Volume 3, Issue 12, December 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Review on Mining Web Log Files for web Analytics and usage Patterns to improve web Organization

Sana Siddiqui, Imran Qadri
CSE, RGPV Bhopal
India

Abstract: The web access log is the best repositories for the knowledge source; it keeps the complete record of even a small event. One can easily identify web usage patterns for various web users. The Web usage pattern analysis is the process of identifying browsing patterns by analyzing the user's navigational behaviour. The web server log files which store the information about the visitors of web sites is used as input for the web usage pattern analysis process. Most of existing websites have a hierarchical content organisation. This way of organizing may be slightly different from the visitor's expectation of organizing the website. In particular, it is often unclear to the visitor at which location a specific document is present. First, these log files are pre-processed and converted into required formats so web usage mining techniques can apply on these web logs. This paper reviews the process of discovering useful patterns from the web server log file of an academic institute. The obtained results can be used in different applications like web traffic analysis, efficient website administration, site modifications, system improvement and personalization and business intelligence etc.

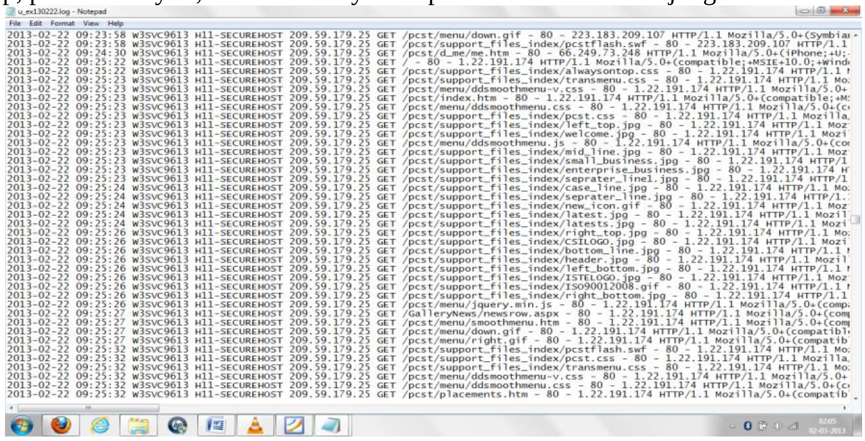
Keywords:

I. INTRODUCTION

With the technological advancements, businesses have gone online. The World Wide Web has, since then, been the ultimate and vast source of information. For example, people today can buy desired things by just clicking on a button in the computer. Because of the growing popularity of the World Wide Web, many websites typically experience thousands of visitor's every day. Analysis of who browsed what can give important insight into, for example, what are the buying patterns of existing customers. Interesting information extracted from the visitors browsing data help analysts to predict, for example, what will be the buying trends of potential customers. Correct and timely decisions made based on this knowledge have helped organizations in reaching new heights in the market. The massive data growth provides several challenges and opportunities to the user and web miners. A data stream is an ordered sequence of items that arrives in timely order. Mining steam data is a significant challenge in web data mining. Web data mining is the process to extract the interesting (nontrivial, implicit, previously unknown and potentially useful) knowledge from huge amount of data. Stream data grows rapidly, so there is an augmented need to perform pre-processing on stream data.

II. WEB LOG MINING

Web Usage Mining addresses the problem of extracting behavioural patterns from one or more web access logs [1]. the entire process can be divided into three major steps. The first step, pre-processing, is the task of accurately identifying pages accessed by web visitors. This is a very difficult task because of page caching and accesses by web crawlers. The second step, pattern discovery, involves applications of data mining algorithms to the pre-processed data to discover patterns. The last step, pattern analysis, involves analysis of patterns discovered to judge their interestingness.



```
2013-02-22 09:23:58 W3SVC9613 H11-SECUREHOST 209.59.179.25 GET /pcst/menu/down.gif - 80 - 223.183.209.107 HTTP/1.1 Mozilla/5.0+(Symbiar
2013-02-22 09:23:58 W3SVC9613 H11-SECUREHOST 209.59.179.25 GET /pcst/support_files_index/pcstflash.swf - 80 - 223.183.209.107 HTTP/1.1
2013-02-22 09:24:30 W3SVC9613 H11-SECUREHOST 209.59.179.25 GET /pcst/d_me/me.htm - 80 - 66.249.73.248 HTTP/1.1 Mozilla/5.0+(iPhone;+U;-
2013-02-22 09:25:22 W3SVC9613 H11-SECUREHOST 209.59.179.25 GET / - 80 - 1.22.191.174 HTTP/1.1 Mozilla/5.0+(compatible;+MSIE+10.0;+Wind
2013-02-22 09:25:22 W3SVC9613 H11-SECUREHOST 209.59.179.25 GET /pcst/support_files_index/alwaysontop.css - 80 - 1.22.191.174 HTTP/1.1 M
2013-02-22 09:25:23 W3SVC9613 H11-SECUREHOST 209.59.179.25 GET /pcst/support_files_index/transmenu.css - 80 - 1.22.191.174 HTTP/1.1 Mo
2013-02-22 09:25:23 W3SVC9613 H11-SECUREHOST 209.59.179.25 GET /pcst/menu/ddsmoothmenu-v.css - 80 - 1.22.191.174 HTTP/1.1 Mozilla/5.0+
2013-02-22 09:25:23 W3SVC9613 H11-SECUREHOST 209.59.179.25 GET /pcst/index.htm - 80 - 1.22.191.174 HTTP/1.1 Mozilla/5.0+(compatible;+M
2013-02-22 09:25:23 W3SVC9613 H11-SECUREHOST 209.59.179.25 GET /pcst/menu/ddsmoothmenu.css - 80 - 1.22.191.174 HTTP/1.1 Mozilla/5.0+(c
2013-02-22 09:25:23 W3SVC9613 H11-SECUREHOST 209.59.179.25 GET /pcst/support_files_index/pcst.css - 80 - 1.22.191.174 HTTP/1.1 Mozilla,
2013-02-22 09:25:23 W3SVC9613 H11-SECUREHOST 209.59.179.25 GET /pcst/support_files_index/left_top.jpg - 80 - 1.22.191.174 HTTP/1.1 Moz
2013-02-22 09:25:23 W3SVC9613 H11-SECUREHOST 209.59.179.25 GET /pcst/support_files_index/welcome.jpg - 80 - 1.22.191.174 HTTP/1.1 Mozi
2013-02-22 09:25:23 W3SVC9613 H11-SECUREHOST 209.59.179.25 GET /pcst/menu/ddsmoothmenu.js - 80 - 1.22.191.174 HTTP/1.1 Mozilla/5.0+(co
2013-02-22 09:25:23 W3SVC9613 H11-SECUREHOST 209.59.179.25 GET /pcst/support_files_index/mid_line.jpg - 80 - 1.22.191.174 HTTP/1.1 Moz
2013-02-22 09:25:23 W3SVC9613 H11-SECUREHOST 209.59.179.25 GET /pcst/support_files_index/small_business.jpg - 80 - 1.22.191.174 HTTP/1
2013-02-22 09:25:23 W3SVC9613 H11-SECUREHOST 209.59.179.25 GET /pcst/support_files_index/enterprise_business.jpg - 80 - 1.22.191.174 H
2013-02-22 09:25:23 W3SVC9613 H11-SECUREHOST 209.59.179.25 GET /pcst/support_files_index/seprater_line1.jpg - 80 - 1.22.191.174 HTTP/1
2013-02-22 09:25:24 W3SVC9613 H11-SECUREHOST 209.59.179.25 GET /pcst/support_files_index/case_line.jpg - 80 - 1.22.191.174 HTTP/1.1 Mo
2013-02-22 09:25:24 W3SVC9613 H11-SECUREHOST 209.59.179.25 GET /pcst/support_files_index/seprater_line.jpg - 80 - 1.22.191.174 HTTP/1.
2013-02-22 09:25:24 W3SVC9613 H11-SECUREHOST 209.59.179.25 GET /pcst/support_files_index/new_icon.gif - 80 - 1.22.191.174 HTTP/1.1 Moz
2013-02-22 09:25:24 W3SVC9613 H11-SECUREHOST 209.59.179.25 GET /pcst/support_files_index/latest.jpg - 80 - 1.22.191.174 HTTP/1.1 Mozil
2013-02-22 09:25:24 W3SVC9613 H11-SECUREHOST 209.59.179.25 GET /pcst/support_files_index/latests.jpg - 80 - 1.22.191.174 HTTP/1.1 Mozi
2013-02-22 09:25:26 W3SVC9613 H11-SECUREHOST 209.59.179.25 GET /pcst/support_files_index/right_top.jpg - 80 - 1.22.191.174 HTTP/1.1 Mo
2013-02-22 09:25:26 W3SVC9613 H11-SECUREHOST 209.59.179.25 GET /pcst/support_files_index/CSILOGO.jpg - 80 - 1.22.191.174 HTTP/1.1 Mozi
2013-02-22 09:25:26 W3SVC9613 H11-SECUREHOST 209.59.179.25 GET /pcst/support_files_index/bottom_line.jpg - 80 - 1.22.191.174 HTTP/1.1 M
2013-02-22 09:25:26 W3SVC9613 H11-SECUREHOST 209.59.179.25 GET /pcst/support_files_index/header.jpg - 80 - 1.22.191.174 HTTP/1.1 Mozil
2013-02-22 09:25:26 W3SVC9613 H11-SECUREHOST 209.59.179.25 GET /pcst/support_files_index/left_bottom.jpg - 80 - 1.22.191.174 HTTP/1.1 M
2013-02-22 09:25:26 W3SVC9613 H11-SECUREHOST 209.59.179.25 GET /pcst/support_files_index/ISTELOGO.jpg - 80 - 1.22.191.174 HTTP/1.1 Moz
2013-02-22 09:25:26 W3SVC9613 H11-SECUREHOST 209.59.179.25 GET /pcst/support_files_index/ISO90012008.gif - 80 - 1.22.191.174 HTTP/1.1 M
2013-02-22 09:25:26 W3SVC9613 H11-SECUREHOST 209.59.179.25 GET /pcst/support_files_index/right_bottom.jpg - 80 - 1.22.191.174 HTTP/1.1
2013-02-22 09:25:26 W3SVC9613 H11-SECUREHOST 209.59.179.25 GET /pcst/menu/jquery.min.js - 80 - 1.22.191.174 HTTP/1.1 Mozilla/5.0+(comp
2013-02-22 09:25:27 W3SVC9613 H11-SECUREHOST 209.59.179.25 GET /GalleryNews/newsrow.aspx - 80 - 1.22.191.174 HTTP/1.1 Mozilla/5.0+(com
2013-02-22 09:25:27 W3SVC9613 H11-SECUREHOST 209.59.179.25 GET /pcst/menu/smoothmenu.htm - 80 - 1.22.191.174 HTTP/1.1 Mozilla/5.0+(com
2013-02-22 09:25:27 W3SVC9613 H11-SECUREHOST 209.59.179.25 GET /pcst/menu/down.gif - 80 - 1.22.191.174 HTTP/1.1 Mozilla/5.0+(compatibl
2013-02-22 09:25:27 W3SVC9613 H11-SECUREHOST 209.59.179.25 GET /pcst/menu/right.gif - 80 - 1.22.191.174 HTTP/1.1 Mozilla/5.0+(compatib
2013-02-22 09:25:32 W3SVC9613 H11-SECUREHOST 209.59.179.25 GET /pcst/support_files_index/pcstflash.swf - 80 - 1.22.191.174 HTTP/1.1 Mo
2013-02-22 09:25:32 W3SVC9613 H11-SECUREHOST 209.59.179.25 GET /pcst/support_files_index/pcst.css - 80 - 1.22.191.174 HTTP/1.1 Mozilla,
2013-02-22 09:25:32 W3SVC9613 H11-SECUREHOST 209.59.179.25 GET /pcst/support_files_index/transmenu.css - 80 - 1.22.191.174 HTTP/1.1 Mo
2013-02-22 09:25:32 W3SVC9613 H11-SECUREHOST 209.59.179.25 GET /pcst/menu/ddsmoothmenu-v.css - 80 - 1.22.191.174 HTTP/1.1 Mozilla/5.0+
2013-02-22 09:25:32 W3SVC9613 H11-SECUREHOST 209.59.179.25 GET /pcst/menu/ddsmoothmenu.css - 80 - 1.22.191.174 HTTP/1.1 Mozilla/5.0+(c
2013-02-22 09:25:32 W3SVC9613 H11-SECUREHOST 209.59.179.25 GET /pcst/placements.htm - 80 - 1.22.191.174 HTTP/1.1 Mozilla/5.0+(compatib
```

Figure 1: Web Log File

Web server records all users' activities of the web site as web servers Logs. Most log files have text format and each log entry is saved as a line of text. There are many types of web logs such as NCSA format, W3C format and IIS format, but they share the same basic information. Log data represented in W3C extended format is shown in figure1. These log data can be used in web site designing, modifying and also to improve the overall performance of web site. After identifying the different web server log data files there is a need to merge the log files.

III. RELATED WORK

There has been considerable work on mining web logs; however, none of them include the idea of using backtracks to find expected locations of web pages.

Web usage mining [3] is referred to the discovery of user access patterns from web usage logs, which records every click made by the users. This information is frequently gathered and automatically stored into access logs through Web server. Web usage mining process is similar to data mining process. The difference is in data collection phase. The data are collected from databases for data mining whereas it is collected from web log files in web usage mining. In conventional data mining techniques information pre-process includes data cleaning, integration, transformation and reduction. But web mining pre-processing categorize into Content pre-processing, Structure pre-processing, Usage pre-processing. Once the data is collected from log files, a three-step process is performed in web usage mining namely data preparation, pattern discovery and pattern analysis.

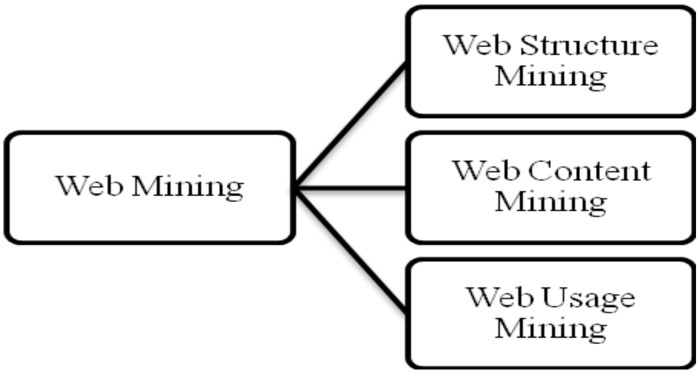


Figure 2: Taxonomy of Web Mining

Maheswara Rao [2] introduced a new framework to separate human user and search engine access intelligently with less time span. And also Data Cleaning, User Identification, Sessionization and Path Completion are designed correctly. The framework reduces the error rate and improves significant learning performance of the algorithm. Perkowitz et al. [4] [5] investigate the problem of index page synthesis, which is the automatic creation of pages that facilitate a visitor's navigation of a website. By analyzing the web log, their cluster mining algorithm finds collections of pages that tend to co-occur in visits and puts them under one topic. They then generate index pages consisting of links to pages pertaining to a particular topic. Spiliopoulou et al. [6] [7] propose a "web utilization miner" (WUM) to find interesting navigation patterns. The interestingness criteria for navigation patterns are dynamically specified by the human expert using WUM's mining language which supports the specification of statistical, structural and textual criteria.

A model called WHOWEDA (Warehouse of Web Data) has been proposed by Sanjay Madria, Sourav S Bhowmick [8] in which a discussion has been performed on various issues in web mining area. Various experiments have been performed for implementing web data as a web personalization tool [9] in which they have categorized the process of web mining in five phases i.e. i) data gathering, ii) data preparation, iii) navigation pattern discovery, iv) pattern analysis and visualization, and v) pattern applications. A model has been proposed to get the benefit of combining the Semantic Web and Web Mining [10]. Accurate Web usage information could help to attract new customers, retain current customers, improve cross marketing/sales, effectiveness of promotional campaigns, track leaving customers and find the most effective logical structure for their Web space [11]. A very good model has been proposed using decision trees which analyses the hyper links of the pages and their hierarchies of arrangements to analyse the page and their structure [08]. Some of the researchers have analysed the pattern using different algorithms like Apriori, Hash tree and Fuzzy and then we used enhanced Apriori algorithm to give the solution for Crisp Boundary problem with higher optimized efficiency while comparing to other algorithm [13]. Few have given the detailed review of web mining as another form of data mining [14]. Another aspect of web mining has been also given using two different views i.e. process-centric view which defined web mining as a sequence of tasks, and data-centric- view which defined web mining in terms of the types of web data that was being used in the mining process [15].

IV. PRE-PROCESSING OF LOG

A Web log [16] is a listing of page reference data sometimes it is referred to as click stream data. Raw web log data is not a suitable format usable by mining applications. Therefore, it is necessary to apply pre-processing techniques that may reformat and cleansed the data for mining application. The process of Web Usage Mining [17] includes three phases namely pre-processing, pattern discovery and pattern analysis. Preprocessing is a primary work in web data mining. Pre-processing [18] consists of data integration, data cleaning, user identification and session identification. It eliminates unnecessary records and validates the important records that are saved into the database, which facilitates effective data mining.

V. PRPOSED APPROACH

The proposed method consists of several phases such as file integration or merging, pre-processing, pattern discovery and pattern analysis. This paper focuses only on pre-processing phase that deals with three major issues such as data cleaning, user identification and session identification. The figure 2 shows the complete web data mining work of this paper.

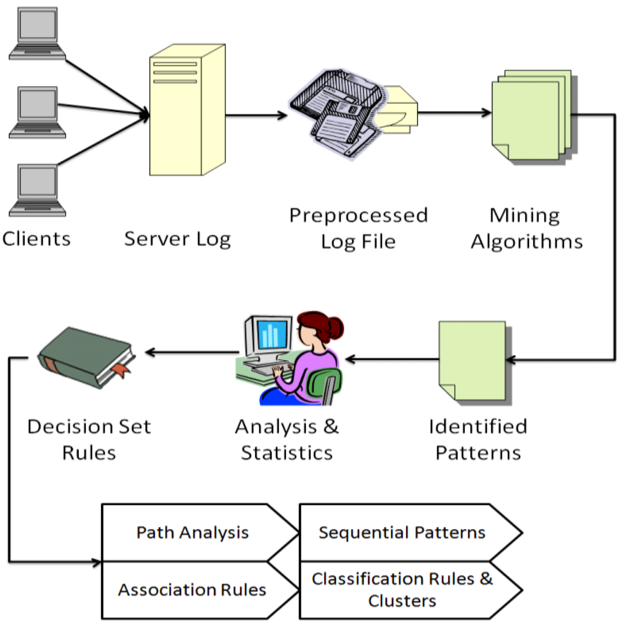


Figure 3: Proposed Methodology

Log File Integration: The integration of content, structure, and user data in other phases of the Web usage mining may also be essential in providing the ability to further analyze and reason about the discovered patterns. The integration of data mining approaches can contribute to create better and more effective intrusion detection system.

Web Usage Mining: Web usage mining attempts to discover knowledge for the data generated by the Web surfer's sessions or behaviors. Web site servers generate a large volume of data from user accesses. This data help us to determine life time value of users, to improve Web site structure design, to evaluate the efficiency of Web services. Clustering and classification on Web server log file is a process that group the users, Web pages, or user requests on the basis of the access request similarities. Association rule mining task is to discover the correlation among a variety of issues like accesses to files, time of accesses, and identities who requested the accesses. The proposed work in this paper will be carried out on different log records, which is taken from the multiple web servers. These web log records are then combined together and undergo for the pre-processing phase. Where the unwanted data and non-relevant entries of log will be removed and then user identification will be done.

Algorithm: Data Cleaning

```
begin
    while(!EOF)
    begin
    readLine();
    Check for keywords (bot, slurp, spider)
    if the line contains keyword
        begin
            botflag=true;
            botcounter++;
        end
    else
        botflag=false;
    end
end
```

After the pre-processing phase, the cleaned log file will be send to the analysis phase, where data mining algorithm will be applied to this log data to identify the log patterns using association rule mining. These association rules and extracted knowledge are then used by the website administrator to organize the web data and pages according to the popularity and usefulness.

Algorithm: Web Usage Mining

```
Inputs: I is set of itemsets, D is multiset of sub set of I Output: all frequent itemsets and all valid association rules in D
•   Level = 1; frequent_sets = ;
•   Candidate_sets = {{i} i  I };
•   While Candidate_sets ≠
•   Scan databade D to compute the frequencies of all sets in candidate_sets
    //An itemset A is closed in a data set D if there exists no proper super-itemset  B such that B has the same support count as A in D. An itemset A is a closed frequent itemset in set D if A is both closed and frequent in D.
•   frequent_sets = frequent_sets  { C  Candidate_sets  frequency(C)  min_fr};
•   level = level + 1;
•   Candidate_sets = (frequent_sets  Candidate_sets)  ( frequency(C)   min_fr and)  (  Candidate_sets  level)  ( number of combination of Candidate_sets  1) ;
•   Candidate_sets = { A  I  A= level ans B  frequent_sets for all B A,  B = level-1};
•   Output frequent_sets;
```

By applying the knowledge rules extracted from the above stated algorithms to the technique of learning user navigation patterns, the information providers would be glad to view the improvement of the effectiveness on their Web sites, which results in adapting the Web site design or by biasing the user's behaviour towards satisfying the goals of the site.

VI. CONCLUSION

With these main concerns, we decide to work for organisation website management and prepare a new reactive approach, which uses the web usage data information, site topology, academic calendar of university, in order to produce more specific process and results for organisation environment. In general this concept may of use to any website organization, A part of this various things may be of help to the system administrator like, analysis of errors helps to know the problems while accessing the website, analysis of references to website during special event will help administrator to know and balance load, analysis of navigational patterns and duration will help the administrator with the knowledge about how to decrease the duration of the user by providing layout change, decrease in duration helps to usage of less bandwidth. While applying pre-processing technique in the sample data, it is found that only 20% of the data contains useful information and the remaining 80% data is not useful for mining. From this 20% data, several frequent patterns are generated using pattern discovery techniques like association rule mining. The result is mainly useful for web page prediction and web site modifications. This work can be extended for massive web log data that is useful for strategic plan.

REFERENCES

1. F. Masseglia, P. Poncelet, and M. Teisseire. Using data mining techniques on web access logs to dynamically improve hypertext structure. In ACM SigWeb Letters,8(3): 13-19, 1999.
2. Maheswara Rao.V.V.R and Valli Kumari.V, 2011. "An Enhanced Pre-Processing Research Framework for Web Log Data Using a Learning Algorithm", Computer Science and Information Technology, DOI: 10.5121/csit.2011.1101, pp.01–15.
3. Anitha.A, 2010. "A New Web Usage Mining Approach for Next Page Access", International Journal of Computer Applications (0975-8887), Vol. 8, No.11, pp.7-10.
4. M. Perkowitz and O. Etzioni. Adaptive web sites: Automatically synthesizing web pages. In Proc. of the Fifteenth National Conf. on Artificial Intelligence (AAAI), pages 727–732, 1998.
5. M. Perkowitz and O. Etzioni. Towards adaptive sites:Conceptual framework and case study. In Proc. of the Eighth Int'l World Wide Web Conf, Toronto, Canada, May 1999.
6. M. Spiliopoulou and L. C. Faulstich. Wum: A web utilization miner. In Proc. of EDBT WorkshopWebDB98, Valencia, Spain, March 1998.
7. M. Spiliopoulou, L. C. Faulstich, and K. Wilkler. A data miner analyzing the navigational behaviour of web users. In Proc. of the Workshop on Machine Learning in User Modelling of the ACAI99, Greece, July 1999.
8. Sanjay Madria, Sourav s Bhowmick, w. -k ng, e. P. Lim, "Research Issues in Web Data Mining"
9. A.Jebaraj Ratnakumar,"An Implementation of Web Personalization Using Web Mining Techniques", Journal Of Theoretical And Applied Information Technology", 2005 - 2010 JATIT
10. Data Engineering and Automated Learning (IDEAL'07), 16-19 December, 2007, Birmingham, UK.
11. J. I. Hong, , J. Heer, S. Waterson, and J. A. Landay,WebQuilt: A proxy-based approach to remote web usability testing, ACM Transactions
12. Naresh Barsagade, "Web Usage Mining and Pattern Discovery: A Survey Paper ", December 8, 2003
13. Thales Sehn Korting, "C4.5 algorithm and Multivariate Decision Trees", Image Processing Division, National Institute for Space Research – INPES" ao Jos´e dos Campos – SP, Brazil S.Veeramalai , N.Jaisankar and A.Kannan, "Efficient Web Log Mining Using Enhanced Apriori Algorithm with Hash Tree and Fuzzy", International journal of computer science & information Technology (IJCSIT) Vol.2, No.4, August 2010
14. Mr. Dushyant Rathod, "A Review On Web Mining ",International Journal of Engineering Research and Technology (IJERT) Vol. 1 Issue 2, April – 2012 , SSN: 2278-0181

15. Jaideep Srivastava, Prasanna Desikan, Vipin Kumar, “Web Mining— Concepts, Applications, and Research Directions”, Page 400-417
16. Castellano.G, A. M. Fanelli and M. A. Torsello, 2007. “Log Data Preparation For Mining Web Usage Patterns”, International Conference Applied Computing, pp.371-378.
17. Doru Tanasa and Trousse B, 2004. “Advanced Data Preprocessing for Intersites Web Usage Mining”, IEEE Intell Syst, Vol.19, No.2, pp.59-65,DOI: 10.1109 /MIS.2004.1274912