ISSN (Online): 1694-0784
ISSN (Print): 1694-0814

Analysis of Server Log by Web Usage Mining for Website Improvement

Navin Kumar Tyagi[1], A. K. Solanki[2] and Manoj Wadhwa[3]

[1]Marathwada Institute of Technology

Bulandshahr, UP, India

[2]Meerut Institute of Engineering and Technology

Meerut, UP, India

[3]Echelon Institute of Technology

Faridabad, Haryana, India

Abstract

Web server logs stores click stream data which can be useful for mining purposes. The data is stored as a result of user's access to a website. Web usage mining an application of data mining can be used to discover user access patterns from weblog data. The obtained results are used in different applications like, site modifications, business intelligence, system improvement and personalization. In this study, we have analyzed the log files of smart sync software web server to get information about visitors; top errors which can be utilized by system administrator and web designer to increase the effectiveness of the web site.

Keywords: *Web server log, Web usage mining, Data mining, User access patterns.*

1. Introduction

World Wide Web (WWW) is very popular and interactive. It has become an important source of information and services. The web is huge, diverse and dynamic. Extraction of interesting information from Web data has become more popular and as a result of that web mining has attracted lot of attention in recent time [1]. Web mining is an application of data mining to large web data repositories [2].It can be divided in to three categories namely web structure mining, web content mining and web usage mining. Cooley et al. [3] introduced the term web usage mining in 1997 and according to their definition; it is the automatic discovery of user access patterns from web servers. Web usage mining is an important technology for understanding user's behaviors on the web and is one of the favorite area of many researchers in the recent time. Obtained user access patterns can be used in variety of applications, for example, one can keep track of previously accessed pages of a user. These pages can be used to identify the typical behavior of the user and to make prediction about desired pages [4]. Thus personalization for a user can be achieved through web usage mining. Mass customization and personalization performed by dynamic Content Web site by making clusters of users with similar access patterns and by adding navigational links [5].

Frequent access behavior for the users can be used to identify needed links to improve the overall performance of future accesses.Prefetching and caching policies can be made on the basis of frequently accessed pages to improve latency time .In addition to modifications to the linkage structure, common access behaviors of the users can be used to improve the actual design of web pages and for making other modifications to a Web site. Moreover, usage patterns can be used for business intelligence in order to improve sales and advertisement.

Many commercial web log analyzer tools are available in the market that analyzes the web server log data to produce different kinds of statistics. In this study, web log expert program has been used to analyze server log data of a website. Program generated different types of reports on server log data that can be useful from the point view of system administrator or web designer to increase effectiveness of the site. It is important to note that preprocessing [6] is a necessary step in web usage mining before applying any technique on usage data to discover user access patterns. As far as mining of knowledge from the data is concerned, quality of data is a key issue. Nearly 80% of mining efforts often spend to improve the quality of data [7]. The data which is obtained from the logs may be incomplete, noisy and inconsistent. The attributes that for in quality data includes accuracy, completeness, consistency, timeliness, believability, interpretability and accessibility. So preprocessing of data is required to make it have the above mentioned attributes and to make it easier for mining.

2.Source of Data for Web Usage Mining

Web data can be classified as content data, structure data, user profile data and usage data. Web usage data is the collection of data that describes the usage of web resources. The usage data which is used for mining purposes can be collected at different levels i.e. Server level, Client level or Proxy level. In this study we will take the case of web server.

2.1 Server Level Collection

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

Modern Web servers like Apache supports combined log format by inserting further variable values. User agent and Referrer are the examples of such variables. When customization was not possible, referring URL’s and user agents stored in different log files namely referrer log and agent log respectively. Data of a typical Web server is shown in figure 1.

```
203.30.5.145 www.acr-news.org – [01/Jun/1999:03:09:21 -0600] “GET /Calls/OWOM.html HTTP/1.0” 200 3942 “http://www.lycos.com/cgi -bin/ pursuit? Query= advertising+ psychology & maxhits =20&cat=dir” “Mozilla/4.5 [en] (Win98; I) “
203.30.5.145 www.acr-news.org - [01/Jun/1999:03:09:23 -0600] “GET/Calls/Image/earthhani.gif. HTTP/1.0” 200 10689 http://www.acr-news.org/Calls/OWOM.html“Mozilla/4.5 [en] (Win98; I)”
203.30.5.145 www.acr-news.org - [01/Jun/1999:03:09:24 -0600] “GET /Calls/Image/line.gifHTTP/1.0” 200 190 “http://www.acr-news.org /Calls/OWOM.html” “Mozilla/4.5 [en] (Win98; I)”
203.30.5.145 www.acr-news.org - [01/Jun/1999:03:09:25 -0600] “GET /Calls/Image/red.gifHTTP/1.0” 200 104 “http://www.acr-news.org /Calls/OWOM.html” “Mozilla/4.5 [en] (Win98; I)”

203.252.234.33 www.acr-news.org - [01/Jun/1999:03:32:31 -0600] “GET / HTTP/1.0” 200 4980 ““
“Mozilla/4.06 [en] (Win95; I)”

203.252.234.33 www.acr-news.org - [01/Jun/1999:03:32:35 -0600] “GET / Image/line.gif HTTP/1.0” 200 190 “http://www.acr-news.org /” “Mozilla/4.06 [en] (Win95; I)”
203.252.234.33 www.acr-news.org - [01/Jun/1999:03:32:35 -0600] “GET / Image/line.gif HTTP/1.0” 200 190 “http://www.acr-news.org /” “Mozilla/4.06 [en] (Win95; I)”
Images/earthani.gif HTTP/1.0” 200 10689 “http://www.acr-news.org /” “Mozilla/4.06 [en] (Win95; I)”
203.252.234.33 www.acr-news.org - [01/Jun/1999:03:32:35 -0600] “GET /CP.html HTTP/1.0” 200 3218 “http://www.acr-news.org /” “Mozilla/4.06 [en] (Win95; I)”
```

Fig1. Web server log

3.Hypertext Transfer Protocol

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. Data Preprocessing

It is important to understand that mining process gives better results with quality data. In order to improve the quality of data, approximately 80% of mining efforts are required [11]. So preprocessing is necessary to build complete and robust data file. Following are the main preprocessing activities.

4.1 Data Cleaning

Irrelevant information which is useless for mining purposes [12, 13,14] can be removed from the HTTP server log files e.g. access performed by spiders, crawlers ,robots(these are automatic agents that surf the Web to collect and store the information e.g. search engine spiders)and files with extension name jpg, gif, css .

4.2 User Identification

IP address, User agents and referring URL fields of log file are used to identify user. There are some problems which can arise in user identification [4]. ISP’s which uses DHCP technology, it is difficult to identify same user through different TCP/IP connections because IP address changes dynamically (single IP address/multiple server session). It is also possible that IP address of a user changes from connection to connection (multiple IP address/single user). Different IP address can be assigned for every single request performed by the user

(Multiple IP address/single server session). Moreover, same user can access the Web by using different browsers from the same host (multiple agent/single users).

4.3 User Session Identification

Log entries of the same user are divided in to sessions or visits. A time out of 30 minutes between sequential requests from the same user is taken in order to close a session.

5. Results

In this study, we have analyzed the log files of Web server of smart sync software (www.smsync.com) with the help of weblog Expert program. The log files consists the data from 8 December 2007 to 15 December 2007. In this duration log files have stored 50 MB data and we have got 6.8 MB data after preprocessing. We have determined different types of errors that occurred in web surfing. Statistics about hits, page views, visitors and bandwidth are shown in table 1. Figure 2 shows the daily errors types. Different types of errors are shown in Table 2. It is clear from the table that 404 (Table 3) is most frequently occurred error. Some other types of client and server errors are shown in Table 4.

Table 1: Summary of statistics

Hits	
Total Hits	30,474
Visitor Hits	29,191
Spider Hits	1,283
Average Hits per Day	3,809
Average Hits per Visitor	8.16
Cached Requests	3,979
Failed Requests	233
Page Views	
Total Page Views	4,435
Average Page Views per Day	554
Average Page Views per Visitor	1.24
Visitors	
Total Visitors	3,576
Average Visitors per Day	447
Total Unique IPs	3,038
Bandwidth	
Total Bandwidth	567.48 MB
Visitor Bandwidth	548.81 MB
Spider Bandwidth	18.67 MB
Average Bandwidth per Day	70.94 MB
Average Bandwidth per Hit	19.07 KB
Average Bandwidth per Visitor	157.15 KB

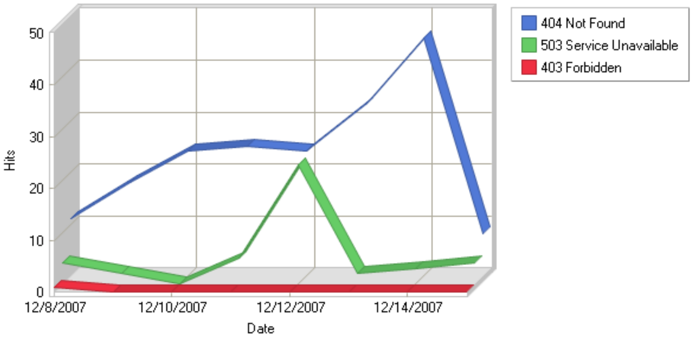

Fig2: Daily errors types

Table 2: Types of errors

Sr. No	**Error**	**Hits**
1	404 Not Found	189
2	503 Service Unavailable	43
3	403 Forbidden	1
	Total	233

Table 3: 404 Errors (Page Not Found)

Sr. No	**Request/Referrer**	**Hits**
1	/smartsyncpro/pad_file.xml No Referrer	38
2	/img/ssp_screenshot.html No Referrer	16
3	/smartsyncpro/history.html No Referrer	10
4	/smartsyncpro/registration.asp No Referrer	7
5	/t.php No Referrer	7
6	/smartsyncpro/screenshots.asp No Referrer	7
7	/)/ No Referrer	5
8	/smartsyncpro/screenshots.php No Referrer	5

Table 4: Other errors

Sr. No	**Error**	**Request/referrer**	**Hits**
1	503 Service Unavailable	/downloads/ssyncpro.exe http://www.smsync.com/downloads	21
2	503 Service Unavailable	/downloads/ssyncpro.exe http://www.listsoft.ru/programs/15371/	4
3	503 Service Unavailable	/favicon.ico No Referrer	4
4	403 Forbidden	/smartsyncpro/.html No Referrer	1

6.Related Work

In recent years, web usage mining is one of the favorite area of many researchers. Web usage mining techniques have been widely used to discover interesting and frequent user navigation patterns from web server logs. A novel approach for classifying user navigation patterns and to predict user's future request was introduced in [15]. In another approach, data from a data warehouse and web data can be used to improve marketing activities [16]. A survey about the different categories of web mining e.g. web content mining, web structure mining and web usage mining has done in [17]. A survey on mining interesting knowledge from web logs is presented in [18]. An overview of soft computing techniques (neural network, fuzzy logic, genetic algorithms) used in web usage mining applications is presented in [19, 20].

7. Conclusions

In order to make a website popular among its visitors, System administrator and web designer should try to increase its effectiveness because web pages are one of the most important advertisement tools in international market for business. The obtained results of the study can be used by system administrator or web designer and can arrange their system by determining occurred system errors, corrupted and broken links. In this study, analysis of web server log files of smart sync software has done by using web log expert program. Other web sites can be used for similar kind of studies to increase their effectiveness. With the growth of web-based applications web usage and data mining to find access patterns is a growing area of research. Data mining techniques like association rules, sequential patterns, clustering and classification can be used to discover frequent patterns.

References

[1] Cooley, R., Mobasher, B., and Srivastava, J, "Web mining: information and pattern discovery on the World Wide Web", International Conference on Tools with Artificial Intelligence, Newport Beach, IEEE, 1997, pp. 558-567.

[2] Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava," Data preparation for mining World Wide Web browsing patterns", Journal of Knowledge and Information System,1999,pp. 1-27.

[3] Robert Cooley, Bam shad Mobasher, and Jaideep Srivastava." Grouping Web page references into transactions for mining World Wide Web browsing patterns", Knowledge and Data Engineering Workshop, New port Beach, CA.IEEE, 1997, pp.2-9.

[4] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, and Pang-Ning Tan, "Web usage mining: Discovery and applications of usage patterns from Web data", SIGKDD Explorations, 2000, Vol.1.pp. 12-23.

[5] F. Masseglia, P. Poncelet, and M.Teisseire,"Using data mining techniques on Web access logs to dynamically improve Hypertext structure", 1999.

[6] Gabriek. Web usage mining and discovery of association rules from HTTP server logs.

[7] David A. Grossman, and Ophir Frieder, Information Retrieval: Algorithms and Heuristics (The Information Retrieval Series) (2nd Edition) (Paperback - Dec 20, 2004)

[8] http://www.w3.org/Daemon/User/Config/ Logging.htm#common-log file-format

[9] James Rubarth-Lay, "Optimizing Web Performance", 1996.

[10] Internet: Hypertext Transfer Protocol Overview, http://www.w3.org/Protocol/rfc2616/rfc216-sec1.html,1995

[11] Ophir Frieder, and David A. Grossman, Information Retrieval: Algorithms and Heuristics. The Information Retrieval Series, 2nd Edition, 2004.

[12] Boris Diebold, and Michael Kaufmann, "Usage based Visualization of web localities", in Australian symposium on information visualization, 2001, pp. 159-164.

[13] Corin R. Anderson," A machine Learning Approach to Web Personalization",Ph. D. Thesis, university of Washington, 2002.

[14] Pang-Ning Tan, and Vipin Kumar, "Discovery of Web robot sessions based on their navigational patterns. Data mining and knowledge discovery", 2002, 6(1), pp. 9-35.

[15] Liu, H., and Keselj, V. ," Combined mining of Web server logs and web contents for classifying user navigation patterns and predicting user's future requests", Data and Knowledge Engineering,2007,Vol 61,Issue 2, pp.304-330.

[16] Arya, S., and Silva, M.," A methodology for web usage mining and its applications to target group identification", Fuzzy sets and systems, 2004, pp.139-152.

[17] R. Kosala, and H. Blockeel," Web mining research: a Survey", SIGKDD Explorations, 2000, 2, pp.1-15.

[18] F.M. Facca, and P.L. Lanzi," Mining interesting knowledge from web logs: a survey", Elsevier Science, Data and Knowledge Engineering, 2005, 53, pp.225-241.

[19] Tug, E., Sakiroglu, and A.M. Arslan, "Automatic discovery of the sequential accesses from web log data

files via a genetic algorithm”, Knowledge based System, 2006, pp.180-186.

[20] S. Pal, V. Talvar, and P. Mitra,”Web mining in soft computing framework: relevance, state of the art and future directions”, IEEE Transactions of Neural Networks, 2002, 13 (5), pp.1163-1177.

Navin Kumar Tyagi received M.Tech in Computer Science and Engineering from Kurukshetra university, Kurukshetra (India) in 1998 and currently pursuing Ph.D in Computer Science and Engineering from Bhagwant University, Ajmer (India).Presently, he is working as Assistant Professor and Head of Computer Science and Information Technology Department in Marathwada Institute of Technology, Bulandshahr (India).He has more than 10 years of teaching experience. His areas of interest include Web usage mining, Software Engineering, Operating Systems and Data Structures.

A.K Solanki received M.E. in Computer Science and Engineering from NIT Allahabad (India) in 1996 and Ph.D in Computer Science and Engineering from Bundelkhand University Jhansi (India) in 2005. Presently, he is working as Professor and Director in Meerut Institute of Engineering and Technology, Meerut (India). He has more than 22 years of teaching experience. Professor Solanki is appointed as an executive committee member of national executive council of Indian society of technical education (ISTE) for three years 2009-2012 for Uttar Pradesh and Uttarakhand Region. He is also the member of selection and inspection committee of AICTE, UPTU and other Universities. Professor Solanki research contribution is in the field of Data warehousing and Web Mining.

Manoj Wadhwa received M.Tech in Computer Science and Engineering from Kurukshetra University Kurukshetra (India) in 1998 and Ph.D from same University in, 2009.Presently; he is working as Professor and Head of Computer Science and Engineering Department in Echelon Institute of Technology, Faridabad (India). He possesses more than ten years experience of Teaching, Research, and Industry. His areas of interest include Software Engineering, Simulation and Modeling and Operating Systems.