Volume 4, Issue 10, October 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Errors in Internet Log files for Website Improvement and Interaction

[1]Saurabh Choudhry*, [2]Prof A. K Solanki
Research Scholar, Department of Computer Science and Engineering, Bhagwant University Ajmer, Rajasthan, India
Head Department of Computer Science Engineering B.I.E.T Jhansi U.P., India

Abstract: ***Web mining is active research area at present. Web mining involves transformation and interpretation of web data in order to visualize and discovery user's access pattern based on user's access pattern or using pattern by data mining, artificial intelligence and knowledge discovery process, based upon the type of knowledge . 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. Web Structure Mining, Web Content Mining, Web Usage Mining. This Paper will provide user's access pattern based on web usage mining . Web usage mining provides support for efficient website design , improve customer satisfaction , improve the performance of web servers and web applications so that we can serve people better than before , in the end of the paper we are trying to elaborate application of web usage mining.***
In this study, we have analyzed the log files of BundelKhand University web server to get information about visitors, top errors which can be utilized by system administrator and web developer to enhance the productivity of the web site.

Keywords: Web Mining ; Web Usage Mining ; Data Mining ; Web Log ; Web Log Analyzer

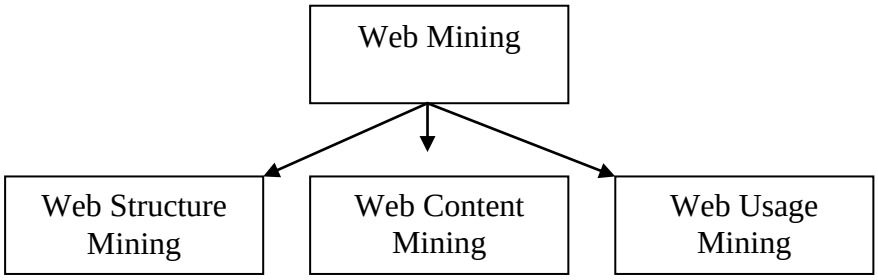
I. INTRODUCTION

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.

CATEGORIES 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



Web Content Mining

Web Content Mining is the process of picking up useful information from the contents of web documents. Content data is the collection of facts a web page is designed to contain. It may contain text, images, audio, video or structured records such as lists and tables. Research activities in this field also involve using techniques from other disciplines such as Information Retrieval (IR) and natural language processing (NLP).

Web Structure Mining

Web structure mining is a process of picking up information from linkages of web pages. It operates on the web's hyperlink structure. Web structure mining is also a process of using graph theory to analyse the node and connection

structure of a web site. The structure of a typical web graph consists of web pages as nodes and hyperlinks as edges connecting between two related pages. In addition, the content within a web page can also be organized in a tree-structured format, based on the various Hyper Text Markup Language(HTML) and eXtensible Markup Language(XML) tags within the page.

Web Usage Mining

Web usage mining is also known as Web log mining. Web usage mining is the process of picking up information from user, how to use web sites. It is an application of data mining techniques to discover interesting usage patterns from web data, in order to understand and better serve the needs of web based applications. Usage data captures the identity or origin of web users along with their browsing behaviour in a web site. Some of the typical usage data collected in a web site includes IP addresses, page references and access time of the users. The web usage data contains the data from 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 data as the results of interactions.

II. SERVER LEVEL COLLECTION

2.1 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

#Software: Microsoft Internet Information Services 7.0
#Version: 1.0
#Date: 2013-07-31 08:55:05
#Fields: date time s-ip cs-method cs-uri-stem cs-uri-query s-port cs-username c-ip cs(User-Agent) sc-status sc-substatus sc-win32-status time-taken
2013-07-31 08:55:04 192.168.50.12 GET /HomePage/Header/banner4.jpg - 443 - 192.168.50.1 Mozilla/5.0+(Windows+NT+5.1)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Chrome/28.0.1500.72+Safari/537.36 200 0 0 7597
2013-07-31 08:55:04 192.168.50.12 GET /App_Themes/Basic/images/top_shadow.png - 443 - 192.168.50.1 Mozilla/5.0+(Windows+NT+6.1;+rv:13.0)+Gecko/20100101+Firefox/13.0.1 200 0 0 234
2013-07-31 08:55:04 192.168.50.12 GET / - 443 - 192.168.50.1 Opera/9.80+(Series+60;+Opera+Mini/6.5.27309/30.3558;+U;+en)+Presto/2.8.119+Version/11.10 302 0 0 358
2013-07-31 08:55:04 192.168.50.12 GET /App_Themes/Basic/images/bul.png - 443 - 192.168.50.1 Mozilla/5.0+(Windows+NT+5.1)+AppleWebKit/535.2+(KHTML,+like+Gecko)+Chrome/15.0.854.0+Safari/535.2 404 0 2 2324
2013-07-31 08:55:04 192.168.50.12 GET /ScriptResource.axd d=jCTqj8ncFwtBXv1EeDlLKnD11z7u_RSrLVZMe-QTPMWFeFdofVjOIJsAvBwnLaWfJpvMrHpziA0GnyyWKgM71nIGXcGoUgOrsHxRPrE5lwwrKY_K6tfCxWTW-IDYbBUF_i1j8vtcztlNLq_pKPlFDQ2&t=ffffffffd2acd832 443 - 192.168.50.1 Mozilla/5.0+(Windows+NT+5.1)+AppleWebKit/535.2+(KHTML,+like+Gecko)+Chrome/15.0.854.0+Safari/535.2 200 0 0 47549
2013-07-31 08:55:04 192.168.50.12 GET /App_Themes/Basic/images/p7sc3_psplay.png - 443 - 192.168.50.1 Mozilla/5.0+(Windows+NT+5.1)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Chrome/28.0.1500.72+Safari/537.36 200 0 0 421
2013-07-31 08:55:04 192.168.50.12 GET /HomePage/Header/banner14.jpg - 443 - 192.168.50.1 Mozilla/5.0+(Windows+NT+5.1)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Chrome/28.0.1500.72+Safari/537.36 304 0 0 655
2013-07-31 08:55:05 192.168.50.12 GET / - 443 - 192.168.50.1 Mozilla/5.0+(Windows+NT+6.1;+WOW64)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Chrome/28.0.1500.72+Safari/537.36 302 0 0 78
2013-07-31 08:55:05 192.168.50.12 GET /index.aspx - 443 - 192.168.50.1 Mozilla/5.0+(Windows+NT+6.1;+WOW64)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Chrome/28.0.1500.72+Safari/537.36 200 0 0 6552

2013-07-31 08:55:05 192.168.50.12 GET /App_Themes/Basic/images/img.jpg - 443 - 192.168.50.1 Mozilla/5.0+(Windows+NT+6.1;+WOW64)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Chrome/28.0.1500.72+Safari/537.36 404 0 2 655
2013-07-31 08:55:05 192.168.50.12 GET /HomePage/Header/banner1.jpg - 443 - 192.168.50.1 Mozilla/5.0+(Windows+NT+5.1)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Chrome/28.0.1500.72+Safari/537.36 304 0 0 561
2013-07-31 08:55:05 192.168.50.12 GET /HomePage/Header/banner2.jpg - 443 - 192.168.50.1 Mozilla/5.0+(Windows+NT+6.1)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Chrome/28.0.1500.95+Safari/537.36 200 0 0 23618
2013-07-31 08:55:05 192.168.50.12 GET /App_Themes/Basic/images/bul.png - 443 - 192.168.50.1 Mozilla/5.0+(Windows+NT+6.1;+WOW64)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Chrome/28.0.1500.72+Safari/537.36 404 0 2 483
2013-07-31 08:55:05 192.168.50.12 GET /App_Themes/Basic/images/bg.jpg - 443 - 192.168.50.1 Mozilla/5.0+(Windows+NT+6.1;+rv:13.0)+Gecko/20100101+Firefox/13.0.1 200 0 0 1544
2013-07-31 08:55:05 192.168.50.12 GET /App_Themes/Basic/js/jquery.nivo.slider.pack.js - 443 - 192.168.50.1 Opera/9.80+(J2ME/MIDP;+Opera+Mini/6.0.24093/30.3558;+U;+en)+Presto/2.8.119+Version/11.10 200 0 0 3088
2013-07-31 08:55:07 192.168.50.12 GET /HomePage/Header/banner2.jpg - 443 - 192.168.50.1 Mozilla/5.0+(Windows+NT+5.1)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Chrome/28.0.1500.72+Safari/537.36 304 0 0 592
2013-07-31 08:55:07 192.168.50.12 GET /HomePage/Header/banner13.jpg - 443 - 192.168.50.1 Mozilla/5.0+(Windows+NT+5.1)+AppleWebKit/535.11+(KHTML,+like+Gecko)+Chrome/17.0.963.12+Safari/535.11 304 0 0 234

2.2 Format of log files from Internet Information server (IIS)

Log files are found in many different formats for example files from internet which are being used in this paper, log files from intranet and ftp log files with different parameters are explained with their filed name how they appear and about their description as below.

2.2.1 Intranet Log file web Site with description

Field	Appears as	Description
Date	2002-05-02	This log file entry was recorded on May 2, 2002.
Time	17:42:15	This log file entry was recorded at 5:42 P.M. UTC. Entries are recorded to the log file when the send completion for t IIS send occurs.
c-ip	172.22.255.255	The IP address of the client.
Cs-username	-	The user was anonymous.
s-ip	172.30.255.255	The IP address of the server.
s-port	80	The server port.
Cs-method	GET	The user issued a GET, or download, command.
Cs-uri-stem	/images/picture. jpg	The user wanted to download the picture.jpg file from the Images folder.
Cs-uri-query	-	The URI query did not occur. (URI queries are necessary only for dynamic pages, such as ASP pages, so this field us contains a hyphen for static pages.)
Sc-status	200	The request was fulfilled with no errors.
csfUser-Agent) (compatible;MSIE+5.5; +Windows+2000+Server)	Mozilla/4.0+	The type of browser that the client used, as represented by the browser.

2.2.2 Internet Log File Format with Description

Field	Appears As	Description
Date	2002-05-24	This log file entry was recorded on May 24, 2002.
Time	20:18:01	This log file entry was recorded at 8:18 P.M. UTC.
c-ip	172.224.24.114	The IP address of the client.
cs-username	-	The user was anonymous.
s-ip	206.73.118.24	The IP address of the server.
s-port	80	The server port.
cs-method	GET	The user issued a **GET,** or download, command.
cs-uri-stem	/Default.htm	The user wanted to download the contents of Default.htm.

cs-uri-query	-	The URI query did not occur.
sc-status	200	The request was fulfilled without error.
sc-bytes	7930	The number of bytes that the server sent to the client.
cs-bytes	248	The number of bytes that the client sent to the server.
time-taken	31	The action was completed in 31 milliseconds.
cs(User-Agent)	Mozilla/4.0+(compatible;+MSIE+5.01;+Windows+2000+Server) represented by the	The type of browser that the client used, as browser.
cs(Referrer) http://62.224.24.114/	The Web page that provided the link to the Web site.	cs(Referrer) http://62.224.24.114/

2.2.3 FTP log file format is also used.

Field	Appear	Description
Time	16:40:23	rc(This) rc(log) rc(file) rc(entry) rc(was) rc(recorded) rc(at) rc(4:40) rc(P.M.) rc(UTC.)
rc(c-ip)	rc(10.152.10.200)	rc(The) rc(IP) rc(address) rc(of) rc(the) rc(client.)
rc(Cs-method)	rc([6994]USER)	rc(The) rc(USER) rc(FTP) rc(command) rc(was) rc(used,) rc(which) rc(requests) rc(a) rc(user) rc(name) rc(and) rc(is) rc(always) rc(followed) rc(by) rc(a) rc(PASS) rc(FTP) rc(command.) rc(6994) rc(is) rc(the) rc(connection) rc(number) rc(corresponding) rc(to) rc(an) rc(anonymous) rc(user.)
rc(Cs-uri-) rc(stem)	rc(Anonymous)	rc(The) rc(user) rc((the) rc(target) rc(of) rc(the) rc(USER) rc(command)) rc(was) rc(anonymous.)
rc(Sc-status)	rc(331)	rc(The) rc(user) rc(name) rc(was) rc(accepted.)
rc(Time)	rc(16:40:25)	rc(The) rc(next) rc(recorded) rc(action) rc(occurred) rc(at) rc(4:40) rc(P.M.) rc(UTC.)
rc(c-ip)	rc(10.152.10.200)	rc(The) rc(IP) rc(address) rc(of) rc(the) rc(client.)
rc(Cs-method)	rc([6994]PA5S)	rc(The) rc(PASS) rc(FTP) rc(command) rc(was) rc(used,) rc(which) rc(supplies) rc(a) rc(password) rc(for) rc(the) rc(user) rc(name) rc(and) rc(is) rc(always) rc(preceded) rc(by) rc(a) rc(USER) rc(command.)
rc(Cs-uri-) rc(stem)	rc(anonymous@example.net)	rc(The) rc(password) rc((the) rc(target) rc(of) rc(the) rc(PASS) rc(command)) rc(supplied.)

III. 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.

IV. 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 . In this paper we are doing cleaning in order to get status of http code with the help of log parser tool.

4.2User Identification

We have used the Internet log file entries to find the Number of hits in a particular time period. 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.

V. RESULTS

In this study, we have analyzed the log files of Web server of BundelKhand University (www.bujhansi.org) with the help of log analyzer program. The log files consists the data from **2013-07-31** to **2013-08-08** . In this duration log files have stored 60 MB data and we have got 4.4 MB data after preprocessing. We have determined different types of errors that occurred in web navigation. Statistics about hits, page views, visitors and bandwidth are shown in Table 1. Table 2 shows the daily errors types. Different types of errors are shown it is clear from the Table that 404 (Table 2) is most frequently occurred error. 404 is a frequently-seen status code that tells a Web user that a requested page is "Not found." 404 and other status codes are part of the Web's Hypertext Transfer Protocol (HTTP), written in 1992 by the Web's inventor, Tim Berners-Lee. He took many of the status codes from the earlier Internet protocol for transferring files, the File Transfer Protocol (FTP). So it is very clear that the mistake is in Html coding which, the web developers should be more care full in opening and closing html tags , so this error can be very easily removed and a newly developed website will more interactive with the end user.

Some other types of client and server errors are shown .

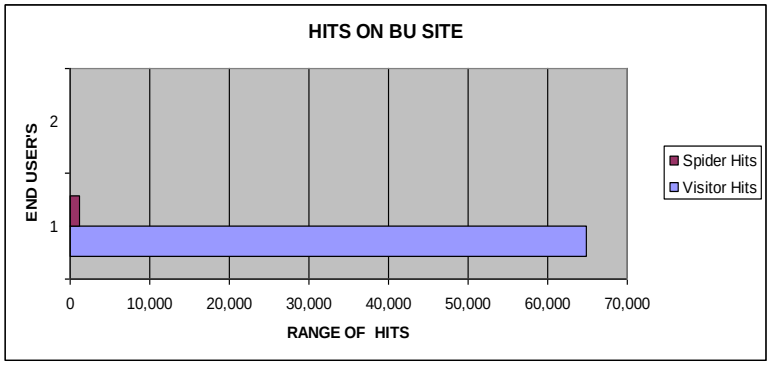
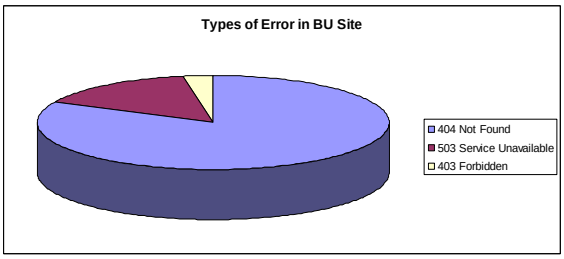


Table 1Hits on BU Site

Visitors Hits	64,876
Spider Hits	1,223
Total	66,099
Average Hits Per Day	8,673
Average Hits Per Visitor	8.16
Cached Requests	4,979
Failed Requests	333
Page Views	
Total Page Views	5,435
Average Page Views per Day	654
Average Page Views per Visitor	1.24
Visitors	
Total Visitors	3485
Average Visitors per Day	447
Total Unique IPs	3,038
Bandwidth	
Total Bandwidth	577.48 MB
Visitor Bandwidth	558.81 MB
Spider Bandwidth	19.67 MB
Average Bandwidth per Day	74.94 MB
Average Bandwidth per Hit	21.07 KB
Average Bandwidth per Visitor	162.15 KB

Table 2 Types of Error which were found in BU web Site are

Sr. No	**Error**	**Hits**
1	404 Not Found	299
2	503 Service Unavailable	55
3	403 Forbidden	10
	Total	364



VI. 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].

VII. CONCLUSIONS

In this fast growing age of IT, people do not have any time ,so end user need user friendly sites which should have the capability of handling error . System administrator and web developer should try to increase site 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 developer and can arrange their system by determining occurred system errors, corrupted and broken links. In this study, analysis of web server log files of Bundlelkhand University web site has been done by using web log expert program. We are lucky to have so many sites to assist us for rectification in our web development style according to the current growth in web.. With the growth of web-based applications web usage and data mining to find access patterns is a growing area of research. Web mining techniques which are based on like association rules, sequential patterns, clustering and classification can be used to predict frequent patterns for daily update.

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 intotransactions 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 InformationRetrieval 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

[20] Tyagi, Navin Kumar; Solanki, A. K.; Wadhwa, Manoj "Analysis of Server Log by Web Usage Mining for Website Improvement" International Journal of Computer Science Issues (IJCSI);Jul2010, Vol. 7 Issue 4, p17

