

requests a specific page from server, at that time entries like gif, JPEG, etc., are also downloaded which are not useful .So for further analysis these are eliminated. The records with failed status code are also eliminated from logs.

The quality of the final results strongly depends on cleaning process. Appropriate cleaning of the data set has profound effects on the performance of web usage mining. The procedures of general data cleaning are as follows:

Web Log Files contains a large number of records which are not suitable for identifying the user's navigational behavior, some of them like

- Accesses from the Web Robots like GoogleBot, Yahoo Slurp and Alta Vista's scooter etc.
- Access for Images, we can identify these records by checking extensions for images ie. jpg, jpeg, png, gif, bmp, ico (for icons) of the accessed file.
- Access for Audio Files, we can identify these in a similar way as for images but considering the extensions .mp3, .wav, .mid, .rm, .ram etc.
- Identifying access records for Video Files, checking the accessed file's extension for .flv, .avi, .mpg, .mpeg, .mp4, .swf etc.
- Access for some important files like JavaScript files and CSS files, these are very important files for any web portal, but no users would take interest in these files.
- Those access records having status made by Administrator of the Web Portal code <200 and >299.
- Access, which is under consideration.

Table 2.Analysis Matrix for Data Cleaning

Authors	Work Done
Mohammad Ala a,AI –Hamami[22]	Classification of Log Files into number of files each one represent a class, Using Decision Tree Classifier.
Youquan He[33],	New approach to find frequent item sets employing Rough set Theory.
G.Castellano,A.M.Fanelli, M.A.Torsello. [10].	LODAP-Four modules are involved namely-Data Cleaning, Data structurisation, Data Filtering, Data Summarization.
Tan,P.N and Kumar[31].	In Data Cleaning removal of outliers or irrelevant data eliminating web robots generated log entries.
N. Kushmerick[19].	Purposed a feature based method, which identifies internet advertisement and remove them.

C. User Identification

User identification is an important issue in preprocessing process of web log file. It is actually how exactly the user have to be distinguished .It depends mainly on the task, how the mining process is executed. In some cases the users are identified with their IP addresses [18].

But the task of user identification is become tedious by the existence of local caches, corporate firewalls and proxy servers. The web log file recorded in CLF as well as ECLF, only records host or proxy IP, So different visible sharing the same host or one proxy server cannot be distinguished. Although some heuristics are used for better identification of the users. In [9] the different methods are grouped into two classes.

Proactive Strategy

1) These are aims at differentiating the users before or during the page request
2) Proactive strategy use cookies or dynamic web pages those are associated with the browser invoking them.

The shortcomings of such methods are that they rely on user's cooperation, but user often denies such cooperation due to privacy concerns or thinking it is not secure and reliable.

Reactive Strategy

1) Reactive strategy works with the recorded log files only, and the different user will be distinguished by their navigational patterns, download timing sequence or some other heuristics based on some assumption regarding their behavior.
2) Web users are also distinguished based on their navigational pattern using clustering [7] [8].

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 4, Issue 2, March 2015

Table 3. Analysis Matrix for User Identification

Authors	Work Done
Suresh R.M and Padmajavall.R[26]	Use accessing behavior is to be constructed as transaction.
Robert.Cooley,Bamshed.Mobasher and Jaideep Srinivastava.[11]	Users are identified using cookies or authentication mechanism
Istvan K,Nagy and Csaba.Gaspar-Papanek.[15]	Using Page Viewing time.
Renata Ivancsy and Sandor JUHASZ.[13]	User Identification using their IP Address.
T.Morzy,M.Wojciechowski[23]	Use are distinguished based on their navigational pattern using clustering Methods.
M.Spiliopoulou and B.Mobasher and B.Berendt M.Nakagawa[29].	Using Proactive and Reactive Strategies for Differentiating users.

D. User Session Definition and Construction

Depending on the popularity of the Web site, a Web log can record thousands or tens of thousands of requests every day. A request is recorded in a log file entry, which contains different types of information, including the IP address of the computer making the request, the user access date and time, the document or image requested, and so on. To find useful patterns (such as association rules or sequential patterns) from this vast amount of information, requests (or log entries) need to be grouped into usage sessions. A session is defined as a group of requests made by a single user for a single navigation purpose. A user may have a single session or multiple sessions during a period of time. So we define User Session as a bounded set of clicks realized in defined time. A session is also considered or identified on the basis of sufficiently long interval of time among two records visits of web page [29].

Proactive strategy that used cookies or dynamic web pages are also used for session identification, The shortcomings of such methods are that they rely on user's cooperation, but user often denies such cooperation due to privacy concerns or thinking it is not secure and reliable.
Two most widely used navigation driven method of user session identification is
1) Maximal Forward Reference transaction identification. [32].
2) Reference Length Method [10].

Many commercial products use 30 minutes as a default timeout, and [32] established a timeout of 25.5 minutes based on empirical data. Once a site log has been analyzed and usage statistics obtained, a timeout that is appropriate for the specific Web site can be fed back into the session identification algorithm.

Table 4 .Analysis Matrix for Session Identification

Authors	Work Done
Robert F.Dell[24]	Using Integer Programming construction of all session simultaneously.
Jose M,Domench and Javier LorenZo.[16]	In this Referrer based method and time oriented heuristics methods are combined.
Baoyao Zhou, Siu Cheung Hui and Alvis C-M.Fong.[2]	Time stamp based method .The default time is 30 minutes.
R.Cooley[9]	Time oriented Heuristics 30 minutes.
Catlegde L,and Pitkow J.[5]	Time oriented Heuristics 25.5minutes to 24 Hrs.
V.Chitra,Dr.Antony Selvadoss Thanamani[6]	This method based on navigation uses web topology in graph format.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 4, Issue 2, March 2015

D. Transaction Identification

Transaction Identification [change the language] some authors propose dividing or joining the sessions into meaningful clusters, i.e. transactions. Pages visited within a session can be categorized as auxiliary or content pages. Auxiliary pages are used for navigation, i.e. the user is not interested in the content (at the time) but is merely trying to navigate from one page to another. Content pages, on the other hand, are pages that seem to provide some useful contents to the user. The transaction generation process usually tries to distinguish between auxiliary and content pages to produce the so called auxiliary-content transactions (consisting of auxiliary pages up to and including the first content page) and the so called content-only transactions (consisting of only content pages). Several approaches, such as transaction identification by reference length [14] and transaction identification by maximal forward reference [32] [14] are available for this purpose.

E. Path Completion

Path completion is the problem of web log preprocessing, when some log entries are not present in the web log files. This occurs because sometime during our searching for information using "back" and "forward" button in the browsers. In this way we open the pages from proxy servers. So the entries for these web links will not be in the web log files maintained by web server. So to discover complete users travel patter, these missing user clicks should be appended .The techniques of achieving this is known as path completion. In path completion, not only the lost pages are appended, but also the time on these pages should be determined, because these lost pages are usually considered as auxiliary pages, their reference length can be estimated by the average reference length of auxiliary pages.

Table 5.Analysis Matrix for Path Completion

Authors	Work Done.
Yan Li, Boqin Feng , Qinjiao Mao[32]	Referrer-based method using proxy servers and local caching.
V.Chitraa, Dr. Antony Selvadoss Davamani[6]	Path completion, finding content path set, and travel path set
Anand Sharma[28]	(1)First 3-Last 2 Visited Pages (2) List Of Directories.

V. PATTERN DISCOVERY

When data preprocessing is completed, the next phase of web usage mining is pattern discovery. The pattern discovery is the most important phase of web usage mining process.The pattern discovery of user's behavior from preprocessed log files is done using various techniques like statistical analysis, association rules, sequential pattern analysis dependency modeling, classification and clustering etc [17].Hybrid dependency pattern exploit lexical relation and provide more accurate results[20]. Location of visitor can be determined using the IP address. Then the server administrator can find the most active countries visited at a particular site and he can provide the useful information relevant to that country.

VI. PATTERN ANALYSIS

The final phase in the Web Usage Mining process is pattern analysis. Pattern analysis is done to extract the only relevant pattern found in the pattern discovery phase. This phase rule out the pattern that are not relevant for our application domain.[17]. Pattern analysis involves the user evaluating each of the patterns identified in the second phase and deriving conclusions from these patterns. The user is generally concerned in finding patterns that provide useful information regarding the users' navigation. Visualization of these results allows them to be more easily interpreted by the user.

VII. APPLICATION OF WEB USAGE MINING

Web usage mining is used in the following areas.

- Web usage mining helps in improving the visualization of a Web site, in terms of content and structure.
- By knowing frequent access behavior for users, important links can be identified to improve the overall performance of future accesses.
- Information of frequently accessed pages can be used for caching.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 4, Issue 2, March 2015

- Web usage mining offers users the ability to analyze massive volumes of click stream or click flow data, integrate the data seamlessly with transaction and demographic data from offline sources and apply sophisticated analytics for web personalization, e-CRM and other interactive marketing programs.
- Web usage mining can be used in e-Learning, e-Business, e-Commerce, e-CRM, e-Services, e-Education, e-Newspapers, e-Government, and Digital Libraries.
- Web usage mining can be used in Customer Relationship Management, Manufacturing and Planning, Telecommunications and Financial Planning.
- Web usage mining can be used in Counter Terrorism and Fraud Detection, and detection of unusual accesses to secure data.

VIII. CONCLUSION AND FUTURE

This paper has tried to provide a complete review of the rapidly growing area of Web usage mining, which is the demand of current technology. In this we focused on the sequence of tasks of preprocessing of web log files. Work done by various researcher in each individual phase of preprocessing is presented .The complete preprocessing task comprises of multiple phases like field extraction ,data cleaning, user identification ,session identification, path completion. During this study, we found that the complete preprocessing strategy is although comprises of multiple phases, but all the phases and technology to implement these phases are interlinked. Detailed research in this area by researchers like Jaideep Srivastava, Bamshad Mobasher, Robert Cooley, Cyrus Shahabi, Ming-Syan Chen, and A.G. Büchner in web mining is described in detail section. In this survey we find that preprocessing stage of web usage mining is still unfolded. There are lots of open issues in preprocessing that are like Session identification, Path completion. Although we find some of the work already done in these areas but still much to be explored. In the future, visual web mining is also emerging as a new concern in the World Wide Web. So preprocessing of visual contents or multimedia should be explored. Now a day's sentiment analysis is also a very emerging field that uses web log files for sentiment analysis and opinion mining. So Preprocessing of data for any decision making process that is based on web log is required. and doing good and quality research in this field is urgently required to make internet a valuable resource for future.

REFERENCES

[1] Berendt B, Spiliopoulou. M., "Analyzing navigation behavior in Web sites integrating multiple information systems." VLDB Journal, Special Issue on Databases and the Web 9, 1 56-75. 2001.

[2] Baoyao Z, Siu C. and Alvis C, Fong M. "An Effective Approach for Periodic Web Personalization,", Proceedings of the IEEE/ACM International Conference on Web Intelligence. IEEE, 2006.

[3] Bhaskaran, V, "Data Preparation Techniques for Web Usage Mining in World Wide Web-An Approach," International Journal of Recent Trends in Engineering, Vol 2, No.4, 2009. (not valid)

[4] Buchner A, Anand S., Hughes J. & Mulvenna M.,"Discovering Internet marketing intelligence through online analytical web usage mining in SIGMOD Record, 27 (4), 54-61, 1998.

[5] Catlegde L. and Pitkow J., "Characterizing browsing behaviors in the world wide Web,", Computer Networks and ISDN systems, 1995.

[6] Chitraa.V, Dr. Davamani A,"A Survey on Preprocessing Methods for Web Usage Data" (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, No. 3, 2010.

[7] Chen J., Sun L, Zaiane O. and Goebel. R., "Visualizing and Discovering Web Navigational Patterns", Seventh International Workshop on the Web and Databases, 17-18 June 2004. Paris.

[8] Ciesielski V, and Lalani A., "Data mining of web access logs from an academic web site", In Proceedings of the Third International Conference on Hybrid Intelligent Systems (HIS'03), 2003.

[9] Cooley R, Mobasher B, Srivastava J., Knowledge and Information System, 1.Springer-Verlag, ISSN 0219-1377, 1999.

[10] Castellano G, Fanelli A., Torsello M.,"Log Data Preparation For Mining Web Usage Pattern".IADIS International Conference Applied Computing, Pg 371-378, 2007.

[11] 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.

[12] Fang Y, Li-Juan W, "Study on data Preprocessing algorithm in Web Log Mining "Proceedings of the Second International Conferences on Machine Learning and Cybernetics, Nov-2003.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 4, Issue 2, March 2015

[13] Ivancsy R, and Juhasz S,"Analysis of Web User Identification Methods." World Academy Of Science, Engineering and Technology. Vol: 1 2007-10-29.

[14] Ismail H. Toroslu, M,"Graph Theoretic Approach for Session Reconstruction Problem", Department of Computer Engineering, Middle East Technical University, 06531, Ankara, Turkey.

[15] Istvan K. Nagy and Csaba G.,"User Behaviour Analysis Based On Time Spent On Web Pages. Web Mining Application in E-Commerce and E-Services, Studies in Computational Intelligence, 2009, Volume172/2009, 117-36, DOI: 10.1007/978-3-540-88081-3_7-Springer.

[16] Jose M. Domenech and Javier L, "A Tool for Web Usage Mining" 8th International Conference on Intelligent Data Engineering and Automated Learning, 2007.

[17] Jozef K, Michal M, Martin D, "User Session Identification Using Reference Length" in 9th International Scientific Conference On Distance Learning In Applied Informatics;2012 pp-175-184.

[18] Joshila G, Maheswari V. and Dhinaharan N,"Web Log Data Analysis and Mining" in Proc CCSIT-2011, Springer CCIS, Vol 133, pp 459-469, 2011.

[19] Kushmeric .N ., "Learning To Remove Internet Advertisements, "In Third Annual Conf. on Autonomous Agents.ACM Press ,NY 1999.

[20] Khan K, Baharudin B and Khan A, "Identifying Product Features from Customer Reviews Using Hybrid Dependency Patterns "International Arab journal of information Technology (IAJIT). Vol-11, no-3.Online journal, Jan 2012.

[21] Marquardt C. Becker K, Ruiz. D "A pre-processing tool for Web usage mining in the distance education domain" in International Database Engineering and Applications Symposium, 2004. IDEAS '04.

[22] Mohammad A:"Adding new level in KDD to make the usage mining more efficient.

[23] Morzy T, Wojcie M, and Zakrazewicz M. "Web Use Clustering" International Symposium On Computer and Information Sciences, 2000

[24] Robert F, Dell P, Roman E, and Juan D., "Web User Session Reconstruction Using Integer Programming," IEEE/ACM International Conference on Web Intelligence and Intelligent Agent, 2008.

[25] Srivastava J, Cooley R, Deshphande M and Tan P,"Web usage mining, Discovery and Applications of usage pattern from Web Data" in SIGKDD Explorations 1(2):12-23, 2000.

[26] Suresh R and Padmajavalli. R."An overview of Data Preprocessing in Data and Web Usage mining,"IEEE, 2006.

[27] Sumathi, C, Vatsa R, "An Overview Of Preprocessing Of WebBLog Files For Web Usage Mining., "Journal of Theoretical and Applied Information Technology, Vol. 34, No. 1, 2011.(not valid)

[28] Sharma A,"NY Web Usage Mining: Data Preprocessing, Pattern Discovery and Pattern Analysis on the RIT Web Data" Rochester Institute of Technology, Rochester, 2008.

[29] Spilopoulou M., Mobasher B. and Berendt B. and Nakagawa M,."Framework for the Evaluation of Session Reconstruction Heuristics in Web Usage Analysis." INFORMS Journal On Computing, 2003

[30] Tanasa,.D, Trousse, B , "Advanced data preprocessing for intersites Web usage mining", Intelligent Systems, IEEE (Volume:19 , Issue: 2), , pp 59 – 65,2004.

[31] Tan. P, and Kumar, "Discovery of Web Data Mining and Knowledge Discovery,6(1),pp.9-35.

[32] Yan L, Boqin F and Qinjiao M, "Research on Path Completion Technique in Web Usage Mining," International Symposium on Computer Science and Computational Technology, IEEE, 2008.

[33] Youquan H,"Decentralized Association Rule Mining On Web using Rough Set Theory" in Journal of communication and computer ,Volume 2,No 7,Jul 2005.(Serial No.8)ISSN 1548-7709,USA.

[34] Zhang H, Liang W," An intelligent algorithm of data pre-processing in Web usage mining", Fifth World Congress on Intelligent Control and Automation, pp 3119 – 3123, Vol.4, 2004.