# *Analysis of visitor's behavior from Web Log using Web Log Expert Tool*

Manoj Kumar
Computer Science and Engineering Department
Madan Mohan Malaviya University of Technology
U.P India
Email - manojkumaritmgida@gmail.com

Mrs. Meenu
Computer Science and Engineering Department
Madan Mohan Malaviya University of Technology U.P
India
Email - myself meenu@yahoo.co.in

**Abstract**

Web usage mining is a data mining technique. There are large amount of data are stored on the internet. When user search any particular information by search engine like Google, Bing etc. is very difficult because the complexity of web pages is increases day by day. Web usage mining plays an important role to solve this problem. In web usage mining we are creating a suitable pattern according to the user's visiting behavior. The goal of this paper is to implement a web log Expert tool on web server log file (an educational institution web log data) to find the behavioral pattern and profiles of users interacting with a web site. The web mining usage pattern of an Technical Institution web data. Web related data is coteries in to three parts namely web log, access log, error log and proxy log data and collect the data in web server and implemented a web log expert. Our experimental results help to predict and identify the number of visitor for the website and improve the website usability. The web related log data are three types, namely proxy log data, web log data, and error log data. We exploration the activity statistic by daily based hourly based week and monthly based report of web usage pattern. The web usage mining is playing an important role to improve the availability of information of your web site.

**Keywords—Web Usage Mining, web server log, web log Analyzer**

## I. INTRODUCTION

Web usage mining plays an important role for extracting useful information and discovering sutable pattern. These pattern are very useful when user search any particular web based application. Web usage mining is some time called web log mining. It have three integrated phases namley pre processing, pattern discovery, and pattern analysis[1]. These three phases are take input data (log server data) one by one. Data Preprocessing: this is the iniseal phase of web usage mining. It takes log data as a input perform some task like user identification, data cleaning, session identification and path completion.Pattern discovery: this is an most important phase it take input from preprocessing output. And perform some task like clustering, statical analysis, association rule and classification and to find interesting pattern.Pattern analysis: oftre performing upper two then analysis is done using

knowledge query tools like SQL or data cubes to perform OLAP operations This research using data from visitor engagement system web log files that generated from web server IIS (Internet Information Services) in 2 Jan 2017 and using the accessed web address page references and access time.

### A. Web Usage Mining

Web mining is categories in to three parts. Web usage mining is the third category in web mining. This is permits for capture the web access information for web pages. This usage data provides the paths leading to accessed web pages. This information is often gathered automatically into access logs via the web server. CGI scripts offer other useful information such as referrer logs, user subscription information and survey logs. This category is important to the overall use of data mining for companies, institutions and there internet based applications and information access. Web usage mining has also three sub steps namely preprocessing, pattern discovery and pattern analysis.

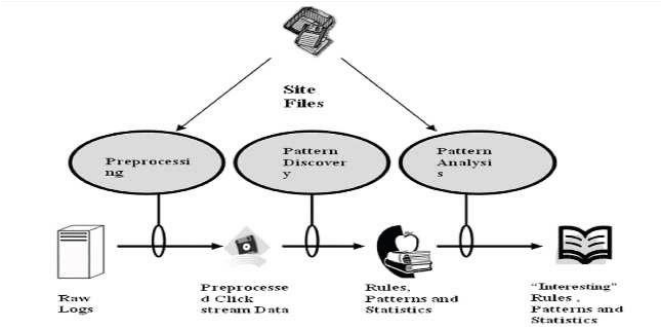


Fig 1. Web usage mining process

**a. Data Preprocessing** It is the in itial phase of web usage mining. The web log data is a raw data it not directly used for information .In this phase, we applying technique to

transform raw data into an understandable format. Real-world data is repeatedly incomplete, erratic, and lacking in certain behaviors or trends, and is likely to contain many errors. Data pre processing is a proven method of resolving such issuse. Data pre processing prepares raw data for further processing.

**b. Pattern Discovery**

The rersults from pre-processing will be used to find frquent user acess pattern . In pattern discovery will be use different data mining technique like as association rule, classifiaction , clustring , and sequential pattern technique to find important information . The results that has been extracted can be represebted in may ways such as graphs , charts , table, etc. .

**c. Pattern Analys**

The outcome of pattern discivery phase is not directly used foranalysis. so , in this phases will develop a technique or tool that can help analysts understand the information has been extracted . tools or techniques that can be used in this phase like visualization techniques , OLAP analysis and knowledge query mechanism.

**B. Web Server Log**

This file is a type of log file which spontaneously formed and preserved by a web server. Each hit to the web site, including image, every observation of HTML document or additional object, is logged. The fresh web log file layout is basically one online of text for every hit to the web site. This comprises data about how was browsing the site. where they come from, and exactly what they were doing on the web site. We are take the log data from an informative instituion web server log. Web server give a article and using this article it can be criate a data log file .This Log files store all information about visitor requests activity webpages , that requests will be added to the existing logg file . log files usually have in clear text format and each log entery will be shorted in a line of text .

## II. PROPOSED METHODOLOGY

Web usage mining is an application of data mining. In this work we are impliment an web log expret tool to finding user visiting behavier on web brausers. We are take input data(web sesrver log ) from an educatoinal inistitution server log. To find the witch type of web resouses are udsed. Like operatnig system, search engens, brausers etc. on the basis of this behavier we are discover a pattern that is give the better result when user serching web based application. This paper present visitor pattern analysis performed through educational institution web log data.

The main aim is to find user activity information

. Visitor browser analysis
. Visitor page view analysis
. Visitor OS analysis
. visitor time analysis

In this work, the web log expert tool reports of the time analysis and page view analysis. The time analysis at the different time of day, day of week, and days of month that the website receives the most visitor. *A Framework*

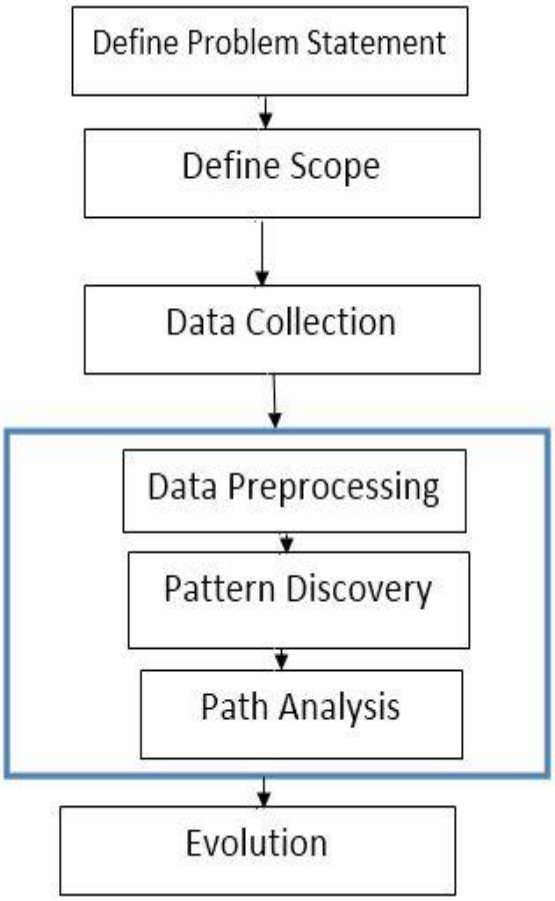


Fig 2. Research Framework

From fig.2 above, framework for this research are as follow: The fist step is to identify the problens that is being experieced by the user behavior on web site. Institution requires the behavior data of visitor's using the application to evaluate application perfomance. For retrieved data, institution can used web usage mining to gather that data.

The next is to determine the benifet of the work that is we want analyzes the behavioral patterns and profiles of users interacting with the web site.

The third step is to find the data from web log file.

The fourth and fifth step is to analyze the log data by using the web log Expert too.

***Web usage mining phases***

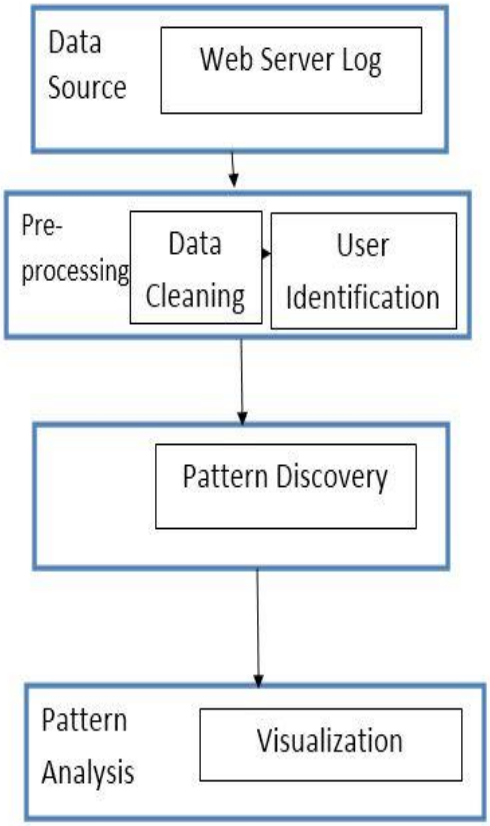


*Fig.3 phases of WUM*

**A .** ***Data Collection***

The source of data collection is the server access log data that is collected to an educational institution. This web server log data has the information of 4 days (2 jan 2017). In this prosses 1.8MB data was transferred. 6957 entries are exist in the log file. In this work we are take server log file of an university,

**Software: Microsoft Internet Information Services 7.0**
**#Version: 1.0**
**#Date: 2017-01-02 12:58:29**
**#Fields: date time s-site name s-computer name s- ip cs-method cs-uri-stem cs-uri-query s-port cs-username c-ip cs-version cs(User-Agent) cs(Cookie) cs(Referer) cs-host sc-status sc-sub status sc-win32-status sc-bytes cs-bytes time-taken**
**2017-01-01 00:00:06 W3SVC770 SVMSVR1 184.172.12.53 GET**

**/GoldenJubilee/45214_MEDC_11292016.pdf - 80 - 180.76.15.26 HTTP/1.1**

**Mozilla/5.0+(compatible;+Baiduspider/2.0;++http://www.baidu.com/search/spider.html)**

for extracting useful information. The web log data contains attributes. These attribues are as follows

**Date-** Format take are Greenwich Mean Time (GMT× 100) is recorded for each hit. The date is presented in form of (Year/Month/Day).

**Time period -** Time period refers of transaction. The format of time is given in (Hour :Minute :Second).

**Server IP-** server IP Address is a static IP provided by internet serves provider. The IP will be a reference for access the information from web server.

**Server port-** it is ports for data transaction.

**Server method (HTTP Request)**- this is persent to a text, video, sound, , html file, pdf etc.

**URL**- this is the path of visitor. This is represent the structures of web site that user want to search.

**Agent Log-** It is providing data on a user's browser, browser version and operating system.

## III. TOOL FOR EXPERIMENT

There are many different types of tools are present to analyze a web server log file and generating the reports .some tools are free available on the internet. Some of tools are Google analyzer, Stat Counter, Deep log Analyzer. in this work we are select Web Log Expert tool [3]. This tool takes log data as a input and generate report of all information for visitors accessing the site. The IP address, time, zone, URL, browser, OS of the user's. The Web Log Expert installation is easy highly user friendly. It will be give the information about the site visitor's activity such as: activity statics, accessed file, paths through the site, information about referring pages, search engines.

Some feature of web log expert is shows below [5].

. it provide a log information of visitor, activity statistics and access activity .

. It supports the IIS and apache logs.

.Automatically detects log format.

. it is reads GZ and ZIP compressed logs.

. Create reports that include text information and charts.

## IV. EXPERIMENTAL RESULTS

The input data (web log data) are collected from an technical Institution. That has five days information of Jan 2017.

1. **General Activity.** The general activity statics of web site shown in the Table-1 there are 60233hits, 59899 visitors,1398IPs, 2797 pages views.

**Table-1 general activity statics of web site**

| | |
|---|---|
| **Hits** | |
| Total Hits | 60,233 |
| Visitor Hits | 59,899 |
| Spider Hits | 334 |
| Average Hits per Day | 60,233 |
| Average Hits per Visitor | 48.31 |
| Cached Requests | 9,500 |
| Failed Requests | 1,404 |
| **Page Views** | |
| Total Page Views | 2,797 |
| Average Page Views per Day | 2,797 |
| Average Page Views per Visitor | 2.26 |
| **Visitors** | |
| Total Visitors | 1,240 |
| Average Visitors per Day | 1,240 |
| Total Unique IPs | 1,195 |
| **Bandwidth** | |
| Total Bandwidth | 5.09 GB |
| Visitor Bandwidth | 5.04 GB |
| Spider Bandwidth | 56.67 MB |
| Average Bandwidth per Day | 5.09 GB |
| Average Bandwidth per Hit | 88.82 KB |
| Average Bandwidth per Visitor | 4.16 MB |

Table shows the information in four types: first Hits, second page view, third visitors and fourth bandwidth. The first of the number of that is Hits: the number of visitor Hits, spider Hits, Average Hits per Day, Average Hits per visitor, Cached Requests, Failed Requests. Second is common statistics the shows table the number of Average page view per day and per visitor. Third is visitor and that is also provide the information of Usual visitor each Day. The last is the Bandwidth of visitor, Spider Bandwidth, Average Bandwidth per Day. This state reduces the info of total usage accessibility of website.

**2 Activity Statistics**

It give the information about daily as well as hourly activity of the visitors log file.

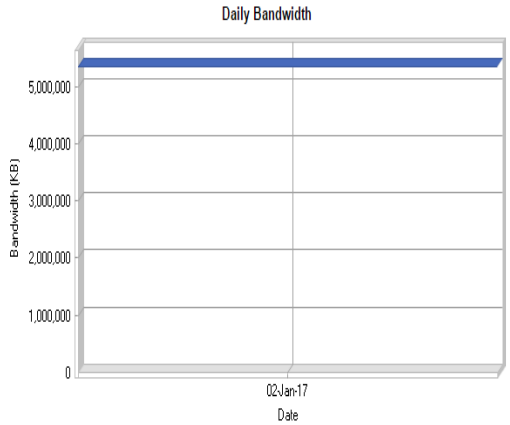


**Fig.4. Daily Website Visitors Report**

The daily activity of web site shown in table 2. The number of hits **63,059,** page view 3,256, and visitor 1,520, Average visit Length 05:17 and bandwidth, 3,660,952.

**Table -2: Daily Activity Statistics of the Website Usage**

Daily Activity

| Date | Hits | Page Views | Visitors | Average Visit Length | Bandwidth (KB) |
|---|---|---|---|---|---|
| Mon 02-Jan-17 | 60,233 | 2,797 | 1,240 | 05:12 | 5,337,730 |
| Total | 60,233 | 2,797 | 1,240 | 05:12 | 5,337,730 |

The number of hits, number of page viewers and in rate of data transferred the best day out of this day of log data in Monday 2/01/2017.

**Activity by Hour of Day**

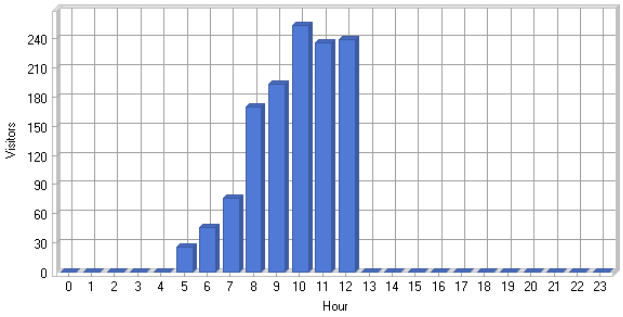


**Fig.5. Hourly Website Visitor Report**

The hourly basis of report of website visitor in the shown figure 5. It shows number of hits per hour, page views per hour, visitor per hour and bandwidth in KB per hour.

**Table -3: show the hourly behavior of visitor on website.**

Activity by Hour of Day

| Hour | Hits | Page Views | Visitors | Bandwidth (KB) |
|---|---|---|---|---|
| 00:00 - 00:59 | 0 | 0 | 0 | 0 |
| 01:00 - 01:59 | 0 | 0 | 0 | 0 |
| 02:00 - 02:59 | 0 | 0 | 0 | 0 |
| 03:00 - 03:59 | 0 | 0 | 0 | 0 |
| 04:00 - 04:59 | 0 | 0 | 0 | 0 |
| 05:00 - 05:59 | 373 | 31 | 26 | 23,292 |
| 06:00 - 06:59 | 1,429 | 95 | 46 | 76,321 |
| 07:00 - 07:59 | 2,207 | 116 | 76 | 205,132 |
| 08:00 - 08:59 | 7,456 | 365 | 170 | 621,035 |
| 09:00 - 09:59 | 9,645 | 393 | 193 | 903,583 |
| 10:00 - 10:59 | 12,315 | 567 | 254 | 984,819 |
| 11:00 - 11:59 | 13,515 | 616 | 236 | 1,458,597 |
| 12:00 - 12:59 | 13,293 | 614 | 239 | 1,064,948 |
| 13:00 - 13:59 | 0 | 0 | 0 | 0 |
| 14:00 - 14:59 | 0 | 0 | 0 | 0 |
| 15:00 - 15:59 | 0 | 0 | 0 | 0 |
| 16:00 - 16:59 | 0 | 0 | 0 | 0 |
| 17:00 - 17:59 | 0 | 0 | 0 | 0 |
| 18:00 - 18:59 | 0 | 0 | 0 | 0 |
| 19:00 - 19:59 | 0 | 0 | 0 | 0 |
| 20:00 - 20:59 | 0 | 0 | 0 | 0 |
| 21:00 - 21:59 | 0 | 0 | 0 | 0 |
| 22:00 - 22:59 | 0 | 0 | 0 | 0 |
| 23:00 - 23:59 | 0 | 0 | 0 | 0 |
| Total | 60,233 | 2,797 | 1,240 | 5,337,730 |

Show tables the accurate information of hourly activity of web site.

**3 Access Activities**

It is provider of information of the most popular page, most downloaded files and most requested image Figure6. Is the daily page access and Figure 7 shows the result of most popular page of web site after analyzing log file.

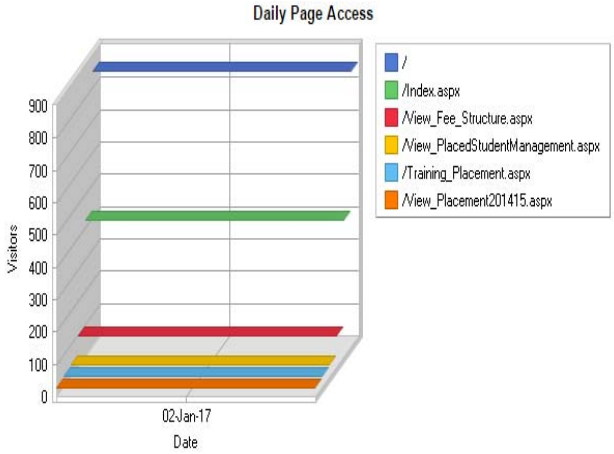


**Fig6.Dailypageaccess**

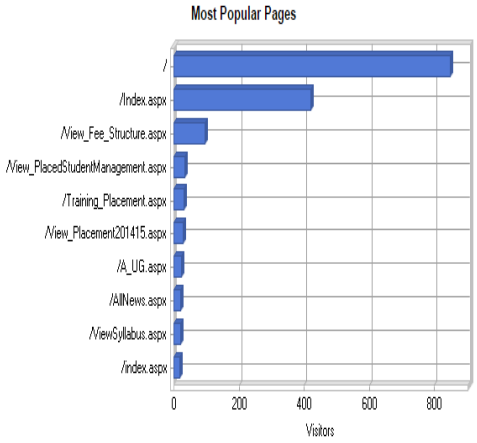


**Fig7. Most Popular Page**

The maximum number of files that are downloaded the most number of time from the file page of the web site shown in figure 8.

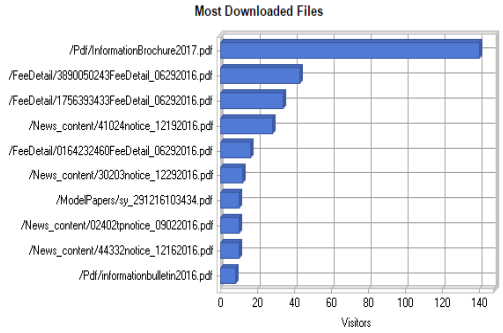


**Fig8. Most downloaded files**

The figure 9 shown the image those have been downloaded most during this month. The most. Next image logo.png.

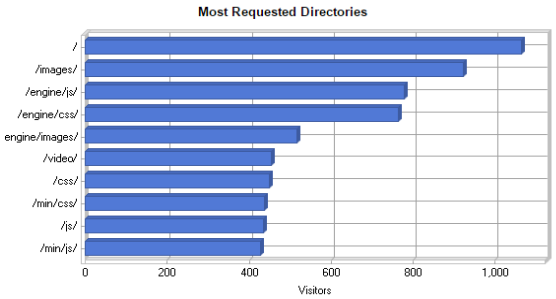


**Fig 9. Most Requested Image**

**4. Search engines**

There are many different type search engines is available for searching web site. The figure 10 is show Google is mostly used by users.

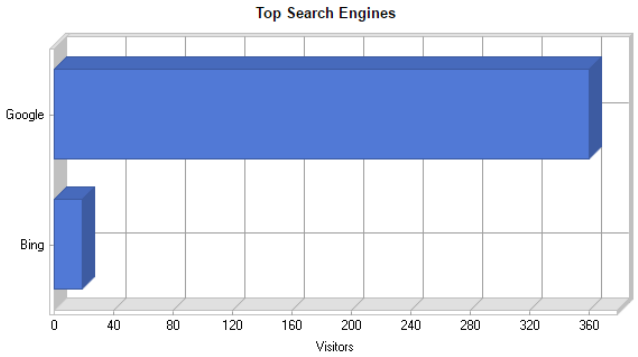


**Fig 10. Google Search Engines**

Top Search Engines

| | Search Engine | Visitors |
|---|---|---|
| 1 | Google | 360 |
| 2 | Bing | 19 |
| | Total | 379 |

**5. Browser**

There many different type browser are available for such as chrome, Android browser etc.fig.11 show Google chrome is the mostly used by users.

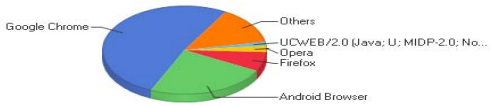


**Fig11. Most used Browsers**

This chart is show the daily SO use of visitor. Hits for Android OS are more than any other OS.

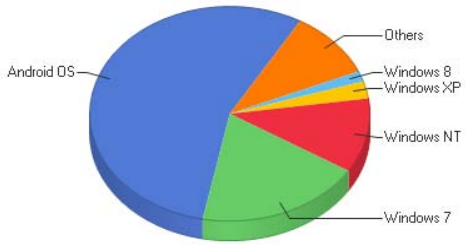


**Fig12. Most used operating system**

## V. CONCLUSION

Thus the conclusion of this paper is to finding a searching behavior of visitor's by using tool and according to this behavior we are maintain the web site. Web-application is an extremely used interface to find remote data, commercial and non-commercial services. Web usage mining is a growing zone with the growth of the web applications to find the web usage pattern. In pattern discovery will be use different data mining technique like as association rule, classifiaction , clustring , and sequential pattern technique to find important information . The results that has been extracted can be represebted in may ways such as graphs , charts , table, etc. The web mining usage pattern of an Technical Institution web data. Web related data is coteries in to three parts namely web log, access log, error log and proxy log data and collect the data in web server and implemented a web log expert. Our experimental results help to predict and identify the number of visitor for the website and improve the website usability.

## VI. REFERENCE

[1]. Dhawan, Sanjeev, and Swati Goel. "Web Usage Mining: Finding Usage Patterns from Web Logs." *American International Journal of Research in Science, Technology, Engineering & Mathematics* (2013): 203-207.

[2] Shukla, Rajesh, Sanjay Silakari, and P. K. Chande. "Web Personalization Systems and Web Usage Mining: A Review." *International Journal of Computer Applications* 72, no. 21 (2013).

[3] Nina, Shahnaz Parvin, Mahmudur Rahman, Khairul Islam Bhuiyan, and Khandakar Entenam Unayes Ahmed. "Pattern discovery of web usage mining." In *Computer Technology and Development, 2009. ICCTD'09. International Conference on*, vol. 1, pp. 499-503. IEEE, 2009.

[4] Nina, Shahnaz Parvin, Mahmudur Rahman, Khairul Islam Bhuiyan, and Khandakar Entenam Unayes Ahmed. "Pattern discovery of web usage mining." In *Computer Technology and Development, 2009. ICCTD'09. International Conference on*, vol. 1, pp. 499-503. IEEE, 2009.

[5] Suneetha, K. R., and Raghuraman Krishnamoorthi. "Identifying user behavior by analyzing web server access log file." *IJCSNS International Journal of Computer Science and Network Security* 9, no. 4 (2009): 327-332.

[6] Baoyao, Zhou. "Intelligent Web Usage Mining." *Nanyang Technological University, Division of Information Systems, School of Computer Engineering* 94 (2004).

[8] Santra, A. K., and S. Jayasudha. "Classification of web log data to identify interested users using Naïve Bayesian classification." *International Journal of Computer Science Issues* 9, no. 1 (2012): 381-387.

[9] Sharma, Anshuman. "Web usage mining using neural network." *international journal of reviews in computing* 9 (2012).

[10] Facca, Federico Michele, and Pier Luca Lanzi. "Recent developments in web usage mining research." In *International Conference on Data Warehousing and Knowledge Discovery*, pp. 140-150. Springer Berlin Heidelberg, 2003.

[11] Turban, Efraim, Ramesh Sharda, Jay E. Aronson, and David King. *Business Intelligence: um enfoque gerencial para a inteligência do negócio*. Bookman Editora, 2009.

[12] Tyagi, Navin Kumar, A. K. Solanki, and Sanjay Tyagi. "An algorithmic approach to data preprocessing in web usage mining." *International journal of information technology and knowledge management* 2, no. 2 (2010): 279-283.

[13] Upadhyay, Akshay, and Balram Purswani. "Web usage mining has pattern discovery." *International Journal of Scientific and Research Publications* 3, no. 2 (2013): 1-4.