Web Usage Pattern Analysis Through Web Logs: A Review

Dilip Singh Sisodia
Department of Computer Sc. & Engg.
National Institute of Technology Raipur, India
Sisodia_dilip@rediffmail.com

Shrish Verma
Department of Information Technology & EC
National Institute of Technology Raipur, India
sverma@nitrr.ac.in

Abstract—**Web server log repositories are great source of knowledge, which keeps the record of web usage patterns of different web users. The Web usage pattern analysis is the process of identifying browsing patterns by analyzing the user's navigational behavior. The web server log files which store the information about the visitors of web sites is used as input for the web usage pattern analysis process. First these log files are preprocessed and converted into required formats so web usage mining techniques can apply on these web logs. This paper reviews the process of discovering useful patterns from the web server log file of an academic institute. The obtained results can be used in different applications like web traffic analysis, efficient website administration, site modifications, system improvement and personalization and business intelligence etc.**

Keywords-Log Repositories,Web usage pattern, Web server log, Knowledge extraction, Web usage mining, browsing patterns, and Useful patterns.

I. INTRODUCTION

Web server log repositories are great source of knowledge, which keeps the record of web usage patterns of different web users. The Web usage mining is the process of applying data mining techniques to discover hidden, valuable and interesting usage patterns from web data, in order to understand and better serve the needs of web-based applications. Digging out hidden and fascinating knowledge from web server logs data is going to more and more appealing day by day so web usage mining research engrossed lot of thoughts in present time [1].

A. Background

Web mining is an appliance of data mining techniques to large web log data repositories [2].This term was coined by Etzioni in 1996[16]. The whole Web mining process is generally divided in to three different but interdependent categories by researchers on the basis of input data used by them like web structure mining, web content mining and web usage mining. In 1971 R.Cooley introduced specific term web usage mining which is defined as the process of automatic discovery of user access behavior patterns from web servers log data [1].

While the web content and structure mining is mainly dependent on the primary data of the web, while web usage mining uses the secondary data which is derived from the interactions of the users with the web. On the basis of spatial location based collection of user interactions record, this data may be further classified in to three different categories: first the web server data (which is a cumulative representation of the usage of a service by all users), second the client side data (which is the complete depiction of usage of all services by a particular client), and third the proxy server data (which is being somewhere in the middle). Other auxiliary information like user registration data, user sessions, browser cookies, user queries, click streams, and any other data generated during the user interactions with web may be used for web usage mining process. The obtained results can be used in various applications like traffic analysis, efficient website administration, adequate hosting resources, site modifications, system improvement and personalization and business intelligence.

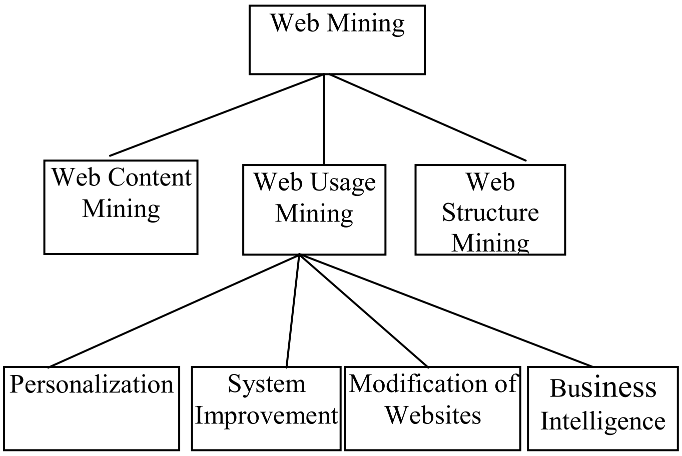

Figure 1: Web Mining Categories

B. Frequently used terms

Following terms are frequently used in web usage mining process: A resource, according to the W3C's Uniform Resource Identifier (URI) specification, can be anything that has identity" Moreover, an URI is a compact string of characters for identifying an abstract or physical resource". Examples are HTML file, an image, and a Web service etc. A Web resource is a resource accessible through any version of the HTTP protocol (for example, HTTP 1.1 or HTTP-NG).A Web server is the server that provides access to the Web resources. A Web page is the collection of Web resources that can be identified by an URI. If the Web page consists of n resources, the first n - 1 are embedded into the nth URI, which identifies the Web page. A page view when a Web browser displays a Web page

978-1-4673-1921-8/12/$31.00 ©2012 IEEE

at a particular moment in time also known as hit. A Web browser is software installed at client side and used to send Web requests, handle the responses, and display requested URIs. A user is a person using a Web browser. A Web request is a request made by a Web client for a Web resource. It can be explicit or implicit. Explicit Web requests are further classified as embedded or user input .implicit web request are generated by web client and it requires embedded resources from web page in order to display that page. A user session consists in a delimited number of a user's explicit Web requests across one or more Web servers. A visit represents a subset of consecutive page views from a user session occurring close enough (measured by means of a time threshold or a semantically distance between pages). An episode is a subset of related clicks which occurs within a user session.

C. Layout of paper

The organization of this paper is as follows: Section 2 provides the overview of Web usage pattern analysis process. Section 3 gives various applications of analyzed patterns. Section 4 briefly describes the major related work in this area. Section 5 describe the common web log content which need to be analyzed. Section 6 gives idea about different Web log formats. Sections 7 discuss the results while section 8 contains the conclusion and future work.

II. OVERVIEW WEB USAGE PATTERN ANALYSIS PROCESS

Web usage mining or web log mining process can be regarded as a five-phase process consisting [17]:

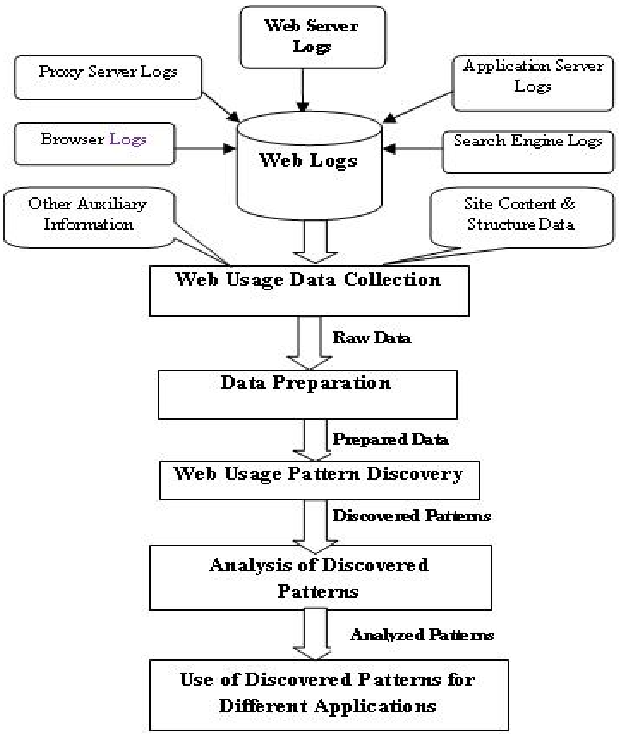
Web Server Logs
Proxy Server Logs
Application Server Logs
Browser Logs
Web Logs
Search Engine Logs
Other Auxiliary Information
Site Content & Structure Data
Web Usage Data Collection
Raw Data
Data Preparation
Prepared Data
Web Usage Pattern Discovery
Discovered Patterns
Analysis of Discovered Patterns
Analyzed Patterns
Use of Discovered Patterns for Different Applications

Figure 2: Web Log Mining Process

A. Usage data collection:

Web usage logs, which record user activities on Web sites, can be collected from i) Web servers, ii) Web proxy servers, and iii) client browsers,

B. Data preparationn:

The data collected from the logs may be partial, deafening and conflicting so the objective of preprocessing is to transfer raw log files in particular format which data mining algorithms can handle easily. The main tasks of preprocessing are:,

1) Data Cleaning – removes log entries that are not needed for the mining process.
2) User Identification- differentiated the Log records according to users for the analysis.
3) Session Identification- the activity of a user from the moment he/she enters the web site until the moment he/she leaves it. Any User can visit the particular website many times during a specific time period. Session identification aims at dividing the multi-visiting user sessions into single ones.
4) Path completion- finds whether there is hyperlink between the previous page and following page.
5) Data integration- stores various data properly and handles data conveniently by making use of database system and database management system respectively.
6) Formatting-Convert the preprocessed data in particular format to smoothly apply the analysis techniques.

C. Pattern discovery:

Statistical methods as well as data mining methods (path analysis, Association rule, Sequential patterns, and cluster and classification rules) are applied in order to detect interesting patterns.

D. Pattern analysis phasey:

In this phase extracted patterns are analyzed through OLAP tools, knowledge management query techniques and intelligent agents to sort out the monotonous rules/patterns.

III. APPLICATIONS OF ANALYZE PATTERNS

We can apply the result of extracted patterns to the following major areas, among other

1) Improving site/page design,
2) Making additional topic or product recommendations,
3) Web personalization, and Learning user behavior
4) Web caching,

Analyzing web log data (or web log analysis) is more popular in e-commerce sites (for profit web sites) than others. In the e-commerce web sites, analyzing web log data may lead to studying patterns of visitors, browsing and purchasing habits. This then may lead to tailoring the design of the web site to fit certain group of visitors or certain other patterns that were discovered from analyzing the data. Web log analysis is not limited to e-commerce sites, instead other organizations may benefit from analyzing data about the visitors of their web sites. Some of the sites that reported benefiting from this type of analysis may include web sites for newsletters, service organizations, as well as other type of web sites [5].

Plenty of tools to analyze web server log data are available with their underlying strengths and limitations. No single tool is sufficient for extracting the all type of knowledge from a given web server log. Apart from the tools accuracy of extracted knowledge patterns are heavily dependent on quality of data. the quality of web log data can improve by applying appropriate preprocessing techniques. Approximately 80% of mining labors often spend to improve the quality of data [7].

Web log quality may improve by removing non human log entries or non contributory information from log data. Session identification heuristics which represents the complete user session, path completion methods to represent complete user navigation path and other formatting of web log server may be applied to enhance quality of data. In this paper we apply all these methods to improve the web log data quality.

IV. RELATED WORK

The past years have seen the tremendous growth in the area of Web Usage Mining research. Since the early papers published in the mid 1990s[1,2,3,12], more than 400 papers on Web Mining have been published; more or less than 150 papers, of the overall 400, have been before 2001; around the 50% of these papers regarded Web Usage Mining. The first workshop entirely on this topic, WebKDD, was held in 1999. Since 2000, more than 150 number of research papers had published on Web Usage Mining which shows a dramatic increase of interest for this area [13]. Basic introduction of web mining and web usage mining is given by [1, 2, and 3]. A detail review of web mining research is presented in [12]. An overview of application soft computing techniques (neural network, fuzzy logic, genetic algorithms) in web usage mining is presented in [14, 15].This paper presents the survey of the recent developments in the area of Web Usage Mining. by applying this method on analyzing a web server log data of an academic institutions web server.

V. RELATED COMMON WEB LOG CONTENTS WHICH NEED TO BE ANALYSED

A Web server log file store the information about the requests made to the Web server in a chronological order. The stored web log file contains huge volume of data. Examine each and every type of data is not useful. Thus examining the most pertinent, relevant and useful information in the weblog data may provide more specific information about the patterns for visitors of the web site. This section lists some items that are commonly presented in web log reports. It explains each item, the meaning of the item presented in the report and also the benefits that it potentially brings to the web site from understanding and analyzing it. The raw log files consists of 19 attributes such as Date, Time, Client IP, Auth User, Server Name, Server IP, Server Port, Request Method, URI-Stem, URI-Query, Protocol Status, Time Taken, Bytes Sent, Bytes Received, Protocol Version, Host, User Agent, Cookies, Referrer. These attributes are part of the following logs:

A. Access/Transfer log

The transfer/access log store the detailed information of each request made from user's web browsers to the server. For example, through the host, it can be determined the geographical location of the host, this includes country, state and city in which the viewer of the page is located. This may help in redefining the page for certain population coming from a specific geographical area.

B. Error logs

The error log contains information about errors and failed requests. If the page contains a links to a file does not exist or if the user is unauthorized to access a particular page or file than request may fail.

C. Agent (Browser) log

This log contains the information about the browsers and operating system used by different users to connect the server. Sometimes, a visitor can access a site from different browsers, counting these visitors differently may help understand the number of visits more precisely

D. Referrer log

When a user access web page by clicking on the link from other site then URL of that site is also recorded in log file which is known as refers log. Or the referrer log contains the URLs of pages on other sites that link to your pages. Further analysis of above variables can generate the following information.

1) Find the user's percentage that accesses the site from a particular domain type (e.g., .com, .edu, .net, .org, .gov). By hits versus accesses analysis.

2) Find the primary clients on the basis of number of hits the server is getting from different user groups.

3) Find the breadth of penetration of the servers. Through number of unique IP addresses visiting the site.

4) Find the optimal time/day by analyzing the quantity of accesses/hits the server receives during specific hours and days of the week. For server maintenance and up gradation.

5) To perform server maintenance and/ or upgrades. Find the average length of a user's session, average time spent by a user on a particular web page, average download times, and navigation pattern of a user through the path analysis of a user on a web site.

The data from Access Logs provides a broad view of a Web servers and users (as indicate by IP address).this type of analysis empowers web server administrators and decision makers to identify potential users of their services and efficient management of their web infrastructure.

VI. WEB LOG FORMATS

World Wide Web consortium is a organization to provide standard format for web server log files, but there exist some other proprietary formats also. For example IIS provides six different log file formats which are used to track and analyze information about IIS-based sites and services such as

- W3C Extended Log File Format
- W3C Centralized Logging
- NCSA Common Log File Format
- IIS Log File Format
- ODBC Logging
- Centralized Binary Logging.

In addition to the six available formats, custom log file format can also be configured. A log file in the W3C extended format contains a sequence of lines containing ASCII characters. Every line in a log file may include either a command or an entry. Commands or directives are begins with # character and contains the information about logging process, software, version etc. But entries are sequence of fields corresponding to a single HTTP transaction. Different fields in entries are separated by white space. If a particular field is not recorded in a log file entry then it is represented by dash "-" marks. The following directives are defined in the W3C Extended format [8]; the following are examples of different log file formats recorded by systems:

1. NCSA Common log file format:

```
172.21.13.45-REDMOND\fred[08/Apr/1997:17:39:04-0800] "GET/scripts/iisadmin/ism.dll?http/serv HTTP/1.0" 200 3401
```

Description of headers- Remote host address, Remote log name (This value is always a hyphen), User name, Date, time, and Greenwich mean time (GMT) offset, Request and protocol version, Service status code (A value of 200 indicates that the request was fulfilled successfully), Bytes sent

2. W3C Extended log format produced by the Microsoft Internet Information Server (IIS):

```
#Software: Microsoft Internet Information Server 4.0 #Version: 1.0 #Date: 2011-07-05 22:48:39 #Fields: date time, c-ip, cs-username, s-ip, cs-method, cs-uri-stem, cs-uri-query, sc-status, sc-bytes, cs-bytes, time-taken cs-version, cs-User-Agent, cs-Cookie, cs-Referrer.
```

```
2011-07-05 22:48:39 206.175.82.5 -208.201.133.173 GET/global/images/topborder.gif - 200540 324 157 HTTP/1.0Mozilla/4.0+(compatible;+MSIE+4.01;+Windows+95)USERID=CustomerA;+IMPID=01234 http://yourturn.rollingstone.com/webx?98@@webx1.html
```

Description of headers: c- Client, s –Server, r –Remote, cs -Client to Server, sc -Server to Client, sr -Server to Remote Server(this prefix is used by proxies), rs-Remote Server to Server(this prefix is used by proxies),x-(Application specific identifier).

This is the identifying information that the client browser reports about itself. More recent entries are appended to the end of the file. This information can be stored in a single file, or distributed into different logs files, such as an access log, error log, or referrer log. Web usage mining research focuses on finding patterns of navigational behavior from users visiting website. The extracted knowledge of user's navigational behavior from web log file, which is recorded in any of the above format, may be used to answer different queries like efficiency of web site in delivering information, users view point about web site structure, prediction of users next visit, fulfillment of needs of different users, user satisfaction, web content personalization and many more such type of information to facilitate web administrators in taking decision. Flourishing websites can be tailored to meet user preferences both in the appearance of information and in significance of the content that best fits the user requirement [18].

VII. RESULTS

In this paper, we are presenting the analysis of the NITR log files of a Web server for period 27/11/2011 to 25/12/2011 with the weblog explorer program. This Analysis is performed directly on the log files and no separate data warehouse is required [9]. Web log analyzer has identified several Web usage access pattern by applying well known data mining techniques to the access logs. This includes descriptive statistic and Association Rules including support and confidence to represent the Web usage and user behavior. The size NITR log file is 57.7MB which consists of more then one lakh entries of particular duration [9]. Table.1 shows summaries statistical information of NITR log file of above duration. Other analysis about trends of visitor's access patterns, bandwidth consumption, referrals, user agents and different types of errors that occurred in web surfing, statistics about hits, page views, visitors and bandwidth are shown. The following figures are showing visitors per week (Fig.3), Browsers v/s visitors (Fig.4) and Visitors v/s response code (Fig.5).

TABLE 1: of NITR Log analysis results Summary

Different Parameter	Value
Unique IP	6
Visitors	683
Hits	198812
Bandwidth	2.86 GB
Pages/Files	696
Errors	16958
Search Engines	2
Search Phrases	52
Referring Sites	19
Referrers	601
Countries	1
Entry Points	39
Spiders	11
Page Views	45758
Average Page Views per Visitor	67

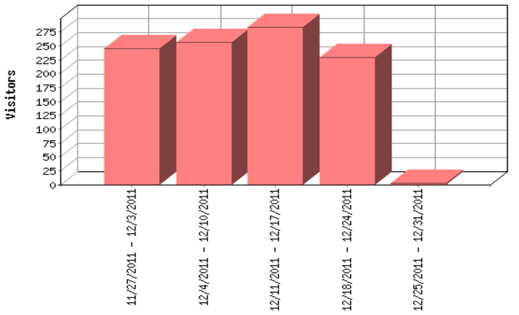

Figure 3: Visitors per week

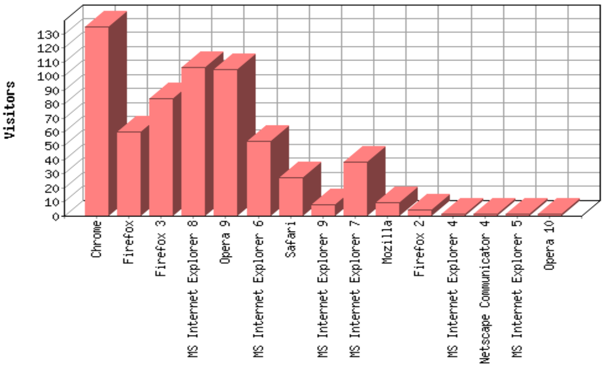

Figure 4: Browsers v/s Visitors

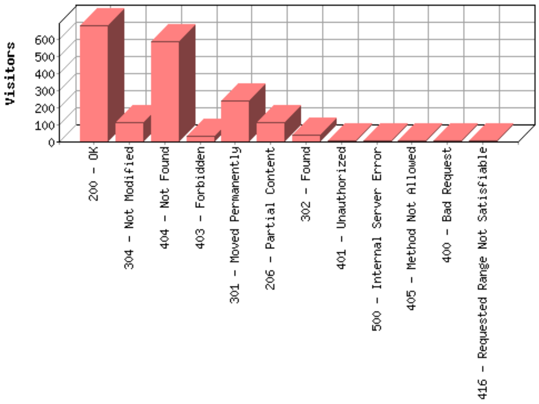

Figure 5: Visitors v/s Response code

VIII. CONCLUSIONS AND FUTURE WORK

Web usage mining process is used to discover hidden and interesting user navigation patterns, which is applicable to find solutions of many real-world problems such as Web page/site improvement, additional product/topic recommendations, user/customer behavior analysis, etc. This paper has provided a review and analysis of current Web usage mining systems, with analysis of the academic institutions web server log file for particular duration using web log analyzer program [9]. The results of analysis may be used to for following purposes to increase popularity of web site amongst its visitors, to increase e usefulness of web pages for medium of revenue generation, for diagnostic purposes such as for detection of system errors, tarnished and wrecked links. Other web server logs may be used for similar kind of studies to increase the effectiveness of web portals or to better understanding of user behavior. The main research challenges of this field are identifying the non human entries made by web robots, design efficient heuristics for user session identification and finding association among different user's access 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] Y. Rubin, Jeffrey H. (2004). Log Analysis Pays Off.(weblogs). Network Computing, 15 (18), 76-78.

[5] Velayathan, Ganesan; Yamada, Seiji (2006). Behavior based web page evaluation. Proceedings of the 16th international conference on World Wide Web, 1317 – 1318. Retrieved May 13, 2010, from ACM Digital Library,http://www.acm.org/dl.

[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] Extended Log File Format, http://www.w3.org/TR/WD-logfile.html

[9] http://www.nitrr.ac.in, http://www.extratrend.com/

[10] 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.

[11] 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.

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

[13] F.M. Facca, and P.L. Lanzi," Mining interesting knowledge from web logs: a survey", Elsevier Science, Data and Knowledge Engineering, 2005, 53, pp.225-24.

[14] Tug, E., Sakiroglu, and A.M. Arslan, "Automatic discovery of the sequential accesses from web log datafiles via a genetic algorithm", Knowledge based System, 2006, pp.180-186.

[15] 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.

[16] Etzioni, The world-wide Web: quagmire or gold mine? Communications of the ACM 39 (11) (1996) 65–68.

[17] Chen Hu, Xuli Zong, Chung-wei Lee and Jyh-haw Yeh, "World Wide Web Usage Mining Systems and Technologies", Systemic, Cybernetics and Informatics Vol. 1 – Number 4.

[18] Dilip Singh Sisodia and S.Verma "Application of Weblogs to Construct Smart Web servers to handle user Traffic efficiently' International Journal of Advanced Computer Eng ineering and Architecture (IJACEA), Vol. 1, No. 1, June 2011 Number 7.