Identifying User Behavior by Analyzing Web Server Access Log File

K. R. Suneetha,
(Research Scholar, Anna University Trichy)
Sr. Lecturer, Dept. of CS&E
Bangalore Institute of Technology.
Vishveshwaraya Technology University.
Bengaluru – 04

Dr. R. Krishnamoorthi,
Professor and Head,
Dept. of Information Technology,
Bharathidasan Institute of Technology
Anna University,
Tiruchirappalli

Summary

Web usage mining is application of data mining techniques to discover usage patterns from web data, in order to better serve the needs of web based applications. The user access log files present very significant information about a web server. This paper is concerned with the in-depth analysis of Web Log Data of NASA website to find information about a web site, top errors, potential visitors of the site etc. which help system administrator and Web designer to improve their system by determining occurred systems errors, corrupted and broken links by using web using mining. The obtained results of the study will be used in the further development of the web site in order to increase its effectiveness.

***Keywords*:**
Web Usage Mining, Web Log Data, Common Log File Format, Data Preprocessing, Pattern Discovery.

1. Introduction

Web mining [1] is the integration of information gathered by traditional data mining methodologies and techniques with information gathered over the World Wide Web. It is used to understand customer behavior, evaluate the effectiveness of a particular Web site, and help quantify the success of a marketing campaign. It also allows looking for patterns in data through content mining, structure mining, and usage mining [2[[3]. Content mining is used to examine data collected by search engines and web spiders. Structure mining is used to examine data related to the structure of a particular Web site and Web Usage Mining is applied to many real world problems to discover interesting user navigation patterns for improvement of web site design by making additional topic or recommendations observing user or customer behavior.

Web usage mining has several applications [4] and is used in the following areas:

- It offers users the ability to analyze massive volume of click stream or click flow data, integrate the data seamlessly, with translation and demographic data from offline sources.
- Personalization for a user can be achieved by keeping track of previously accessed pages. These pages can be used to identify the typical browsing behavior of a user and subsequently to predict desired pages.
- By determining access behavior of users, needed links can be identified to improve the overall performance of future accesses.
- Web usage patterns are used to gather business intelligence to improve customer attraction, customer retention, sales, marketing, and advertisements cross sales.
- Web usage mining is used in e-Learning, e-Business, e-Commerce, e-Newspapers, e-Government and Digital Libraries. The information gathered through Web mining is evaluated by using traditional data mining parameters such as clustering and classification, association, and examination of sequential patterns.

In this paper we have concentrated on preprocessing stage of web usage mining and then analyzed the preprocessed data for performance improvement of web site design. The contents of the paper is ordered as follows, section 2 refers sources of web logs, web log structure, status codes of Hyper Text Transfer Protocol, section 3 briefs related work, section 4 explains main stages of web usage mining and process of preprocessing , experimental results shown in section 5 , conclusions and future work are mentioned in section 6.

2. Web Logs

A Web log file [5] records activity information when a Web user submits a request to a Web Server. The main source of raw data is the web access log which we shall

Manuscript received April 5, 2009
Manuscript revised April 20, 2009

refer to as log file. As log files are originally meant for debugging purposes.

A log file can be located in three different places: i) Web Servers, ii) Web proxy Servers, and iii) Client browsers. And each suffers from two major drawbacks:

i) Server-side logs: These logs generally supply the most complete and accurate usage data, but their two major drawbacks are:
- These logs contain sensitive, personal information, therefore the server owners usually keep them closed.
- The logs do not record cached pages visited. The cached pages are summoned from local storage of browsers or proxy servers, not from web servers.

ii) Proxy- side logs: A proxy server takes the HTTP requests from users and passes them to a Web server then returns to users the results passed to them by the Web server. The two disadvantages are:
- Proxy-server construction is a difficult task. Advanced network programming, such as TCP/IP, is required for this construction.
- The request interception is limited, rather than covering most requests.
- The proxy logger implementation in Web Quilt, a Web logging system performance declines if it is employed because each page request needs to be processed by the proxy simulator.

iii) Client-side logs: Participants remotely test a Web site by downloading special software that records Web usage or by modifying the source code of an existing browser. HTTP cookies could also be used for this purpose. These are pieces of information generated by a web server and stored in the users' computers, ready for future access. The drawbacks of this approach are:
- The design team must deploy the special software and have the end-users install it.
- The technique makes it hard to achieve compatibility with a range of operating systems and Web browsers.

NASA web server log file of 195 MB is considered for the purpose of analysis.

2.1 Web Log Structure

Web Server logs are plain text (ASCII) files, that is independent from the server platform. There are some distinctions between server software, but traditionally there are four types of server logs: Transfer Log, Agent Log, Error Log and Referrer Log.

The first two types of log files are standard. The Referrer and Agent Logs may or may not be "turned on" at the server or may be added to the Transfer log file to create an "Extended" Log File format.

A Web log [6] is a file to which the Web server writes information each time a user requests a resource from that particular site. Most logs use the format of the common log format. The following is a fragment from the server logs for loganalyzer.net.

66.249.65.107 - - [08/Oct/2007:04:54:20 -0400] "GET /support.html HTTP/1.1" 200 11179 "-" "Mozilla/5.0(compatible;Googlebot/2.1;+http://www.google.com/bot.html)".

This reflects the information as follows:

- 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

2.1.1 Status Codes Of Hyper Text Transfer Protocol

The Hypertext Transfer Protocol (HTTP) is an application-level protocol has been in use by the World-Wide Web since 1990. The first version of HTTP, referred to as HTTP/0.9, was a simple protocol for raw data transfer across the Internet. HTTP/1.0, as defined by RFC 1945. Status codes of Hypertext Transfer Protocol are shown in Table 1 to indicate error conditions as well as successful transmission of data.

Table 1. Status Codes of Hypertext Transfer Protocol [19]

101	Switching Protocols	404	Not Found
200	OK	405	Method Not Allowed
201	Created	406	Not Acceptable
202	Accepted	407	Proxy Authentication Required
203	Non-Authoritative In formation	408	Request Time-Out
204	No Content	409	Conflict
205	Reset Content	410	Gone
206	Partial Content	411	Length Required
300	Multiple Choices	412	Precondition Failed
301	Moved Permanently	413	Request Entity Too Large
302	Moved Temporarily	414	Request-URL Too Large
303	See Other	415	Unsupported Media Type
304	Not Modified	500	Server Error
305	Use Proxy	501	Not Implemented
400	Bad Request	502	Bad Gateway
401	Unauthorized	503	Out of Resources
402	Payment Required	504	Gateway Time-Out
403	Forbidden	505	HTTP Version not supported.

3. Related work:

This section presents the related work in this domain, now a days web usage mining is one of the emerging area where data analysation is most important to track user behavior in order to better serve users.

Many researchers [10] [11] are working on data preprocessing which involves user identification, session identification, path completion, transaction identification etc. This helps the organization to determine the value of specific customers, cross marketing strategies across products and the effectiveness of promotional campaigns, etc. The paper [12] presents an efficient incremental web traversal pattern mining algorithm. It utilizes the previous mining results and finds new patterns from the inserted or deleted part of web logs such that the mining time can be reduced. Tsuyoshi et. al [13] described a method for clarifying users' interests based on an analysis of the site-keyword graph. The method is for extracting sub graphs representing users' main interests from a site keyword graph which is generated from web log data. The author [14] analyzed the log file using statistical analysation method and provided a tool for a better understanding and interpretation of the preprocessed statistical results produced from web log data, The paper [15] explains about web server log files, problems of dealing with log data, lessons and metrics based on e-commerce, deficiencies of web server and presents statistics to overcome the issues. Detailed characterization study of user sessions is presented [16] [17] [18] and provides preliminary results on different aspects, including requests per session, number of pages requested per session, session length and intersession times.

4. Analyzing Of System Errors Using Web Usage Mining

The three main stages of web usage mining [7] are data preprocessing, pattern discovery and pattern analysis. Data preprocessing involves removal of unnecessary data. In pattern discovery data mining techniques are used in order to extract patterns of usage from Web data. Pattern discovery is the key process of the Web mining, which covers the algorithms and techniques from several research areas, such as data mining, machine learning, statistics and pattern recognition. The techniques such as statistical analysis, association rules, clustering, classification, sequential pattern and dependency modeling are used to discover rules and patterns The knowledge that can be discovered is represented in the form of rules, tables, charts, graphs, and other visual presentation forms for characterizing, comparing, predicting, or classifying data from the web access log. Pattern Analysis is the final stage of the Web usage mining. The aim of this process is to extract the interesting rules or patterns from the output of the pattern discovery process by eliminating the irrelative rules or patterns.

Our work mainly focuses on analysation of web log file; hence the access log file contents need to be preprocessed. Data preprocessing [8] is process of eliminating irrelevant fields from access log file. Fig. 1 shows the overview of

data preprocessing. It comprises Data cleaning, Identification of Users.

4.1 Preprocessing

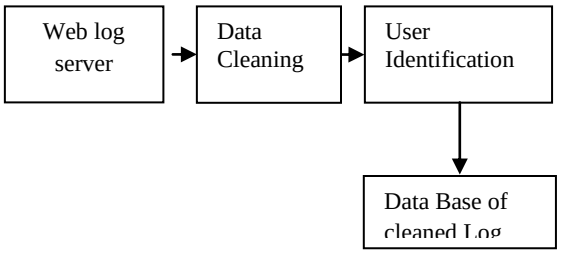

Fig.1 Overview of Data Preprocessing

Data cleaning is the first step performed in the web usage mining process some low level data integration tasks may also be performed at this stage such as combining multiple logs incorporating referrer logs, etc. The entries which are irrelevant in data analyzing and mining are removed. In data cleaning process, the entries that have status of "error" or "failure" should be removed, then some access records generated by automatic search engine agent should be identified and removed from the access log and also this process removes requests concerning non-analyzed resources such as images, multimedia files, and page style files. For example, requests for graphical page content (*.jpg & *.gif images) and requests for any other file which might be included into a web page or even navigation sessions performed by robots and web spiders. By filtering out useless data, the log file size will be reduced to use less storage space and to facilitate upcoming tasks. For example, by filtering out image requests, the size of Web server log files reduced to less than 50% of their original size after the data cleaning, the log entries must be partitioned into logical clusters using one or a series of transaction identification modules.

4.1.1 User Identification

User Identification means identifying individual users by observing their IP address. To identify unique users we propose some rules: If there is new IP address , then there is a new user, if the IP address is same but the operating system or browsing software are different, a reasonable assumption is that each different agent type for an IP address represents a different user.

NASA server log file of seven days data is processed in our work Table2 gives complete idea of percentage of reduction compare to original size.

Table2: Results of Preprocessed data

Server Log File	NASA Jul-95
Duration	1 - 7days
Original Size	69.84MB
Reduced Size After Preprocessing	19.23MB
Percentage in Reduction	72.47
Total No. of Unique Users	8660

5. Experimental Results

In this paper we have analyzed NASA server log file of size 195MB, various analysis has been carried out to identify the user behavior. The errors which arise in Web surfing were determined. The general profiles of users are shown in Table 3 and Table 4.

By observing the Table 3 the system administrators can able to guess the most active day , least active day, number of hits on most active day as well as least active day, and also can able to predict the most preferable date when to shut down the server. The table 4 shows status code of commonly occurrence of errors, failures, while transmission of information. The graphs drawn in Figure 2 shows overall visitors logged in, Figure3 gives an idea about total Hits by excluding failures and Figure 3 indicates range of unique users.

Table 3: User Profile.

Day	No. of Entries	No of IP Address	No of Unique Users	No.of Hits	Failures
1	64567	7597	576	20893	2931
2	60264	6630	613	21995	2463
3	89565	19193	1766	17390	1738
4	65536	9340	868	8545	795
5	6535	17638	1039	12564	1421
6	68342	24706	2033	22398	2304
7	87233	33657	1765	2,324	2897

Table 4: Top Errors found

302	Moved Temporarily	414	Requested URL too Large
303	See Other	415	Unsupported Media Type
304	Not Modified	500	Server Error
305	Use Proxy	501	Not Implemented
400	Bad Request	502	Bad Gateway

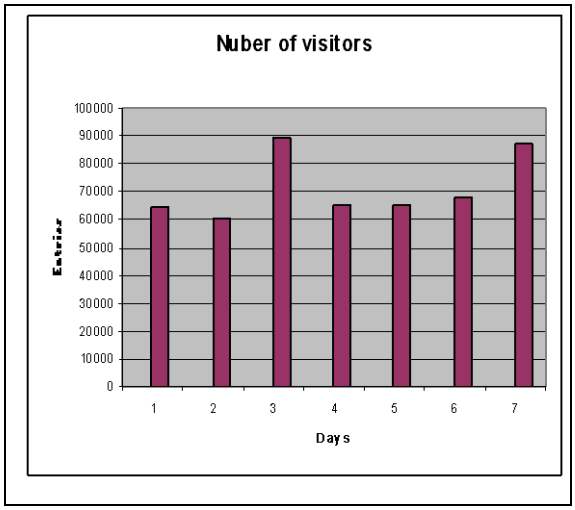

Figure 2: overall visitors for the time period

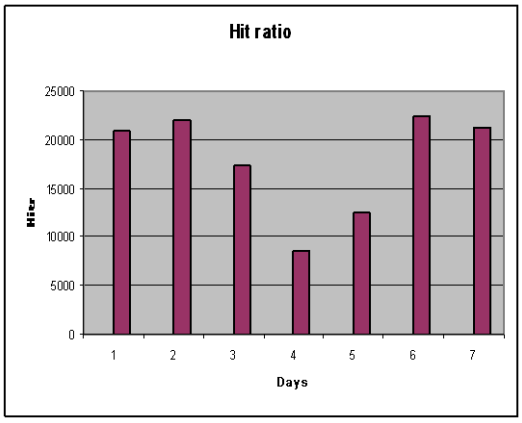

Figure3: Total number of Hit ratio

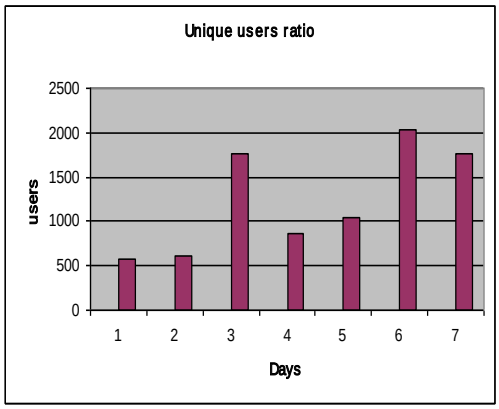

Figure 4: Total number of unique visitors

6. Conclusions and Future work

The web pages are one of the most important advertisement tools in international area for foundation, institutions, etc. Therefore, the suitability to W3C standards, content and design of web pages are very important for system administrator and Web designer. These features have deep impact on the number of visitors. Especially, the number of visitors is acceptable as the measure of the affectivity and quality for a commercial foundation or a university. So web analyzers have to analysis their server log files to determine systems error In this study, the user access log files of NASA Web server were analyzed to help system administrator and Web designer to arrange their system by determining occurred systems errors, corrupted and broken links. Similar studies can be done for any others web sites to increase their performances. Web usage patterns and data mining can be the basis for a great deal of future research. More research needs to be done in E-Commerce, Bioinformatics, Computer Security, Web Intelligence, Intelligent Learning, Database Systems, Finance, Marketing, Healthcare and Telecommunications by using Web usage mining.

6.1 Future work

Instead of tracking the behavior of overall users (interested or not interested) in order to redesign the web site to support users. The data mining techniques like Association , Clustering , and Classification can be applied only on to the group of interested regular users to find frequently accessed patterns which results in less time consumption and less memory utilization with high accuracy and performance.

References

[1] R. Kosala, H. Blockeel, Web mining research: a survey, SIGKDD: SIGKDD explorations: newsletter of the special interest group (SIG) on knowledge discovery & data mining, ACM 2 (1), 1–15, 2000.
[2] Lizhen Liu, Junjie Chen, Hantao Song, "The Research of Web Mining", Proceedings of the 4th World Congress on Intelligent Control and Automation, June 10-14, Shanghai/China, 2002.
[3] S. Pal, V. Talwar, P. Mitra, Web Mining in soft computing framework: relevance, state of the art and future directions, IEEE Transactions on Neural Networks 13 (5), 1163–1177, 2002.
[4] Naresh Barsagade "Web usage Mining and Pattern Discovery "a Survey Paper" –Dec 8 2003.
[5] 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.
[6] W.W.W.Consortium the Common Log File format http://www.w3.org/Daemon/User/Config/
Logging.html#common-logfile-format, 1995.
[7] Bamshad Mobasher, Robert Colley, and Jaideep Srivastava "Automatic Personalization Based on Web Usage Mining" August 2000/Vol.48,No.8 Communications of the ACM.
[8] R.M.Suresh, R.Padmajavalli. "An Overview of Data Preprocessing in Data and Web Usage Mining". 2006IEEE
[9] Bamshad Mobasher et.al "Effective Personalization based on Association rule Discovery from Web usage data" WIDM01 3rd ACM workshop on Web Information and data management, November 9 2001, Atlanta 2001.
[10] Zhang Huiying, Laing Wei "An Intelligent Algorithm of Data Pre-processing in Web Usage Mining " Proceedings of the 5th world Congress on Intelligent Control and Automation, June15-19, 2004 Hangzhou, P.R.China.
[11] Doru Tanasa et.al Advanced data preprocessing for inter sites Web Usage mining IEEEE computer society 2004.
[12] Show-Jane Yen, Yue-Shi Lee and Min-Chi Hsieh. "An Efficient Incremental Algorithm for Mining Web Traversal Patterns" Proceedings of the 2005 IEEE International Conference on e-Business Engineering (ICEBE'05).
[13] Tsuyoshi Murata and Kota Saito"Extracting Users Interests from Web Log Data" Proceedings of the 2006 IEEE/WIC/ACM International Conference of Web Intelligence (WI 2006 Main Conference Proceedings) (WI'06) 2006 IEEE.
[14] A-Nikos Koutsoupias "Exploring Web Access Logs with Correspondence Analysis" 2nd Hellenic Conf. on AI, SETN-2002, 11-12 April 2002, Thessaloniki, Greece, Proceedings, Companion Volume, pp. 229-23
[15] Kohavi "Mining E-Commerce Data: The Good, the Bad, and the Ugly" *KDD 2001*, Aug 26-29, San Francisco, CA. Copyright 2001 AC.
[16] Bettina Berendt, Bamshad Mobasher, Myra Spiliopoulou, Jim Wiltshire "Measuring the Accuracy of Sessionizers for Web Usage Analysis" KDD'99 Workshop on Web Usage Analysis and User Pro_ling WEBKDD'99, San Diego, CA, Aug. 1999. ACM. Springer, LNCS series.
[17] Martin Arlitt "Characterizing Web User Sessions" Internet and Mobile Systems Laboratory HP Laboratories Palo Alto HPL-2000-43 May, 2000.
[18] Maristella Agosti Giorgio Maria Di Nunzio "Web Log Mining: A Study of User Sessions "UNIVERSITY OF PADUA Department of Information Engineering.10th DELOS Thematic Workshop on Personalized Access, Pro¯le Management, and Context Awareness in Digital Libraries Corfu, Greece, 29{30 June 2007.
[19] Internet: Hypertext Transfer Protocol Overview, http://www.w3.org/Protocols/, http://www.w3.org/Protocols/rfc2616/rfc2616-sec1.html, 1995.
[20] Kurt Thearling "An Introduction to Data Mining", Computer Society of India Communications, Oct 2006, pages 21-25.
[21] Mooney, R., and Bunescu, R."Mining Knowledge from Text Using Information Extraction." To appear in a special issue of SIGKDD Explorations on Text Mining and Natural Language Processing's, 2005.
[22] Jiawei Han, Micheline Kamber, Simm Fraser University "Data Mining Concepts & Techniques" Academic Press, 2001.

Dr. R Krishnamoorthi is currently working as Professor and Head of Information Technology Department, Bharathidasan Institute of Technology, Anna University, Tiruchirapplli. He has received his Ph.D. in Image Processing from Indian Institute of Technology, Karagpur in the year 1995. He has authored several books in Computer Science and has published many research papers in reputed journals, International and National conferences. His research interest includes Software Engineering, Image Compression, Image Encryption and Authentication, Pattern Recognition and Knowledge Discovery & Management.

Suneetha K. R is a research student at Anna University, Tiruchirappalli, Tamil Nadu. She has received her M.Tech. Degree in Computer Science & Engg. from B.M.S.C.E under Visvesvaraya Technological University, Karnataka and B.E degree in Computer Science & Engg. from P.E.S.C.E under Mysore University. She is currently working as a Sr. Gr. Lecturer in Computer Science & Engineering Department, Bangalore Institute of Technology, Bangalore. Her research interest includes Knowledge Discovery from Web Usage Data, Classification, and Intelligent Agents.