# Application of Web Usage Mining for Administration and Improvement of Online Counseling Website

**Neeraj Kandpal[1,*], Prof. H P Singh[2], M. S. Shekhawat[3]**

[1]*Research Scholar, Suresh Gyan Vihar University, Jaipur, Rajasthan, India.*
[2]*Pro-President( Academics), Suresh Gyan Vihar University, Jaipur, Rajasthan, India.*
[3]*Department of Physics, Govt. Engineering College, Bikaner, Rajasthan, India.*

## Abstract

The successful website must take care of issues related to user satisfaction and security of website. This may be in terms of easiness of user access to contents of website or quick traversal through the website. Website administrator concerns about security and integrity of the website. The behavior of the website at the peak time of operations is the real test for the site administrator. This key information can be obtained by analysis of server log data of the website. The analysis of weblog data is the key to the success of any website. It gives the insight to activities of visitors and provides a way to the improvement of the website for better performance. The analysis of weblog data also provides valuable information for website security and development.

**Keywords:** Web Usage Mining, Web log files, Web usage mining process, Web Log mining tools.

## 1. INTRODUCTION

There are lots of data available on the World Wide Web and continuously increasing in a gigantic manner on every coming day. But this data is useful only in the form of information. The absolute goal of research and development of modern information technology should be to bridge the gap between data and information. The methods and techniques used to transform data into information collectively come under the Data Mining. Data mining is the broad concept basically contains mathematical methods. These methods applied to the raw data for automatically prediction of future events in terms of various distinct patterns, which are actually previously unknown trends and behavior. Data Mining sometimes also referenced as Knowledge Discovery in Data (KDD). Kantardzic, M. defined Data Mining as the computer-based techniques to extract knowledge and information from a huge amount of available data [1]. He suggests that using techniques of data mining, valuable and peculiar information can be obtained in the desired form from the huge amount of data.

## 1.1 Related Research Work:

With the increase in digital data in every field, web usage mining becomes the favorite area of research. There are many Web Usage Mining Tools available, which are used for analysis of weblog data. Vinod Kumar et al gave a comparative analysis of various offline and online web usage mining tools[2]. Similarly, S. Bhuvaneswari and T. Anand presented a comparative study of various Web log analyzer tools [3]. In this paper different Web Log analysis tools viz. Google analytics, Deep log analyzer, Weblog expert, Webalizer, Piwik, Open Web Analytics, AWStats compared for different features. This Paper also provides a comparative illustration of general statics tables and graphs for different web usage mining tools [3]. Researchers used Web Log Expert tool to analyze the web server log files of the website [4,5]. According to researchers information mined through Web Log Mining is very useful for presuming users interest and his behavior while traversing the website [6].

Nanhay Singh et al analyzed web log data from the "NASA" web server using Web log Explorer tool [7]. They concluded by analyzing server data that between JPG and GIF image files, JPG files use more bandwidth. So lots of bandwidth can be saved if web administrator prefers GIF files for the image media than JPG files. Thus extracted usage pattern is very useful for the web administrator for utilizing the resources of the website [7]. According to Navin Kumar Tyagi et al analysis of Web Log Files gives information about all types of errors, which can be used by the web designer to enhance the dynamism of the website [8]. Researcher [9] used the Web Log Data of jewelry website to improve the efficiency of the website.

According to Amit Pratap Singh et al, clustering and classification methods of web log mining can be used for online crime investigation like internet fraud, hacking, virus spreading, child pornography distribution, web attacks (like XPath injection, SQL injection and XSS attacks), cyber terrorism etc.[10]. According to same researchers identities of cybercriminals can be accessed by using clustering and classification techniques of weblog mining whereas crime

patterns and visualization of the network is possible by neural networks, decision trees genetic algorithm, support vector machine etc.

Olatunde Abiona et al used Webalizer web analyzer tool for analysis of access logs of six months for three different proxy servers [11]. They find that although the use of proxy server increases the security of the website but it decreases the performance, as dynamic pages cannot be processed by proxy servers. Juan D. Vela'squezm has combined eye-tracking technologies and web usage mining for the study of the behavior of visitor [12]. Using eye tracking tool he seized the ocular moments of the user visiting the website to discover the web objects, which are most significant to him. These web objects defined as Website Key objects. Researcher successfully extracted Website Key objects by Applying web mining algorithms taking website's weblogs, pages of website and interest of users in web objects as input [12].

Jayanti Mehra et al proposed a preprocessing algorithm for reducing size of web log files[13]. Researchers given novel method for recommendation system using domain knowledge and web usage[14].

Sonia Sharma et al used Web Mining Tools for recognizing behavioral patterns and various user characteristics for improving the structure of websites of different universities [15]. Andrea Dominique Cortez et al used the weblog data collected from the personal computers of four users using Chrome extension Log Catcher [16]. They found 12 distinct features, which indicate the user behavior and their browsing patterns. Some of these patterns are the number of unique tabs open, type (social, news, sports etc.) of first website open, most open webpage, Idle time between sessions, days of the week, time of day, content type etc. [16]. Neeraj kandpal et al presented a survey paper on different web mining tools used for web data analysis [17]. Same researchers elaborated the processes involved in web usage mining [18].

## 1.2 Web Mining

Web Mining is termed as the application of data mining in which mathematical techniques are applied to data collected from web server logs. So Web Mining may be defined as exposure of data mining techniques to weblog data to find the probability of future events like potential cyber-attack or incessant tries by an intruder to crack the password of the website.

Web mining is part of the data mining process broadly divided into three parts i.e. Web Content Mining, Web Structure Mining and Web Usage Mining.

*Web Content Mining:*

Web Content Mining is the extraction of contents of web documents exist in the web page like text, images, audio, video or other information about the structure of web page like tables, lists etc. for generation of useful information and knowledge from it. For this web content mining uses techniques of Information Retrieval (IR) and Natural Language Processing (NLP). Image Recognition is the latest area of web content mining. There are two points of view of Web content mining viz. Information retrieval view and Database view. Information retrieval view defines the representation of extracted information of web contents in terms of specific knowledge. Database sees the website as a reservoir of information and stores information in a database like a view.

*Web Structure Mining:*

It is the procedure to describe the structure of the website including the position of different nodes, hyperlinks used in the website. It uses the graph theory to explain the structure of the webpage. Web structure mining can be branched into i) mining of document structure of website and ii) finding patterns from hyperlinks used in the website. The algorithm PageRank used to rank search results assured by a total number of links pointed towards a specific node.

*Web Usage Mining:*

Web usage mining is the data mining process of weblog data. It gives the accurate picture of the behavior of users of the website. Analysis of this data is very useful for the research and development process related to the website. Threats of cyber-attack like an attempt of hacking, phishing etc. can be early detected by proper use of Web Usage Mining.

## 1.3 Web Log Mining Tools

There are many tools available for Web Usage Mining. Some of these are Webalizer, Analog Stats, AWStats, Web Log Explorer, Web Log Expert etc. We used different weblog mining tools like Web Log Expert, Webalizer, AWStats, Analog Stats for analysis of Web Log Data of website 'rajeducon.com'[21]. Web Log Expert is a widely used log analyzer [22]. It provides information about the website's visitor activity statistics, accessed files by visitors, visitors paths through the site, information about referring pages, search engines, browsers, operating systems etc. The program produces information-rich reports that include text, tables and charts.

## 1.4 Introduction to Counseling website 'rajeducon.com' used for Data Collection and online Counseling Process:

In this paper, we have done an analysis of server log data obtained from website 'rajeducon.com', which is an official website for Counseling for posting of teachers in government schools of Rajasthan, India. The Counseling process conducted by Directorate of secondary education, Rajasthan, Bikaner, India for Department of Secondary Education, Government of Rajasthan. In the Counseling process,

successful candidates find a chance to choose the school of their priority for posting according to their merit number in required competition examination or according to departmental seniority for promotion.

Rajasthan is the biggest state of India situated at the northwestern part of it. It has an area of 342,239 sq. km and comprises of 33 districts having nearly 44700 villages. Rajasthan has an atypical rhomboid shape and stretches lengthwise 869 km. from west to east and 826 km. from north to south. Department of Secondary Education, Government of Rajasthan runs more than 14280 schools across the state and nearly 4800000 students studying from class one to class twelve in these schools. There are nearly 230000 teachers posted in these schools. Website 'rajeducon.com' used to online display of vacancies at the day of posting by Counseling for transparency in posting and also for real-time information sharing to all candidates appeared in Counseling for posting in various schools.

Website 'rajeducon.com' is used for real-time updation of vacant posts available for Counseling. All the vacant posts available in schools uploaded 1-2 days before counseling. At the day of counseling as the candidate of higher rank selects a school, the same is updated real-time as unavailable for remaining candidates at the website. This gives accurate and real-time information for the rest of the candidates so that they can change their priority. Many times counseling process for one subject goes on more than one day, as once counseling of lecturer (Hindi) conducted up to eight days. Candidates from all Rajasthan come for Counseling to Bikaner according to predefined schedules. Any candidate at any place can have a close watch to the whole process by using the website 'rajeducon.com'. The website 'rajeducon.com' is also used by government and education officers for monitoring purpose.

We have used server log data of counseling website ''rajeducon.com'' for different time period range from January-2017 to June-2018. Analysis of this data has been done by using different web data mining tools like Web Log Expert, Webalizer, AWStats, Analog Stats etc.

**1.5 Web Log/Usage Data Files Used for Analysis:**

Web Log Data of user's navigation obtained from a server used in weblog mining. Web Log Data or Web Usage data can be obtained from three different places [20], a web server, a web proxy server or in the client browser. So types of web usage data files are Web Server Log Files, Web Proxy Server Log Files and Client Browser Log Files. Web Server Log Files are most accurate one as it is free from cached data. It contains transfer logs, agent logs, error logs and referrer logs. Web Proxy Server Log Files have the disadvantage of cached data. Client Browser Log Files resides at user browser. The web server stores all the movements of visitors at the website. Weblog data is very useful in website designing and improvement, the creation of personalized websites, business and support services, network traffic flow analysis, target marketing, resolution of speed and security-related issues and many more.

Web Log File Example from counseling website 'rajeducon.com':

Various terms in web log file in Table-1 are:

❖ 106.199.59.48 - Unique IP address of Visitor.

❖ - - - Unidentified user id field.

❖ [30/Jun/2017:18:15:46 +0530] - Date, time and zone entry of web log.

❖ GET - The request method of visitor.

❖ /image/12.jpeg - Resource requested by user.

❖ HTTP/1.1 - Protocol used to request page.

❖ 200 - Status code returned to user by server. Current status code shows success.

❖ 56775 - The bytes sent to user from server as requested by user.

❖ "http://'rajeducon.com'/" - URL of the reference page.

❖ "Mozilla/5.0 (Linux; Android 4.4.2; en-us; SAMSUNG SM-T231 Build/KOT49H) AppleWebKit/537.36 (KHTML, like Gecko) Version/1.5 Chrome/28.0.1500.94 Safari/537.36" - Browser name, it's version and the operating system used.

**Table 1:** Access Log File

| SNo | Sample Weblog |
|---|---|
| 1. | 106.199.59.48 - - [30/Jun/2017:18:15:46 +0530] "GET /image/12.jpeg HTTP/1.1" 200 56775 "http://'rajeducon.com'/" "Mozilla/5.0 (Linux; Android 4.4.2; en-us; SAMSUNG SM-T231 Build/KOT49H) AppleWebKit/537.36 (KHTML, like Gecko) Version/1.5 Chrome/28.0.1500.94 Safari/537.36" |

**Figure 1:** Web Log File

Web Server Log Files are plain text files which store the information about each and every visit of the user of the website. Log Files are of four types [20]. (i) Access Log Files contains all information generated by each and every single click stream of a visitor at the website. Example access log file of website 'rajeducon.com' given in Table-1. (ii) Agent Log Files contains information about user's browser, operating system and version of the browser. (iii) Error Log Files contains information about any error encountered during the visit of the user.  (iii) Referrer Log Files gives information of referring sites from where user diverted to the website. It is used for promotional or security purpose. Out of four log files, Agent and Referrer log files not always included in the server system.

There are three types of formats described for web log files [19]. (i) W3C extended log File Format or Common Log format (CFL) is ASCII text file format, which can be modified according to requirements of the concerned website. Fields necessary for a website can only be included and rest may be excluded in W3C web log file format. Whereas (ii) Microsoft IIS (Internet Information Services) and (iii) NCSA (National Center for Superconducting Applications) log file formats are the fixed ASCII text log file formats i.e. we can't modify these formats according to our need.

## 2.  ISSUES RELATED TO 'rajeducon.com'

On the day of Counseling the website, 'rajeducon.com' became very slow, particularly during day time. This was reported by many users. Complaints about slow opening are prominent among others. There are some complaints about errors 404 and 508 at the day of Counseling. Once we have encountered the complete breakdown of 'rajeducon.com' due to exhaustion of resource provided by the hosting provider. We find from analysis of reports of the server that all entry processes used by visitors and because of that now entry of new user becomes very difficult. We also want to know about most valuable page of our website to publish achievements of education department.

Due to the above reasons there is slower access to online vacant posts, which is the main purpose of the Counseling website.

1434

## 3.  RESULT AND DISCUSSIONS

Data from counseling website 'rajeducon.com' have been taken for a different period of time between January-2017 to June-2018. Weblog data used for analysis extracted for months of important Counseling activities, which contains a relatively larger volume of traffic. This data have been analyzed by using different weblog mining tools viz. Web Log Expert, Webalizer, AWStats, Analog Stats.

The important results visible from analysis of weblog data are:

1) According to AWStats for the month of June-2017 website 'rajeducon.com' encountered a whopping 6,476,322 hits using the total bandwidth of 207.80 GB. In this duration 2,477,059 pages fetched by the 117,748 unique users in total 379,934 visits. It is not included the traffic generated by robots, worms and or replies with special HTTP status code.

2) According to the reports of Web Log Experts web mining tool about most requested directories for July-2017, directory http://rajeducon.com/image/ accessed 73235 times out of total 156717.

3) From the reports of June-2018, it is found that nearly 70% bandwidth of total bandwidth consumed, exhausted by image files used as scroll images at the front page of the website.  It's whopping 30 GB out of total 42 GB data used in June-2018 month.

4) According to daily Statistics reports from Webalizer web mining tool for June 2017, the lots of hits attracted by the old counseling links made for the earlier counseling processes. This report also claims that pdf documents used for introduction of counseling process and contacts using some bandwidth.

5) A report from Webalizer and AWStats Web Mining tool suggests that most traffic generated in between 10 am and 2 pm.

6) It is found that a large number of users (nearly 86%) accessed the website using mobile Figure-2.

7) Different robots and spiders activities visible in reports generated by Web Log Expert Tool and other tools. Maximum hits encountered by Google bot. Hits from various countries other than India is due to traffic from various bots Figure-3.
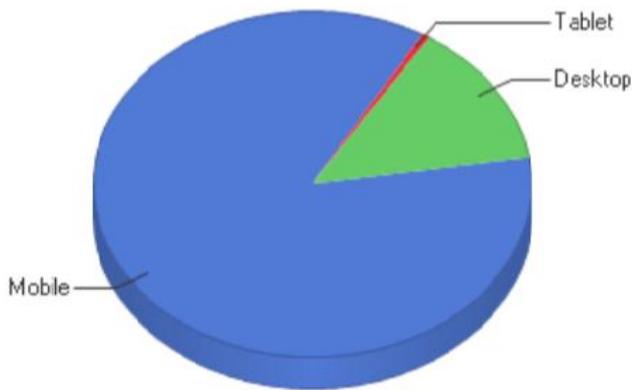


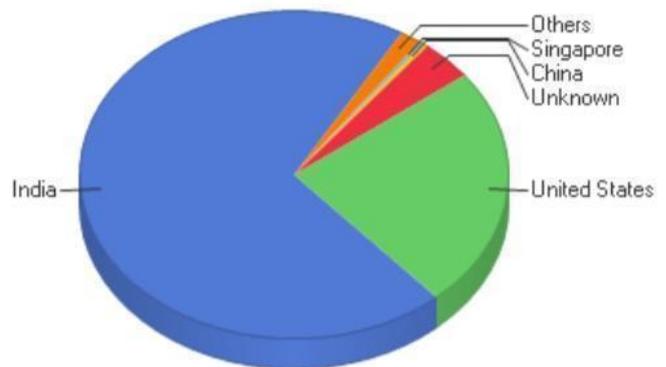**Figure 2:** Device used to access website



**Figure 3:** Hits from different countries

## 4.  ANALYSIS OF RESULTS

By analysis of different reports, we can suggest that-

1. On the day of Counseling due to the heavy traffic (data of the day of Counseling) website becomes slower. By analyzing server log data for the day of Counseling, following measures has been taken.

   a. Image files consumed lots of data which results in the slow opening of pages. At the day of Counseling when a lot of users accessing website simultaneously we have temporarily disabled the scroll image code to speed up the loading of the homepage of website 'rajeducon.com'.

   b. Links of the PDF files meant for the introduction of process and contact information removed from the home page temporarily.

   c. Some links of earlier Counseling processes also targeted by the visitors just for curiosity again resulting loss of bandwidth. So links of earlier Counseling have been also deactivated at the peak time of operations for saving bandwidth for normal operations of the online Counseling process.

Comparison Table-2, Showing drastic improvement in performance of website 'rajeducon.com' by the implementation of results of research in terms of increase in visitors count. Analysis has been made for the same counseling session and for the same period of time. (After July 2017, we are regularly implementing the measures as given in point-1. For research and development purpose, we have created earlier conditions for selected months and equal time duration. Sample data for analysis taken from the Visitors log of cpanel of website 'rajeducon.com')

**Table 2:** Comparison table

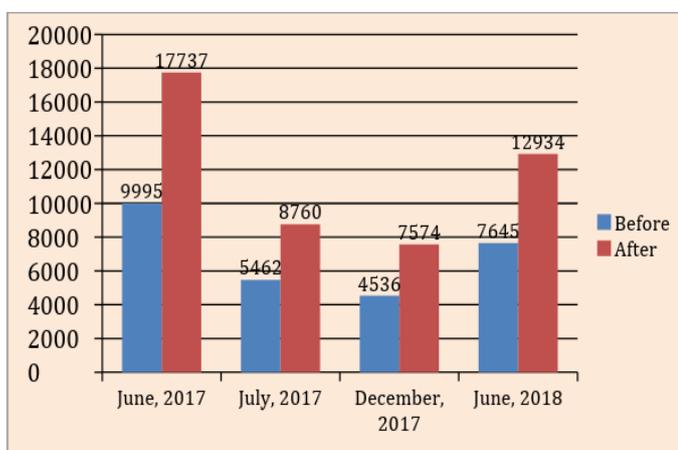| S.No. | Month and Year of Data Collection. | Visitors before applying measures | Visitors after applying measures | % increase in Visitor |
|---|---|---|---|---|
| 1. | June-2017 | 9995 | 17737 | 77 |
| 2. | July-2017 | 5462 | 8760 | 60 |
| 3. | December-2017 | 4536 | 7574 | 66 |
| 4. | June-2018 | 7645 | 12934 | 69 |



**Figure 4:** Comparison Curve

It is obvious from the Table-2 and Figure-4 comparison curve that there is an increase in visitor count after application of measures suggested by the research.

2. Due to larger volumes of mobile users, modification of site has been done to make the site more mobile user-friendly.

3. Detailed analysis of error logs from the web server of 'rajeducon.com' has been done.  Pages appeared in error logs corrected or removed for better performance of the website 'rajeducon.com'.

4. Lots of users encountered the error 404 due to the old counseling links. The error 404 appears, when a visitor tries a URL, which has been already removed. We have edited a common error code page for error 404 to inform the user that they are using an expired link to access the website.

5. The information about maximum traffic may be used for variable advertisement rates purpose for revenue generation.

6. The information about most visited page is useful for advertisement purpose.

## 5.  CONCLUSION

The analysis of web log data of website 'rajeducon.com' reveals many hidden facts which are useful for the improvement of website. We have successfully reduced the load on website on peak time of operations by deactivating scrolled image tabs and pdf documents from homepage of the website as suggested by results of research. The links of old counseling removed from homepage to enhance traversal speed of user. Using results of analysis of weblog data from different web mining tools, the program code improved to remove loops formed, which were exhausting the entry processes. Speed of traversal through website 'rajeducon.com' increased significantly by applying results of research work. Reports of entry and exit pages are very important for analysis of usefulness of pages for visitors. There is a need for modification of the page mostly appearing in the list of exit page.

## REFERENCES

[1] Kantardzic, M. (2011). Data Mining: Concepts, Models, Methods, and Algorithms, Second Edition. Institute of Electrical and Electronics Engineers. John Wiley & Sons, Inc.

[2] Vinod Kumar, Ramjeevan Singh Thakur, A Brief Investigation on Web Usage Mining Tools (WUM), Saudi Journal of Engineering and Technology. DOI:10.21276/sjeat.2017.2.1.1

[3] S. Bhuvaneswari and T. Anand, A Comparative Study of Different Log Analyzer Tools to Analyze User Behaviors, International Journal on Recent and Innovation Trends in Computing and Communication, Volume: 3 Issue: 5, 2015, 2997 - 3002.

[4] Arvind K. Sharma and P.C. Gupta, ANALYSIS OF WEB SERVER LOG    FILES TO INCREASE THE EFFECTIVENESS OF THE WEBSITE USING WEB MINING TOOL, International Journal of Advanced Computer and Mathematical Sciences, Vol 4, Issue 1, 2013, pp1-8.

[5] Karthik, S Padmaja & Sheshasaayee, Ananthi. (2014). WEB SERVER LOGS TO ANALYZING USER BEHAVIOR USING LOG ANALYZER TOOL. IJARSE. 3. 514-525.

[6] Navjot Kaur and Himanshu Aggarwal, Web Log Analysis for Identifying the Number of Visitors and their Behavior to Enhance the Accessibility and Usability of Website, International Journal of Computer Applications , Volume 110 – No. 4, p-0975 – 8887, January 2015.

[7] Nanhay Singh, Achin Jain and Ram Shringar Raw, COMPARISON ANALYSIS OF WEB USAGE MINING USING PATTERN RECOGNITION TECHNIQUES, International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.4, July 2013.

[8]     Navin Kumar Tyagi, A. K. Solanki and Manoj Wadhwa, Analysis of Server Log by Web Usage Mining for Website Improvement, International Journal of Computer Science Issues, Vol. 7, Issue 4, No 8, July 2010.

[9]     Jigar H. Jobanputra, Banshi D. Soni, Enhancing The Efficiency of the Website by Mining Web Server Log, International Journal for Scientific Research & Development, Vol. 3, Issue 12, 2016.

[10]    Amit Pratap Singh , Dr. R. C. Jain, A Survey on Different Phases of Web Usage Mining for Anomaly User Behavior Investigation, International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Volume 3, Issue 3, May – June 2014.

[11]    Abiona, O., Oluwaranti, A., Oluwatope, A., Bello, S., Onime, C., Sanni, M. and Kehinde, L. (2014), Proxy Server Experiment and Network Security with Changing Nature of the Web. Int. J. Communications, Network and System Sciences, 7, 519-528.

[12]    Juan D. Vela'squez, Combining eye-tracking technologies with web usage mining for identifying Website Keyobjects, Engineering Applications of Artificial Intelligence 26 (2013) 1469-1478.

[13]    Jayanti Mehra, Dr. R S Thakur, An Effective method for Web Log Preprocessing and Page Access Frequency using Web Usage Mining, International Journal of Applied Engineering Research, Volume 13, Number 2 (2018) pp. 1227-1232.

[14]    Nguyen, Sang & Lu, H.Y. & Lu, Jie. (2014). Web-Page Recommendation Based on Web Usage and Domain Knowledge. Knowledge and Data Engineering, IEEE Transactions on. 26. 2574-2587. 10.1109/TKDE.2013.78.

[15]    Sonia Sharma and Munishwar Rai, Customer Behaviour Analysis using Web Usage Mining, International Journal of Scientific Research in Computer Science and Engineering, Vol.5, Issue.6, pp.47-50, December (2017)..

[16]    Andrea Dominique Cortez, Paolo Gabriel Gamab, Joseph Emmanuelle Julian, Benjamin Gabriel Tan, and Jocelynn Cu, Extracting Features from Web Logs for Web Usage Mining, Presented at the DLSU Research Congress 2017 De La Salle University, Manila, Philippines June 20 to 22, 2017.

[17]    Neeraj Kandpal, Prof. Ripu Ranjan Sinha, M. S.Shekhawat , A Survey on Web Usage Mining:Process, Application and  Tools, Suresh Gyan Vihar University Journal of Engineering & Technology, Vol . 3, Issue 1, 2017, pp 19-25.

[18]    Neeraj Kandpal, Prof. Ripu Ranjan Sinha, M. S. Shekhawat, A Study of Processes Involved in Web Usage Mining, International Journal of Allied Practice, Research and Review, Vol. III, Issue XII, p.n.01-05, Dec, 2016.

[19]    Shaily Langhnoja, Mehul Barot, Darshak Mehta,"Pre Pre-Processing: Procedure on Web Log File for Web Usage Mining", International Journal of Emerging Technology and Advanced Engineering, Volume 2, Issue 12, December 2012.

[20]    Hanane EZZIKOURI, Mohammed ERRITALI, Mohamed OUKESSOU, "Pretreatment of web log files", Journal of Information Sciences and Computing Technologies, Volume 2, Issue 1, February, 2015.

[21]    Counseling Website of Department of Secondary Education, Rajasthan, India; https://www.rajeducon.com/

[22]    Web log Analyzer tool; https://www.weblogexpert.com/