International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Knowledge Retrieval from Web Server Logs Using Web Usage Mining

Naresh Kumar Kar[1], H. R. Sharma[2], Asha Ambhaikar[3]

[1, 2, 3]Rungta College of Engineering and Technology, Kohka-Kurud Road, Bhilai, India

Abstract: ***Now a days the data of internet is increasing day by day and web-based applications are very common. The main problem that faces any website admin or any web application system is data increase per-second, which is stored in different types and formats in server log files about users, their future needs and maintains the structure and content of website or web services according to their previous data. The main aim of web usage mining is discovering useful information or knowledge from usage data registered in log files, based on primary kinds of data used in the mining process. Using one of the web mining techniques, this paper uses a web usage mining technique to find knowledge from web server log files where all user navigation history is registered.***

Keywords: Web usage mining; navigation patterns; log file.

1. Introduction

In Internet based business Web Usage Mining (WUM) is an active field of research and is most likely to generate new knowledge. WUM applications are being used in some famous websites to understand customer's profiles and their performance in terms of strengths and weaknesses of their website. This paper gives a brief introduction of WUM lifecycle, a data mining technology and WUM implementation [1]. The main problem that faces any website admin or any web application is data increasing per-second, which is stored in different type and format in the server log file. Learning about the users and expect or predict their needs in the future and maintaining the structure and content of the website or web service according to their previous data. The paper problem is the presence of large number of users in the websites, the presence of huge amount of data about them and the need of the web master and site administration to technology enables them to make decision fast.

Datasets of server log files have been selected for this study to implement WUM technique. The main purpose of using WUM is gathering information about user navigation patterns. This information can be exploited later to improve the web site from the user's point of view. The results produced by the mining of web logs can be used for various purposes, for example to personalize the delivery of web content, to improve user navigation through prefetching and caching web design and customer satisfaction.

2. Related Work

Now a days, WUM has been one of the best areas of researches. WUM technique being used widely to discover user navigation patterns from web server logs. DNS traces user interests by applying a keyword matching approach, across corresponding domain names, or by operating on top of a search engine. To identify online user and their behavior or interests across network [2]. Analyze of website errors which help system administrator and web designer to improve their system by determining systems errors, corrupted and broken links by using WUM [3]. A novel approach for classifying user navigation patterns and predicting user future requests is introduced in [4]. In another approach, data from a data warehouse and web data can be used to improve marketing activities [5].

3. Source of Data for WUM

3.1 Proxy server logs

A web proxy is a caching mechanism which lies between client browsers and Web servers. It helps to reduce the load time of web pages as well as the network traffic load on both sides (i.e. server and client). Proxy server logs contain HTTP requests from multiple clients to multiple web servers. This may serve as a data source to discover the usage pattern of an anonymous user groups, sharing a common proxy server.

3.2 Web Server logs

Log file registered information about user request history (i.e. Information about the request, including client IP address, request date/time, page requested, HTTP code, bytes served, user agent, and referrer are typically registered). These data can be combined into a single file, or separated into distinct logs, such as an access log, error log, or referrer log. Server logs do not typically collect user specific information. These files are not usually accessible to general Internet users, only to the webmaster or other administrative entity [7].

3.3 Browser logs

Collect client-side data the JavaScript and Java applets are used. This implementation of client-side data collection requires user cooperation, either in enabling the functionality of the JavaScript or Java applets in modified browser [7].

4. Log File Format

The Common Log Format [W3C] is the most popular log file formats and an extended version. The W3C provides standard

format (Common Log Format) for web server log files (see Fig 1). Information obtained from log files are explained as follows:

Browser Type: This gives the information of the type of browser that was used for accessing the website.

Visitor Referral Website: The referral website gives the information or URL of the website which is being referred to by the particular website in consideration.

Number of Hits: This number of times any resource is accessed in a Website. When a web page is uploaded from a server the number of "hits" or "page hits" is equal to the number of files requested.

Number of Visitors: It's a user who navigates to website and browses one or more pages.

Platform: This information gives the type of operating system used to access the website [8].

Cookies: A message given to a web browser by a web server. The browser stores the message in a text file called cookie. The main purpose of cookies is to identify users and possibly prepare customized web pages for them [7].

Visitor Referring Website: The referring website gives the information or URL of the website which referred the particular website in consideration.

Time and Duration: The time and duration for how long the website was accessed by a particular user.

Path Analysis: Path analysis gives the analysis of the path a particular user has followed in accessing contents of a website.

Visitor IP Address: This information gives the Internet Protocol (IP) address of the visitors who visited the website in consideration.

5. WUM Process

Web Usage Mining is a process similar to exploration in the data mining with different data sources type and tools used. Fig (3.1) Web Usage Mining process is a series of steps which can be summarized as follows:

5.1 Data Collection

Within this stage, usage data from various sources are collected from web servers, clients connected to a server, or from middle sources such as proxy servers and packet sniffers.

Data of a typical web server is shown in fig (1), in following there is a sample data of the first raw of the log file:
1.2012-10-12 00:01:21 W3SVC4045 C27384-57916 70.87.39.67 GET /robots.txt - 80 - 24.232.136.71 HTTP/1.1 Mozilla/4.0 + (compatible; + MSIE + 6.0; + Windows + NT + 5.1.

Figure 1: An example of raw log file.

The log file is a customizable ASCII text-based format. The field prefixes in the file are defined as follows:
s : Server actions.
c :Client actions
sc:Server-to-Client actions.
cs Client-to-Server actions.
Variable fields appear as: Date, time, s-sitename, s-computername, s-ip, cs-method, cs-uristem, cs-uriquery, s-port, cs-username, c-ip, cs- version, cs-useragent, cs-cookie, cs-referer, cs-host, sc- status, sc-substatus, sc-win32status, sc-bytes, and cs-bytes [7].

5.2 Data Preprocessing

Data preprocessing describes any type of processing performed on raw data to prepare it for another processing procedure. Commonly used as a preliminary data mining practice, data preprocessing transforms the data into a format that will be more easily and effectively processed for the purpose of the user. This is the stage where data are cleaned from noise, their inconsistencies are resolved, and they are integrated and consolidated in order to be used as input to the next stage of pattern discovery. The techniques that are used here can provide client data elaboration [9]. The different tasks of data preprocessing are:

Data Cleaning: The first step in data preprocessing is to clean the raw web data. During this step the available data are examined and irrelevant or redundant items are removed from the dataset. Irrelevant records are deleted during data cleansing. Since the target of WUM is to get traversal pattern [10]. Below are two kinds of unnecessary records which should be removed:

- The records with filenames extension of GIF, JPEG, CSS.
- The records with noisy data or uncompleted queries are removed.

Session Identification: The identification of user sessions also received significant attention in WUM process, as sessions encode the navigational behavior of the users and they are most important for pattern discovery. A user session is a delimited set of pages visited by the same user within the duration of one particular visit to a web site [11].

User Identification: The identification of individual users who access a website is one of the most important issues for the success of a personalized website. The simplest approach is to assign a different user to each different IP identified in the log file. Cookies are also useful for identifying the visitors of a website by storing an ID, which is generated by the web server for each user visiting the website

[11].

5.3 Pattern Discovery

In this stage, knowledge is discovered by classifying users according to their navigational activities. The goal of classification is to identify the distinguishing characteristics of predefined classes, based on a set of instances, e.g. users of each class [9]. Classification is the technique to map a data item into one of several predefined classes. This requires extraction and selection of features that best describes its properties of a given class or category [1].

5.4 Pattern Analysis

This is the final step in the WUM process. After the preprocessing and pattern discovery, the obtained usage patterns are analyzed to filter uninteresting information and extract the useful information. The methods like SQL (Structured Query Language) processing and OLAP (Online Analytical Processing) can be used. Relational query languages (SQL) allow users to pose queries for data retrieval. High-level data mining query languages need to be developed to allow users to describe data mining tasks by facilitating the specification of the relevant sets of data for analysis, the domain knowledge, the kinds of knowledge to be mined, and the conditions and constraints to be enforced on the discovered pattern. Pattern analysis enables us to do the automatic detection of patterns in data from the same source and make predictions of new data coming from the same source [12].

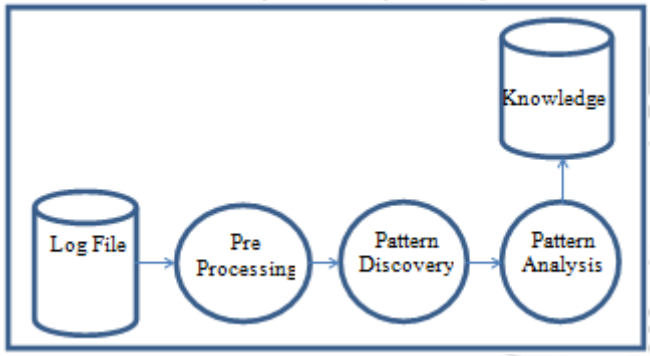

Figure 2: Web usage mining Process

6. Results

This study analyses the web server logs Fig (1) with the help of Deep log Analyzer program. The analyzed results can be seen as knowledge to answer these questions. How to extract knowledge from incomplete data structure? Is the log data that is gathered about the users enough to understand them? What is the optimal structure and content of a web site in order to attract the maximum interest of visitors? What does the user want to do? What is the suitable method and technique of web mining to extract knowledge?

Many different types of results occur according to WUM technique used. Visualization of WUM results should be expressed in high-level languages. The (visual) knowledge can be easily understandable and usable by humans. Implementation of WUM steps displays an example results below. For an analysis accessed resources to know user preferences, see Fig (3). For an analysis site navigation to understand user behavior, see Fig (4). For an analysis visitor activities to know of visit, see Fig (6). For an analysis of website popularity, search engines, and phrases used by visitors, see Fig (7). For an analysis, in terms of version, of browsers, of OSs, screen resolution, JavaScript Support Flash plug-ins or adds-on, all of which used for website optimization, see Fig (8). For diagnostics and error correction, see Fig (9). Following are some examples of results divided according to user information registered in log files:

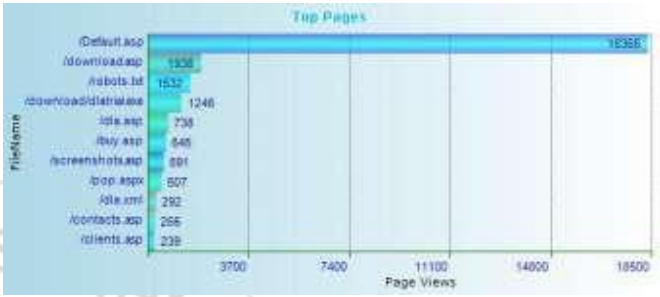

Figure 3: Top visited pages

Figure 4: Popular paths through site.

Figure 5: Visitors stay length.

Figure 6: Visits by hour of day.

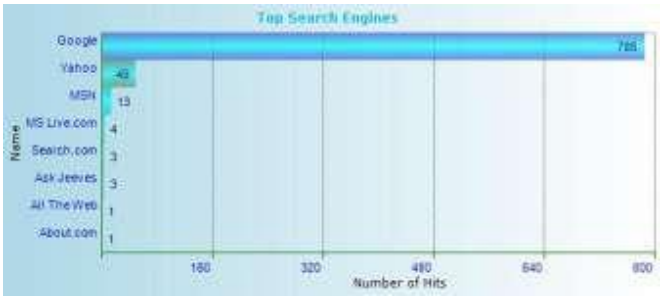

Figure 7: Top search engines.

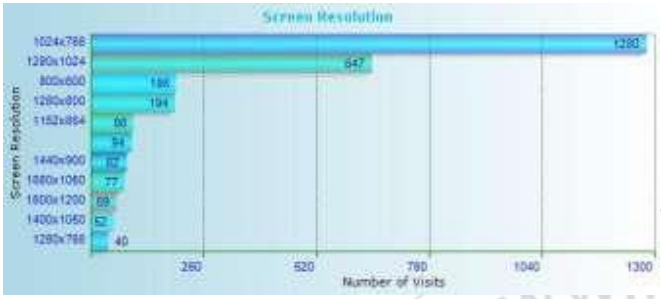

Figure 8: Popular paths through site.

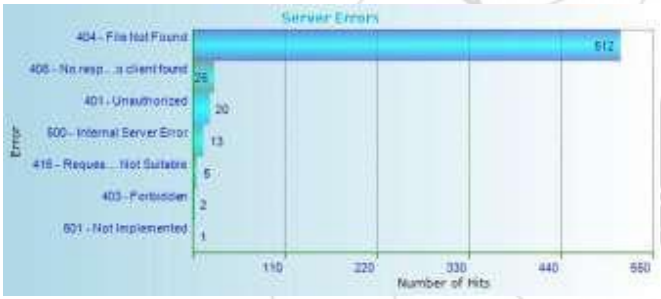

Figure 9: Server Errors.

7. Conclusion

In order to make a website popular among its visitors, website administrator and web designer should try to increase its effectiveness by understanding users and learning about them from registered usage information in log files. WUM technique is very good to extract knowledge from unstructured data. The obtained results of WUM can be used by web administrator or web designer to arrange their website by determining system errors, user's preferences, technical information about users, and corrupted and broken links. We also recognize that large amount of information stored in unstructured sources is calling for attention to develop innovative approaches to solve challenges of how to impart knowledge encoded in ontology into the unstructured data and how to explore more meaningful ways to utilize the knowledge?

Finally, this paper has important aspects data exploration and analysis of activity and preferences of users and this aspect is ignored by a lot of institutions, despite its importance in building a strong relationship between web admin and the users.

References

[1] Guandong Xu,"Web Mining Techniques for Recommendation and Personalization", Victoria University, Australia, March2008.
[2] Bamshad Mobasher "Data Mining for Web Personalization", LCNS, Springer-Verleg Berlin Heidelberg, 2007.
[3] Dr. R. Krishnamoorthi and K. R. Suneetha,"Identifying User Behavior by Analyzing Web Server Access Log File", International Journal of Computer Science and Network Security, April 2009.
[4] 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.
[5] K R. Kosala, and H. Blockeel," Web mining research: a Survey", SIGKDD Explorations, 2000, 2, pp.1-15.
[6] "Log files formats", www.kavach.mobishastra.com.
[7] S. K. Pani, "Web Usage Mining: A Survey on Pattern Extraction from Web Logs", International Journal of Instrumentation, Control & Automation, January 2011.
[8] Dr. G. N. Purohit, "Page Ranking Algorithms for Web Mining", International Journal of Computer Applications, January 2011.
[9] Faten Khalil," Combining Web Data Mining Techniques for Web Page Access Prediction", University of Southern Queensland, 2008.
[10] Yu-Hai tao, Tsung-Pei Hong andYu-Ming Su, "Web usage mining with intentional browsing data, ", Expert Systems with Applications, Science Direct,2008.
[11] Dr.D.Suresh Babu,"Web Usage Mining: A Research Concept of Web Mining", International Journal of Computer Science and Information Technologies, 2011.
[12] Renáta Iváncsy, István Vajk,"Frequent Pattern Mining in Web Log Data", Acta Polytechnica Hungarica, January 2006.
[13] Raju G.T. and Sathyanarayana P. "Knowledge discovery from WebUsage Data: Complete Preprocessing Methodology," IJCSNS 2008.