ISTANBUL UNIVERSITY –
JOURNAL OF ELECTRICAL & ELECTRONICS ENGINEERING

YEAR : 2007 (379 - 386)
VOLUME : 7
NUMBER : 2

ANALYZING OF SYSTEM ERRORS FOR INCREASING A WEB SERVER PERFORMANCE BY USING WEB USAGE MINING

Resul DAS [1] **Ibrahim TURKOGLU [2]** **Mustafa POYRAZ [3]**

[1] Firat University, Department of Informatics, 23119, Elazig/TURKEY
[2] Firat University, Department of Electronics and Computer Science, 23119, Elazig/TURKEY
[3] Firat University, Department of Electrical-Electronics Engineering, 23119, Elazig/TURKEY

[1] E-mail: rdas@firat.edu.tr [2] E-mail: iturkoglu@firat.edu.tr [3] E-mail: mpoyraz@firat.edu.tr

ABSTRACT

Web usage mining is to analysis Web log files to discover user accessing patterns of Web pages. The user access log files present very significant information about a web server. This paper is deal with finding information about a web site, top errors, link errors between the pages, etc. from the web server access log files. The aim of this study is to analysis the web server user access logs of Firat University to help system administrator and Web designer to improve their system by determining occured systems errors, corrupted and broken links by using web using mining. We found useful information about activity statistics like top errors, client errors, server errors within the visited pages etc. in a web server. 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 Logs, Web Server, Knowledge Discovery.*

1. INTRODUCTION

The World Wide Web (WWW) is increasingly growing with the information transaction volume from Web servers and the number of requests from Web users in Internet. Analyzing of web server access logs is one of the application areas of web mining. With the rapid growth of the WWW, it becomes more important to find the useful information from these huge amounts of data. The Web also contains a rich and dynamic collection of hyperlink information and Web page access and usage information [1].

In recent years, web usage mining techniques have been widely used for discovering interesting and frequent user navigation patterns from Web server logs. Most research activities in web mining have centered on web usage mining [2-7]. A project aiming an automatic classification of web user navigation patterns and propose a novel approach to classifying user navigation patterns and predicting users' future requests was introduced in [3]. Another publication provides such a methodology that is based on suggestions from literature and own experience from various web mining projects. Its application in a Chilean Bank shows how a combined use of data from a data warehouse and web data can contribute to improve marketing activities [5].

There are a number of papers which provide an overview of what has happened in the area of Web Mining since 1996. Ref. [8] defines Web

Received Date : 12.07.2007
Accepted Date: 05.12.2007

Mining, providing the categorization in Web Content Mining, Web Structure Mining, and Web Usage Mining. Ref. [6, 9-14] presents a survey of the research in the area of Web Usage Mining. Recently, [7, 11, and 16] have presented an overview of the Soft Computing techniques (e.g., neural networks, fuzzy logic, genetic algorithms, and rough sets) used in Web usage mining applications.

In this study, the web server user access logs of Firat University were analyzed to help system administrator and Web designer to improve their system by determining occured systems errors, corrupted and broken links.

The paper is organized as follows: In Section 2, we describe background of this study that is application areas of web usage mining and status codes of Hypertext Transfer Protocol. In Section 3, the procedures of realized study are presented in details. The overall results of the application for Firat University are evaluated, too. In Section 4, shows results of our application by occurring systems errors, corrupted and broken links. Finally, in Section 5 concludes this paper and introduces future research.

2. BACKGROUND

2.1.Application Areas of Web Usage Mining

In this section, we present information about the application areas of Web Usage Mining. The purpose of Web Usage Mining (WUM) is to reveal the knowledge hidden in the log files of a web server [15]. Web Usage Mining extracts user access patterns by applying data mining techniques to web server logs. Web usage mining use web logs in a web server records traces of every hit into log files which can be controlled by the web master. In other words, WUM works on the secondary web data such as Web server access logs, proxy server logs, browser logs, user profiles registration data, user sessions or transactions, cookies, user queries, bookmark data, mouse clicks and scrolls and any other generated by the interaction between users and the web [7, 19]. The main application areas of WUM are shown in Fig. 1. These are *Personalization, system improvements, Site Modification, Business Intelligence* and *Usage Characterization* [5]. A detailed overview about these areas can be found in [4].

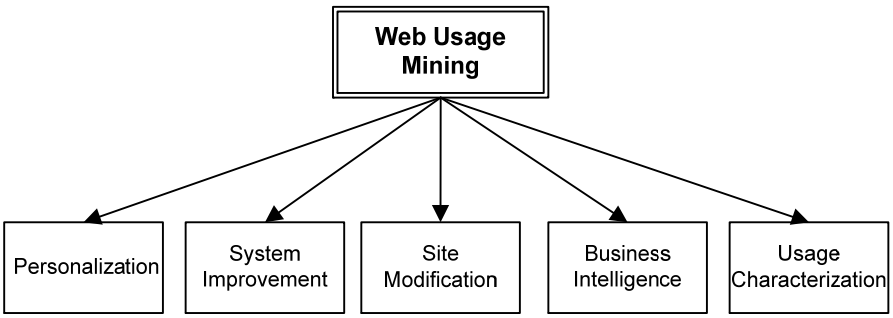

Figure 1. Main Application Areas for Web Usage Mining

2.2. Status Codes of Hypertext Transfer Protocol

The Hypertext Transfer Protocol (HTTP) is an application-level protocol for distributed, collaborative, hypermedia information systems. HTTP 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, improved the protocol by allowing messages to be in the format of MIME-like messages, containing Meta information about the data transferred and modifiers on the request/response semantics. However, HTTP/1.0 does not sufficiently take into consideration the effects of hierarchical proxies, caching, the need for persistent connections, or virtual hosts. In addition, the proliferation of incompletely-implemented applications calling them "HTTP/1.0" has necessitated a protocol version change in order for two communicating applications to determine each other's true capabilities [23]. Status codes of Hypertext Transfer Protocol are shown in Table 1.

3. WEB USAGE MINING

Recently, web usage mining has been combined in order to provide a better understanding of the requirements of a visitor when entering a web site. Analyzing the log files in relation to the content of each page offers much more information than pure log file analysis. Based on techniques from information retrieval a similarity measure between visitors of a web site has been suggested combining content with usage of the respective site [24].

3.1. Structure of Web Logs

We have the following four types of web data i.e. data generated by visits to a web site: log files, cookies and query data. Within the first one of these types there are amongst others there are different kinds of log files of particular importance: Access log, error log, referrer log, agent log.

In this work, we analyzed to user access log files by using WUM. *Access Logs* could be stored in Common Log file format (CLF) [20] or in Extended Log file format (ELF) [21]. We used CLF type of access files.

Table 1. Status Codes of Hypertext Transfer Protocol [23]

100	Continue	403	Forbidden
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 Information	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
		505	HTTP Version not supported

Table 2. A single line in a common log file format

10.1.1.18 - - [31/May/2007:08:01:38 +0300] "GET /perweb/default.asp HTTP/1.1" 304 1642 "http://www.firat.edu.tr" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1)"

Table 3. A single line from server log and its description [20]

10.1.1.18	Remote hostname (or IP if DNS hostname is not available)
-	The remote log name of the user (RFC931)
-	The username as which the user has authenticated himself.
[31/May/2007:08:01:38 +0300]	Date and time of the request.
"GET /perweb/default.asp HTTP/1.1"	The HTTP request line exactly as it came from the client.
304	The HTTP status code returned to the client.
1640	The content-length of the document transferred.
http://www.firat.edu.tr	Referrer URL
"Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1)"	User agent

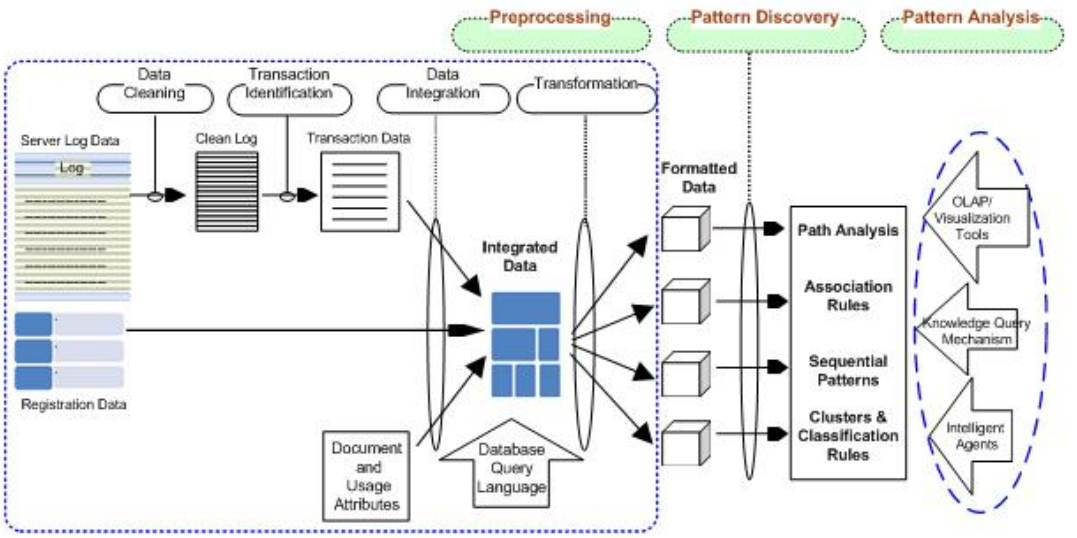

Figure 2. A General Architecture for Web Usage Mining

3.2.Preprocessing

The preprocessing of Web usage mining is usually complex. Purpose of preprocessing is to offer structural, reliable and integrated data source to pattern discovery. It consists of four steps.

Data cleaning: Data cleaning is the first step performed in the preprocessing of Web usage mining. The entries which are irrelevant in data analyzing and mining are removed. In data cleaning process, firstly, entries that have status of "error" or "failure" should be removed. Secondly, some access records generated by automatic search engine agent should be identified and removed from the access log.

*Transaction Identification***:** After the data cleaning, the log entries must be partitioned into logical clusters using one or a series of transaction identification modules. This module can be called as either a *merge* or a *divide* module the aim of of transaction identification is creating meaningful clusters of references for each user. Both types of modules take a transaction list and possibly some parameters as input, and output a transaction list that has been operated on by the function in the module in the same format as the input [12].

Data Integration: The integration of content, structure, and userdata in other phases of the Web usage mining may also be essential in providing the ability to further analyze and reason about the discovered patterns.

Tranformation: Once the domain-dependent data transformation phase is completed, the resulting transaction data must be formatted to conform to the data model of the appropriate data mining task. For example, the format of the data for the association rule discovery task may be different than the format necessary for mining sequential patterns [12].

3.3. Pattern Discovery

In this step, 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 [4]. Zaiane et al. [25] proposed the use of On-Line Analytical Processing (OLAP) technology in web usage mining. OLAP and the data cube structure offer a highly interactive and powerful data retrieval and analysis environment. 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. Visualization can also be used in web usage

mining, and it presents the data in the way that can be understood by users more easily.

3.4. Pattern Analysis

The final stage of the Web usage mining is Pattern analysis, as described in Fig. 2. 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. Many of Web usage mining tools, e.g. WebMiner, Nihuo Web Log Analyzer [17], MiDAS, have incorporated a SQL-like Web mining language, which firstly provides some objective criteria, supporting and confidence for example. In addition, OLAP to data cube makes for understanding data from various aspects. Visualization assists an analyst to better apprehend navigation pattern and to predicate trends of data [14].

4. AN EXAMPLE OF WEB USAGE MINING APPLICATION: www.firat.edu.tr

In this study, the user access log files which are stored in Web server of the University of Firat were analyzed by using Nihuo Web Log Analyzer program. The errors which arise in Web surfing were determined. The user access log files contain the information from 30 March 2007 to 20 June 2007. In this 83 days period, 6.86 GB data were stored which were belong to different 287 IP users and their 5229 times access to Web server. The general profiles of these users are shown in Table 4. The determined errors types are tabulated in Table 5. The most encountered error in Web server is 404 not found. Target URL addresses which are to cause this error are determined and tabulated in Table 6. Other client and server errors are tabulated in Table 7 and Table 8, respectively.

Table 4. Summary for General Profile: Activity Statistics

Summary of Activity	
Average Number of Visits per Day on Weekdays	16
Average Number of Hits per Day on Weekdays	701
Average Number of Visits per Weekend	10
Average Number of Hits per Weekend	399
Most Active Day of the Week	Wednesday
Least Active Day of the Week	Sunday
Most Active Date	Monday, 18 June, 2007
Least Active Date	Saturday, 19 May, 2007
Hits	
Total Hits	50962
Average Hits per Day	614
Average Hits per Visit	42.05
Cached Requests	21
Failed Requests	6
Page Views	
Total Page Views	5
Average Page Views per Day	63
Average Page Views per Visit	4.32
Visits	
Total Visits	1
Average Visits per Day	14
Total Unique IPs	287
Total Visitor Stay Length	141045:00:21
Average Visitor Stay Length	6:58
Bandwidth	
Total Bandwidth	158.667,78 MB
Average Bandwidth per Day	1.911,66 MB
Average Bandwidth per Hit	3,19 KB
Average Bandwidth per Visit	134,03 KB

Table 5. Top Errors

No	Top Errors	Hits	% of Total
1	404 Not Found	6532	97,64 %
2	501 Not Implement	95	1,42 %
3	500 Internal Server Error	39	0,59 %
4	403 Forbidden	18	0,27 %
5	406 Not Acceptable	3	0,05 %
6	405 Method Not Allowed	1	0,03 %

Table 6. 404 (Page Not Found) Errors

No	Target URL/Referrer	Hits	% of Total
1	/images/firufak_over.jpg	2877	43,70 %
2	/inc/strbkgde.gif	1665	25,29 %
3	/strbkgde.gif	765	11,62 %
4	/favicon.ico	135	2,06 %
5	/ogrotomasyon/	65	0,99 %

Table 7. Other Client Errors

No	Errors	Target URL/Referrer	Hits
1	403 Forbidden	/fiha/fihav3/haberresim/	2.369
2	403 Forbidden	/ilahiyat/Tr/AkademikKadro/	2.127
3	403 Forbidden	/ilahiyat/Tr/	1.474
4	403 Forbidden	/matematik/index.htm	1.236
5	403 Forbidden	/ilahiyat/Tr/Yonetim/	1.123

Table 8. Server Errors

No	Errors	Target URL/Referrer	Hits
1	501 Not Implemented	/kimmuh	43
2	500 Internal Server Error	/Default.asp	24
3	501 Not Implemented	/duyuru	10
4	501 Not Implemented	/fenbilimleri	9
5	501 Not Implemented	/kararlar	5

5. CONCLUSION

The web pages are one of the most important advertisement tool in international area for foundation, instutiations etc. Therefore, the suitability to W3C standards [23], content and desing 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 effectivity and quality for a commercial foundation or a university. So web analyzers have to analysis their server log files to determine systems error to increase their Web pages performance.

In this study, the user access log files which are stored in Web server of Firat University were analyzed by using Nihuo Web Log Analyzer program to help system administrator and Web designer to arrange their system by determining occurred systems errors, corrupted and broken links. These results are shown in Table 5, Table 6, Table 7 and Table 8. Similar studies can be done for any others web sites to increase their performances. Web usage and data mining to find patterns is a growing area with the growth of Web-based applications. Application of web usage data can be used to better understand web usage, and apply this specific knowledge to better serve users. 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.

ACKNOWLEDGEMENT

We thank, Firat University Computer Center, Elazig, Turkey for providing the Web server user access log files to us.

REFERENCES

[1] Gündüz, Ş., "Recommendation models for Web users: User interest model and click-stream tree", PhD. Thesis, Institute of Science and Technology, Istanbul Technical University, TURKEY, 2003.

[2] Oosthuizen, C., Wesson, J., Cilliers, C., "Visual Web mining of Organizational Web Sites", Proceedings of the Information Visualization (IV'06), IEEE Computer Society, 2006.

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

[4] Srivasta, J., Cooley, R., Deshpande, M., and Tan,P., "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data" SIGKDD Explorations. 1(2), 12-23, 2000.

[5] Araya, S., Silva, M., Weber, R., "A Methodology for web usage mining and its applications to target group identification", Fuzzy sets and systems 148, 139–152, 2004.

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

[7] Tuğ, E., Şakiroğlu, A.M., Arslan, A., "Automatic discovery of the sequential accesses from web log data files via a genetic algorithm", Knowledge-Based Systems 19, 180-186, 2006.

[8] 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.

[9] Cooley, R., Mobasher, B., Srivastava, J., "Web Mining: Information and Pattern Discovery on the World Wide Web", Tools with Artificial Intelligence, Ninth IEEE International Conference on 3-8 November 1997, Page(s): 558 – 567, USA, 1997.

[10] 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.

[11] 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.

[12] Cooley, R., Mobasher, B., Srivastava, J., "Data Preparation for mining World Wide Web Browsing Patterns", Knowledge and Information Systems 1, 1-27, 1999.

[13] Junjie Chen, Wei Liu, "Research for Web Usage Mining Model", International Conference on Computational Intelligence for Modeling Control and Automation – Intelligent Agents, Web Technologies and Internet Commerce, (CIMCA-IAWTIC'06), 2006.

[14] Feng Zhang, Hui-You Chang, "Research and Development in Web Usage Mining System-Key Issues and Proposed Solutions: A Survey", Proceedings of the First International Conference on Machine Learning and Cybernetics, 986–990, Beijing, 4-5 November 2002.

[15] Eirinaki, M., Vaziargiannis, M., "Web mining for web personalization", ACM Transactions on Internet Technology (TOIT) 3(1):pp 1-27, 2003.

[16] Khasawneh, N., Chan, C.C., "Web Usage Mining Using Rough Sets", IEEE Annual Meeting of the North American Fuzzy Information Processing Society – (NAFIPS'05), 2005.

[17] Internet: Nihuo Web Log Analyzer (NWLA), http://www.nihuo.com/, 2007.

[18] Cooley, R., "Web Usage Mining: Discovery and Application of Interesting Patterns from Web Data", PhD thesis, University of Minnesota, 2000.

[19] Wang Bin, Liu Zhijing, "Web Mining Research", Proceedings of the Fifth International Conference on Computational Intelligence and Multimedia Applications (ICCIMA'03), IEEE Computer Society, 2003.

[20] Internet: WWW Consortium, http://www.w3.org/Daemon/User/Config/Logging.html#common-logfile-format, 1995.

[21] Internet: W3C Extended Log File Format, http://w3c.org/TR/WD-logfile.html, 1996.

[22] Xiaozhe Wang, Ajith Abraham, Kate A. Smith, "Intelligent web traffic mining and analysis", Elsevier Science, Journal of Network and Computer Applications 28, 147-165, 2005.

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

[24] J.D. Velezquez, H.Yasuda, T.Aoki, R.Weber, "A New similarity measure to understand visitor behavior in the web site", IEICE Trans, Inform. Systems E87-D (2), 389-396, 2004.

[25] O. R. Zaiane, M. Xin, and J. Han, "Discovering Web access patterns and trends by applying OLAP and data mining technology on Web logs", in Proc. Advances in Digital Libraries Conference (ADL'98), Santa Babara, CA, April, 1998.

Resul DAS was born in Elazig, Turkey, 1975. He received his BS degree at Firat University from Department of Computer Science, 1999. He had his M.S.E.E degree from Computer Systems, Institute of Physical Sciences, Firat University 2002 and currently he is a Ph.D student in Firat University. His research interests are Knowledge Discovery, Web Mining, Computer Networks and Network Technologies, Distance Education Technologies. Now, he is working as an instructor in Department of Informatics at Firat University.

Ibrahim TURKOGLU was born in Elazig, Turkey, 1973. He received the B.S., M.S. and Ph.D. degrees in Electrical-Electronics Engineering from Firat University, Turkey in 1994, 1996 and 2002 respectively. He is working as an assistant professor in Electronics and Computer Science at Firat University. His research interests include artificial intelligent, pattern recognition, intelligent modeling, radar systems and biomedical signal processing.

Mustafa POYRAZ was born in Malatya, Turkey, 1953. He received the B.S. and M.S. degrees in Electrical Engineering from K.T.U. of TURKEY in 1973 and 1974. He received the Ph.D. degree in Electrical-Electronics Engineering from Firat University in 1981. Since December 1977, he has been Firat University. He has published more than 35 technical articles.