Enhancing The Efficiency of the Website by Mining Web Server Log Files

Jigar H. Jobanputra[1] Banshi D. Soni[2]
[1,2]Assistant Professor
[1]Department of M.C.A [2]Department of BE (I.T)
[1,2]SRK Institute of Management and Computer Education

Abstract— Web mining - is the application of data mining techniques to discover patterns from the World Wide Web. Web mining can be divided into three different types – Web usage mining, Web content mining and Web structure mining. Web usage mining is the technique that mines the web log. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site. The web log can be examined by either the client side or server side. In order to gain better efficiency, we can use server web log. It contains many useful information such as name of user, IP Address, number of bytes transferred timestamp, URL etc, so it is very useful in analysis of user behavior and from that it is also useful in the effective maintenance of website. A web server log file is the pool of huge data about client's behavior. In order to increase the efficiency of a website, the website owner should study the user's interest. This can be traced using Web Log Expert or other web mining tools. Web Log Expert is the tool that creates the quick report based on server log file of any website. In this paper we have studied the web log file of a jewellery website and extracted useful patterns of the user interest.

Key words: web mining, web usage mining, web server log mining

I. INTRODUCTION

with the incredible usage of internet now the life of individual become so convenient that consumers prefer e-commerce application/website to shop. So, e-commerce website produces large amount of data. The data has very useful hidden information regarding the consumer's behavior. This information is useful for trading website to find interesting patterns about the consumer behavior which may help to attract & entertain users, implement new marketing strategy, improve the efficiency of website, to compose accessible structure of website. These all useful information of consumers can be mined using web usage mining [1]. Web usage mining is used in the following areas [2]

- Web usage patterns can be used to gather business intelligence to improve consumer attraction, Customer retention, sales, marketing and advertisement, cross sales.
- Web usage mining focuses on techniques that could predict user behavior while the user interacts with the Web.
- Performance and other service quality attributes are crucial to user satisfaction and high quality performance of a web application is expected.
- Web usage mining of patterns provides a key to understanding Web traffic behavior, which can be used to deal with policies on web caching, network transmission, load balancing, or data distribution.
- Web usage and data mining is also useful for detecting intrusion, fraud, and attempted break-ins to the system.

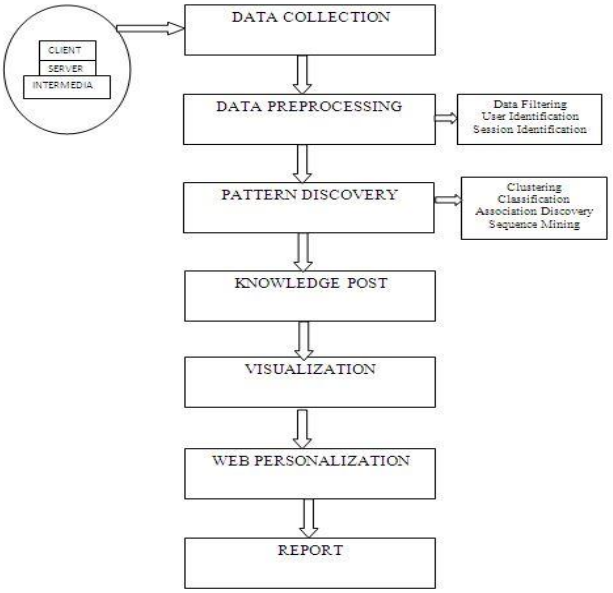


Fig. 1:

A. *Steps of web usage mining*

1) Data collection
The very first step of web usage mining is to collect a data to mine the data is a web log document which consist the log of every visit on the web.
2) Data Integration
Coordinate different web logs to a single document.
3) Data Pre-processing
Cleaning and organizing information to get ready for pattern extraction.
4) Pattern Extraction
Extraction of user patterns which will help to study user behaviour.
5) Pattern Analysis and visualization
Pattern Analysis will help to find interesting rules from set of patterns.
6) Pattern Application
Apply the patterns in the real world problems.

B. *Web Logs*

A Web log file [3] records activity information when a Web user submits a request to a Web Server. The web log file is the main source of raw data which we shall refer to as log file. The web log file can be stored in three different places: 1). Web servers 2). Web proxy servers 3). Client browsers

C. Web servers

a web server generally stores the most complete and accurate usage data.

D. Web proxy server logs

A proxy server takes the HTTP requests from users and passes them to a web server then returns to the users the results passed by the web server. Web proxy server can also stores the log files.

E. Client side logs

clients can download the special software that records the web usage HTTP cookies could also be used for this purpose. These are a piece of information generated by a web server and stored in user's computer.

Web log file of jewelers' website is considered for the purpose of analysis.

F. Description of log file

A file produced by a Web server to record activities on the Web server. It usually has the following features[5]:

- The log file is text file. Its records are identical in format.
- Each record in the log file represents a single HTTP request.
- A log file record contains essential information about a request: the client side host name or IP address, the date and time of the request, the requested file name, the HTTP response status and size, the referring URL, and the browser information.
- A browser may shoot multiple HTTP requests to Web server to display a single Web page. This is because a Web page not only needs the main HTML document; it may also need additional files, like images and JavaScript files.
- The main HTML document and additional files all require HTTP requests.
- Each Web server has its own log file format.
- If your Web site is hosted by an ISP (Internet Service Provider), they may not keep the log files for you, because log files can be very huge if the site is very busy. Instead, they only give you statistics reports generated from the logs files.
- Remote IP address or domain name: An IP address is a 32-bit host address defined by the Internet Protocol; a domain name is used to determine a unique Internet address for any host on the internet. One IP address is usually defined for one domain name.
- Authuser: Username and password if the server requires user authentication
- Entering and exiting date and time.
- Modes of request: GET, POST or HEAD method of CGI (Common Gateway Interface).
- Status: The HTTP status code returned to the client, e.g., 200 is "ok" and 404 are "not found".
- Bytes: The content-length of the document transferred.
- Remote log and agent log.
- Remote URL.
- "request:" The request line exactly as it came from the client.
- Requested URL

G. Types of Log Analyzer Tools

There are many web log analyzer tools are available and some of them are open source. Here the list of some free available web log analyzer tool [4].

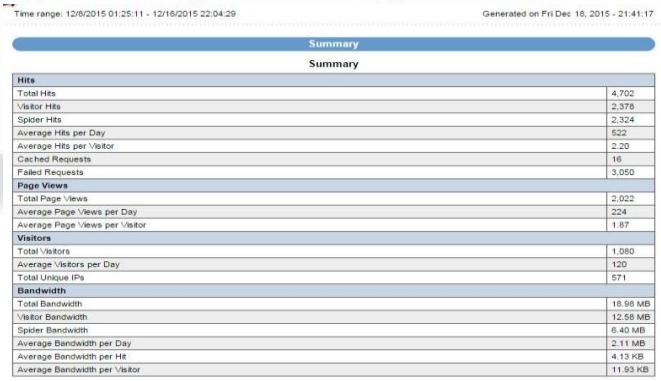
- Web log Expert
- Deep log Analyzer
- Google Analytics
- AWstats
- W3Perl
- Visitors
- Analog
- Webtrax
- Weblog Expert Lite

H. Example with snapshot analysis

In this study we have analyzed the web log file of web server with the help of weblog Expert analyzer tool. The sample log files consist the data form 8th dec, 2015 to 16th dec, 2015.The web log file have stored 1393 KB of data. The weblog Analyzer tools are required as they help extensively in analyzing the information about visitors, hits, page views, bandwidth, browsers used, operating systems etc, which can be utilized by system administrator and web designer to increase the effectiveness of the website.

I. General Statistics

In this section we get general information regarding the website.


Time range: 12/8/2015 01:25:11 - 12/16/2015 22:04:29 Generated on Fri Dec 18, 2015 - 21:41:17

Summary

Summary

Hits	
Total Hits	4,702
Visitor Hits	2,378
Spider Hits	2,324
Average Hits per Day	522
Average Hits per Visitor	2.20
Cached Requests	16
Failed Requests	3,050
Page Views	
Total Page Views	2,022
Average Page Views per Day	224
Average Page Views per Visitor	1.87
Visitors	
Total Visitors	1,080
Average Visitors per Day	120
Total Unique IPs	571
Bandwidth	
Total Bandwidth	18.98 MB
Visitor Bandwidth	12.58 MB
Spider Bandwidth	6.40 MB
Average Bandwidth per Day	2.11 MB
Average Bandwidth per Hit	4.13 KB
Average Bandwidth per Visitor	11.93 KB

Fig. 2:

J. Activity Statistics

in this section the activities performed by user will be displayed.


Time range: 12/8/2015 01:25:11 - 12/16/2015 22:04:29 Generated on Fri Dec 18, 2015 - 21:41:17

Daily

Daily Visitors

Daily Activity

Date	Hits	Page Views	Visitors	Average Visit Length	Bandwidth (KB)
Tue 12/8/2015	354	164	100	00:35	1,506
Wed 12/9/2015	508	244	114	00:33	2,136
Thu 12/10/2015	668	200	126	00:06	2,456
Fri 12/11/2015	630	360	110	04:46	2,639
Sat 12/12/2015	580	224	104	00:26	2,625
Sun 12/13/2015	436	182	122	00:02	1,619
Mon 12/14/2015	452	228	132	00:16	1,372
Tue 12/15/2015	624	246	168	00:26	3,649
Wed 12/16/2015	250	154	102	00:07	1,024
Total	**4,702**	**2,022**	**1,080**	**00:46**	**19,435**

K. Access Statistics

This section will display the statistics regarding the accessed pages by users. This section is very useful to know the user interest.

Most Popular Pages

	Page	Hits	Incomplete Requests	Visitors	Bandwidth (KB)
1	http://umajewellers.in/	730	0	674	6,972
2	http://umajewellers.in/ Photogallery.aspx	252	0	210	1,915
3	http://umajewellers.in/ index.aspx	14	0	12	135
4	http://umajewellers.in/ ContactUs.aspx	16	0	12	148
5	http://umajewellers.in/ Default.aspx	10	0	8	48
6	http://umajewellers.in/ Treditional.aspx	8	0	6	95
7	http://umajewellers.in/ Style.aspx	8	0	6	62
8	http://umajewellers.in/ Beauty.aspx	8	0	6	96
9	http://umajewellers.in/ Index.aspx	4	0	2	38
	Total	1,050	0	N/A	9,512

Fig. 4:

Most Requested Directories

	Directory	Hits	Incomplete Requests	Visitors	Bandwidth (KB)
1	http://umajewellers.in/	1,054	0	706	9,512
2	http://umajewellers.in/ images/	54	0	40	1,308
3	http://umajewellers.in/c ss/	110	0	12	575
	Total	1,218	0	N/A	11,396

Fig. 5:

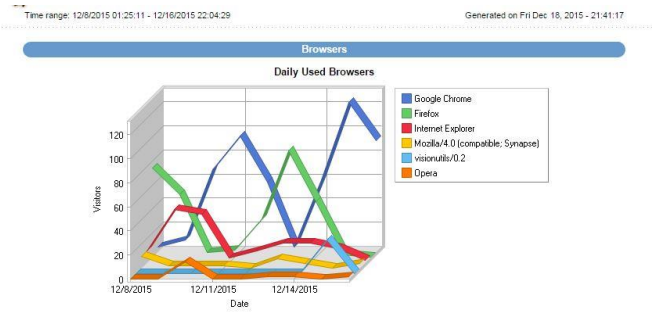
L. Visitors

This section will display the host information along with their hits and bandwidth.

Fig. 5:

M. Browsers

This section will display the information regarding the browser used to visit the web site. And also operating system on client machine.



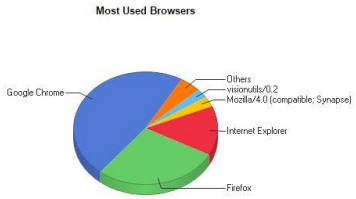


Fig. 6:

Most Used Browsers

	Browser	Hits	Visitors	% of Total Visitors
1	Google Chrome	1,228	516	47.51%
2	Firefox	536	304	27.99%
3	Internet Explorer	314	154	14.18%
4	Mozilla/4.0 (compatible; Synapse)	32	32	2.95%
5	visionutils/0.2	32	28	2.58%
6	Opera	32	20	1.84%
7	Others	18	10	0.92%
8	Mozilla/5.0	6	6	0.55%
9	Food Research	8	4	0.37%
10	Wada.vn Vietnamese Search	2	2	0.18%
11	ltx71 - (http://ltx71.com/)	2	2	0.18%
12	Seoterritory.com bot	2	2	0.18%
13	Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/534.34 (KHTML, like Gecko) Qt/4.8.1 Safari/534.34	110	2	0.18%
14	Mozilla/5.0 (compatible; spbot/4.4.2; http://OpenLinkProfiler.org/bot)	40	2	0.18%
15	Mozilla/4.0 (compatible; Win32; WinHttp.WinHttpRequest.5)	16	2	0.18%
	Total	2,378	1,086	100.00%

Fig. 7:

Most Used Operating Systems

	Operating System	Hits	Visitors	% of Total Visitors
1	Windows 8	786	356	32.96%
2	Windows 7	744	300	27.78%
3	Windows 8.1	458	256	23.70%
4	Others	158	86	7.96%
5	Windows XP	74	50	4.63%
6	Windows Vista	22	12	1.11%
7	Mac OS	12	8	0.74%
8	Linux	116	8	0.74%
9	Windows 2000	2	2	0.19%
10	Windows 10	4	2	0.19%
11	Windows Server 2003	2	0	0.00%
	Total	2,378	1,080	100.00%

Fig. 8:

II. CONCLUSION

The internet is very huge and common platform to attract the consumers for a product or service. When users visit the website they leave their footsteps in forms of web log file. And this web log will help the website owner or designer to understand the consumer interest or behavior. Web site owner can improve the efficiency of website using these reports and can attract more consumers.

REFERENCES

[1] Yanduo Zhao, "The Review of Web Mining in E-Commerce", ICCIS, 2013, 2013 Fifth International Conference on Computational and Information Sciences (ICCIS), 2013 Fifth International Conference on Computational and Information Sciences (ICCIS) 2013, pp. 571-574, doi:10.1109/ICCIS.2013.158

[2] Naresh Barsagade "Web usage Mining and Pattern Discovery "a Survey Paper" –Dec 8 2003.

[3] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang Ning Tan "Web usage mining: Discovery and Applications of usage patterns from web data" SIGKDD Explorations- vol-1, issue-2 Jan 2000 pages 12-33

[4] S.Padmaja, Dr.Ananthi Sheshasaayee "WEB SERVER LOGS TO ANALYZING USER BEHAVIOR USING LOG ANALYZER TOOL" (PAGE: 514-525)
[5] http://www.w3.org/Daemon/User/Config/
[6] http://www.weblogexpert.com/

