Enrolment Comparison for students in Bundelkhand University through out the Globe using Web usage mining

Saurabh Choudhry[1], Prof (Dr.) A .K. Solanki [2]

[1] Research Fellow, Department of CSE, Bhagwant University,Ajmer. (Rajasthan). India

[2] Head , Director B.I.E.T Jhansi. (U.P.).India

Abstract

With an explosive growth of the World Wide Web, websites are playing an important role in providing an information and knowledge to the end users. Web usage patterns are an important aspect to discover hidden and meaningful information. It will be big challenge in web mining when the volume of traffic is large and the volume of web data is still in the growing phase. To face the challenge an intelligent approach of web traffic analysis has been highlighted in this paper. This paper will be used to better serve the web based application.

KEYWORDS:-Web Usage Mining, Log Data, Pattern Recognition, Web log explorer, Data pre-processing

1. Introduction

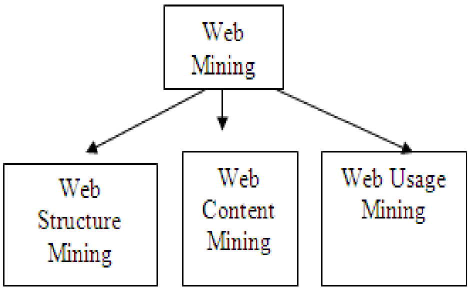
Today billions of customers visit millions of web sites daily for consumer information, financial management, education and many other services. When customer visits the websites customer leave mass of information in the log file. Stored data in log files becomes useful only when it is analysed and turn into information that can be used in future. The web mining is the use of data mining techniques to automatically discover and extract information from the user (web log) data which was left by user while working on web site. In this we extract user's visiting characteristics, then extract the user's using pattern.

Web mining is classified into three categories: Web Usage Mining (WUM), Web Content Mining (WCM), and Web Structure Mining (WSM). Among these, WUM is applied on usage data and it is being used at large scale by the organizations to study the behaviour of their web users. In WUM the user's web log is collected for inferring the useful information by analyzing it. Before the process of pattern recognition the log file data has to go through three stages i.e., pre-processing, pattern discovery and pattern analysis. This paper mainly considers web usage mining which is the process of extracting useful information from server logs i.e. user's log files . The rest of the paper is organized as follows: What is web mining is given in section 1.Section 2 describes about data collection Section 3 presents status codes . In section 4 description of parameters used in web usage mining using pattern recognition is given in detail using log files which plays an important role in extraction of useful patterns. Experimental result analysis of the problem is carried out in section 5. Finally section 6 concludes the paper.

1.1CATEGORIES OF WEB MINING

In this section we present taxonomy of web mining. . The web mining is the use of data mining techniques to automatically discover and extract information from World Wide Web documents and services in which at least one of structure or usage (web log) data is used in the mining process. Web mining can be divided into three distinct categories.

Taxonomy of Web Mining



2 Server Level Collection

2.1 Log Files

In this paper, log file data of BUNDELKHAND UNIVERSITY JHANSI web server is used to extract useful patterns. Before the application of pattern recognition, initially log file data is being pre-processed to remove any unwanted entries so that the patterns extracted are useful and relevant. Log files can be classified into three categories depending on the location of their storage.

- **Web Server Log Files:** These log files resides in web server and notes activity of the user browsing website. There are four types of web server logs i.e., transfer logs, agent logs, error logs and referrer logs.
- **Web Proxy Server Log Files:** These log files contains information about the proxy server from which user request came to the web server.

- **Client browser Log Files:** These log files resides in client's browser and to store them special software are used.

2.2 Access log files at server side are the basic information source for Web usage mining. These files record the browsing behavior of site visitors. Data can be collected from multiple users on a single site. Log files are stored in various formats such as Common log [8] or combined log formats. Following is an example line of access log in common log format.
123.456.78.9-[25/Apr/1998:03:04:41-0500]
"GET/HTTP/1.0" 200 3290

This line consist the following fields.

- Client IP address
- User id ('-'if anonymous)
- Access time
- HTTP request method
- **.** Path of the resource on the Web server
- Protocol used for the transmission
- Status code returned by the server
- Number of bytes transmitted

There are mainly three types of log file formats that are used by majority of the servers.

Common Log File Format: It is the standardized text file format that is used by most of the web servers to generate the log files. The configuration of common log file format is given below in the box.

Combined Log Format: It is same as the common log file format but with three additional fields i.e., referral field, the user agent field, and the cookie field.

Multiple Access Logs: It is the combination of common log format and combined log file format but in this format multiple directories can be created for access logs.

3. Hypertext Transfer Protocol with Status code

In this paper we are using error status of http code .Hypertext Transfer Protocol (HTTP) [9] is a standard method for transmitting information through the Internet. A Web is interconnections between hypermedia documents and these documents are delivered by hypertext transfer protocol. Transfer Control Protocol (TCP) work as a transport layer for hypertext transfer protocol to retrieve distributed hypermedia. HTTP is a very simple protocol. Initially a connection is established between client and server .Client issue a request to server .Server processes the request, returns a response and then closes the connection. A method (GET, PUT, POST, etc.) is used to get an object. HTTP request specifies a method, the object to which method is to be applied, and a string specify HTTP level (e.g. HTTP/1.0) that client can accept. Object types and methods the client or server supports may be specified in MIME, RFC-822 format. A HTTP status code is returned by server to the client as a response. Such status codes of Hypertext Transfer Protocol are listed in [10].Some of them are 100(Continue), 200(OK), 300(Multiple Choice), 400(Bad Request), 403(Forbidden), 404(Not Found), 503(Out of Resources) etc. In this study we have mainly focused on 403,404 and 503 status codes.

4. Log Files Parameters

Log files contain various parameters which are very useful in recognizing user browsing patterns.
Below is the list of some of the parameters.

- **User Name:** Identifies the user who has visited the website and this identification normally is IP address.
- **Visiting Path:** It is the path taken by the user while visiting the website.
- **Path Traversed:** It is the path taken by the user within the website.
- **Time Stamp:** It is the time spent by user on each page and is normally known as session.
- **Page Last Visited:** It is the page last visited by the user while leaving the website.
- **Success Rate:** It is measured by downloads and copying activity carried out on the website.
- **User Agent:** It is the browser that user uses to send the request to the server.
- **URL:** It is the resource that is accessed by the user and it may be of any format like HTML, CGI etc.
- **Request Type:** It is the method that is used by the user to send the request to the server and it can be either GET or POST method.

5. EXPERIMENTAL RESULT ANALYSIS

Pattern recognition is defined as the act of taking in raw data and making an action based on the "category" of the pattern [12] or it can be defined as the process for observing patterns of interest (e.g. most used file type) from entire data (e.g. web server log data). Pattern recognition can also be used to make important decisions about the patterns. Finding or recognizing patterns from web logs requires the log data to go through three stages i.e., pre-processing, pattern discovery and pattern analysis to investigate the pattern of the file types accessed by the users during the browsing of the BUNDELKHAND UNIVERSITY JHANSI website. The results can be analyzed in terms of the following browsing patterns. Details of the user access log files that are used to analyze the user browsing patterns are shown below.Details of Log File used

Log File Details

Log File Name BUNDELKHAND UNIVERSITY JHANSI-HTTP Logs

Log Duration and Data Range one months - Jul 01 to Jul 31(2013) Size of the Log File 20.7 MB gzip compressed, 205.2 MB uncompressed URL ftp://ita.ee.lbl.gov/traces/BUNDELKHAND UNIVERSITY JHANSI_access_log_July2013.gz

Since the size of considered data is very huge and it is very difficult to represent in this paper, therefore we have taken very few data from the BUNDELKHAND UNIVERSITY JHANSI database. The sample of BUNDELKHAND UNIVERSITY JHANSI web log data is presented in the fig.1.

The stages used to carry out the experimental work on log file from BUNDELKHAND UNIVERSITY JHANSI web server using WLE tool to recognize important and useful patterns are explained as follows.

Table 1 Summary for General Profile: Activity Statistics

Unique IP	272
Visitors	555
Hits	31576
Errors	2980
Pages/Files	180
Countries	6
Entry Points	34
Referrer site	1

Summary of Activity

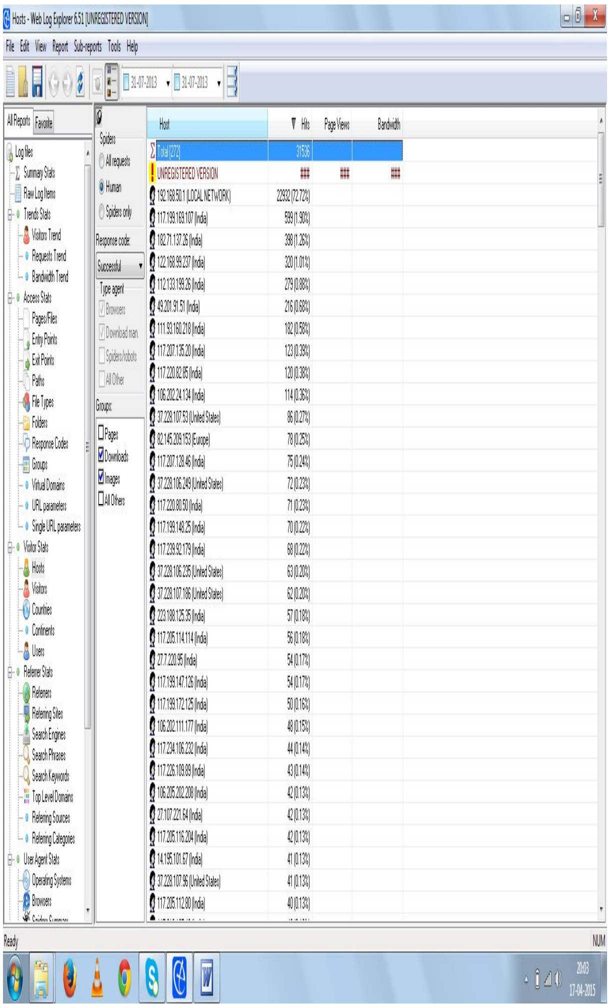
5.1 Pattern Discovery

Pattern Discovery is used to extract patterns of usage from web data [13]. This method uses data mining techniques and algorithms to find out useful information. Knowledge extracted can be represented in many ways such as graphs, charts, tables, forms etc. Techniques used in pattern discovery are given as follows:

5.1.1. Converting IP Address to domain name

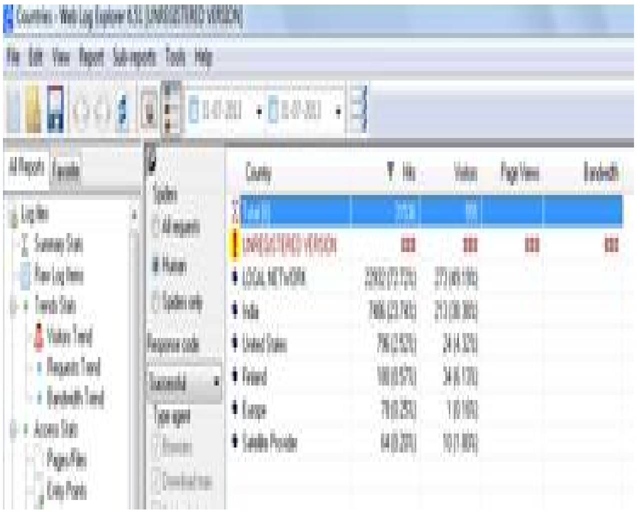
This Technique is useful in discovering important facts about the visitor such as Country of Visitor by looking domain name extension such as ".in" domain specifies that the user is from India. Finding such information about visitors helps in customizing website according to visitor's interest. In this study the log file is analyzed and results extracted are as follows.

- IP address (37.228.107.53) map to user from USA
- IP address (85.145.209.153) maps to user from EUROPE
- IP address(117.199.169.107)maps to user from INDIA


Hosts - Web Log Explorer 6.51 [UNREGISTERED VERSION]

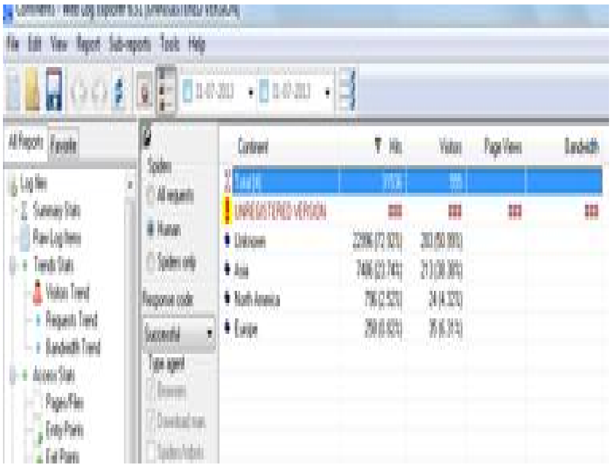
Host	Hits	Page Views	Bandwidth
Total [272]	31536		
UNREGISTERED VERSION	###	###	###
192.168.50.1 (LOCAL NETWORK)	22932 (72.72%)		
117.199.169.107 (India)	599 (1.90%)		
182.71.137.26 (India)	398 (1.26%)		
122.168.99.237 (India)	320 (1.01%)		
112.133.199.26 (India)	279 (0.88%)		
49.201.91.51 (India)	216 (0.68%)		
111.93.160.218 (India)	182 (0.58%)		
117.207.135.20 (India)	123 (0.39%)		
117.220.82.85 (India)	120 (0.38%)		
106.202.24.134 (India)	114 (0.36%)		
37.228.107.53 (United States)	86 (0.27%)		
82.145.209.153 (Europe)	78 (0.25%)		
117.207.128.46 (India)	75 (0.24%)		
37.228.106.249 (United States)	72 (0.23%)		
117.220.80.50 (India)	71 (0.23%)		
117.199.148.25 (India)	70 (0.22%)		
117.239.92.179 (India)	68 (0.22%)		
37.228.106.235 (United States)	63 (0.20%)		
37.228.107.186 (United States)	62 (0.20%)		
223.188.125.35 (India)	57 (0.18%)		
117.205.114.114 (India)	56 (0.18%)		
27.7.220.95 (India)	54 (0.17%)		
117.199.147.126 (India)	54 (0.17%)		
117.199.172.125 (India)	50 (0.16%)		
106.202.111.177 (India)	48 (0.15%)		
117.234.106.232 (India)	44 (0.14%)		
117.226.109.89 (India)	43 (0.14%)		
106.205.202.208 (India)	42 (0.13%)		
27.107.221.64 (India)	42 (0.13%)		
117.205.116.204 (India)	42 (0.13%)		
14.195.101.67 (India)	41 (0.13%)		
37.228.107.96 (United States)	41 (0.13%)		
117.205.112.80 (India)	40 (0.13%)		

HOST


Countries - Web Log Explorer 6.51 [UNREGISTERED VERSION]

Country	Hits	Visitors	Page Views	Bandwidth
Total [6]	31536	555		
UNREGISTERED VERSION	###	###	###	###
LOCAL NETWORK	22932 (72.72%)	271 (48.83%)		
India	7466 (23.74%)	213 (38.38%)		
United States	796 (2.52%)	24 (4.32%)		
Finland	180 (0.57%)	34 (6.13%)		
Europe	78 (0.25%)	1 (0.18%)		
Satellite Provider	64 (0.20%)	10 (1.80%)		

Country Visitors



Continent	Hits	Visitors	Page Views	Bandwidth
Total	[illegible]	[illegible]		
UNREGISTERED VERSION	[illegible]	[illegible]	[illegible]	[illegible]
Unknown	[illegible]	[illegible]		
Asia	[illegible]	[illegible]		
North America	[illegible]	[illegible]		
Europe	[illegible]	[illegible]		

Continent Visitors

6. CONCLUSION

In this paper, we study the web usage mining with pattern recognition techniques and carried out experimental work on web log data collected from BUNDELKHAND UNIVERSITY web server to find out useful browsing patterns Pattern recognition can be used by the web administrator to optimize web site performance It was found that students are interested to get enrol themselves in University Program. This extracted usage pattern will help Administrators managing the website resources in better way. Where I.P is converted to domain for better comprehension.

Removal of error when published information is used by student in web site will definitely attract more students to get enrolled in BU.

As already we have information students from different part of globe are interested to come to India for higher Studies.

The results or findings from this study are surely useful for web administrator in order to improve web site performance through the improvement contents, structure, presentation and delivery. So that more and more students get attracted towards India for Higher Studies.

REFERENCES

[1]. J. Srivastava et al., “Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data,” ACM SIGKDD Explorations, vol. 1, no. 2, 2000, pp. 12-23.

[2]. L.K Joshila Grace, V. Maheswari, Dhinaharan Nagamalai, “Analysis of Weblogs and Web User in Web Mining,” International Journal of Network Security & Its Applications (IJNSA), Vol. 3, No. 1, January 2011.

[3]. Rajni Pamnani, Pramila Chawan: “Web Usage Mining: A Research Area in Web Mining” Department of computer technology, VJTI University, Mumbai.

[4]. Yang Bin, Dong Xianguin, Shi Fufu, “Research of Web Usage Mining based on Negative Association Rules” International Forum on Computer Science-Technology and Applications, 2009.

[5]. M Eirinaki and M Vazirgiannis, “Web Mining for Web Personalization,” ACM Trans. Internet Technology, vol. 3, no. 1, 2003, pp. 1-27. International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.4, July 2013

[6]. Resul Das, Ibrahim TURKOGLU, “Extraction of Interesting Patterns Through Association Rule Mining for Improvement of Website Usability,” Journal of Electrical & Electronics Engineering, Vol.9, No. 2, 2009.

[7]. Kobra Etminani, Mohammad-R. Akbarzadeh-T., Noorali Raeeji Yanehsari, “Web Usage Mining:user’s navigational patterns extraction from web logs using Ant-based Clustering Method,” IFSAEUSFLAT, 2009.

[8]. Ms. Dipa Dixit, Mr. Jayant Gadge, “Automatic Recommendation for Online Users Using Web Usage Mining,” IJMIT, Vol. 2, No. 3, August 2010.

[9]. Anshuman Sharma, “Web Usage Mining Using Neural Network,” International Journal of Reviews in Computing (IJRIC), Vol. 9, 2012.

[10]. Juan Julian Merelo Guervos et al., “Weblog Recommendation Using Association Rules,” IADIS International Conference on Web Based Communities 2006.

[11]. Web Log Explorer Tool - http://www.exacttrend.com/WebLogExplorer/

[12]. Richard O. Duda, Peter E. Hart, David G. Stork 2001 Pattern classification (2nd edition), Wiley, New York, ISBN 0-471-05669-3.

[13]. Resul Das, Ibrahim TURKOGLU, Mustafa POYRAZ, “Analyzing of System Errors for Increasing a Web Server Performance by using Web Usage Mining,” Journal of Electrical & Electronics Engineering, Vol. 7, No. 2, 2007

Author Profile:

(Ph.D. Scholar)Saurabh Choudhry Pursuing Ph.D. from Bhagwant University, Ajmer Rajasthan

Prof.(Dr.) Anil Kumar Solanki did his Ph.D. in CSE from Bundelkhand University. He has published good number of papers in National and International journals.