ijecs
open access www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3 Issue 6 June, 2014 Page No. 6739-6741

Web Usage Behavior and Navigational Structure mining

[1]*Mr.Sujeet Kumar Tiwari*, [2]*Mrs. Neelu Tiwari*
[1,2]Computer Science and Engineering
Lakshmi Narain College of Technology Jabalpur(M.P)
Email id-sujeet.tiwari08@gmail.com
Email id-neel_31@rediffmail.com

Abstract: - The Web can be defined as a depot of varied range of information present in the form of millions of websites dispersed around us. Often users find it difficult to locate the appropriate information fulfilling their needs with the abundant number of websites in the Web. Hence multiple research work have been conducted in the field of Web Mining so as to present any information matching the user's needs. The need to fetch the navigation pattern for the website helps in aiding the owners business.

Keywords:- Structure Mining, Logs usage,

1. INTRODUCTION

"If the web site would be a car, hyperlinks would be the engine, because without them, we are not going anywhere"

Our focus is on the data mining solutions for the WWW, and that shall be done by web usage behavior, and analysis of the link associated, showing the different URL's. Navigational patterns can be used for different purposes, they can show how users of the web site behave in general or extract different (groups of) users' behaviors in order to adjust the web site to the need of a specific users group. The above pinpoints that data mining gives possibilities. However, the question is how these should be used.

Having many web sites to chose from, user can ditch one web site if it is to hard to browse, designers are challenged to fulfill the goal of users' navigational requirements.

The problems that need to be solved are to identify where software engineers can find measures of the website and how to use them.

Facing the fact that the users are the ones who evaluate the website, the designer should strive to validate design assumptions with the actual usage of the website. This type of assessment is only possible after "releasing" the website, since external quality of any software can be measured only starting with the moment when the software product is being used. This leads to the problem of retrieving useful information from the usage logs and to make a relevant interpretation of it.

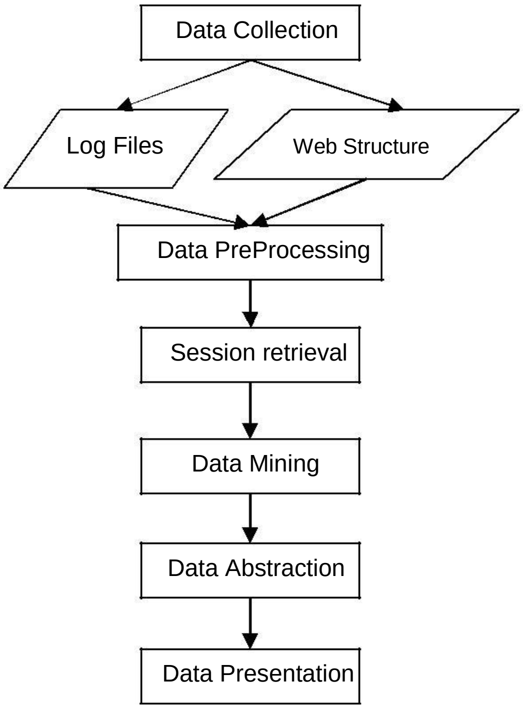
2. METHODOLOGY

The server log is the chief source of data for analyzing user course-plotting patterns which contains the information collected by Web servers. Now several kinds of access pattern mining can be done depending on the analyst needs. One of the ways suggested to get the server logs processed is by utilizing web mining techniques so as to gain information about web usage. Hence promoting varied web mining techniques to assist web administrator in indulging himself into user's site usage patterns .Upon analyzing the server access logs and user registration data one can get useful information about a better structure of a website so as to create effectual presence for the organization. Also such data helps business organizations to get the customer's cross-marketing strategies across products and efficacy of promotional campaigns amongst all other things.

The experiment under study shall go through major step listed below:

1. Data Collection
2. Data Cleaning
3. Data Mining
4. Data Abstraction
5. Data Presentatio

3. EXPERIMENTAL FLOWCHART



4. HYPERLINKS MEASUREMENT

Data used for the hyperlinks measurements can be divided in three types, these are:

− content,

− log files, and

− web structure

Maintaining website requires logging different measures during its existence. The web site structure also allows for conducting several measures. Web administrators in order to perform fixes or web site updates measure traffic on the server, knowing when the traffic is low; they can perform their tasks with low probability to angry the web site users. Traffic measures are possible because of the recorded log files. Log files contain all requests send to the web site server. Request is "questions" to the server for different resources.

Since there is possibility to log time of the request, time measure is also available for the web site measurement activities.

1 THE WEBSITES TRAFFIC LOG

"A Web server when properly configured, can record every click that users make on a Web site". For each click in the visit path, the server adds to the log file information about user request.

The logs collect data on the server in the files of specific format. Measures hold information about web site usage by recording how users visit the web site and how active they are.

Depending on the log format structure, different data is stored. Usually logs contain data such as: client's IP address, URL of the page requested, time when the request was send to the server etc. This data is used later as the basis of usage behavior discovery.

Very important fact is that logs can contain additional information which is navigational data. Navigational data is information extracted from the pre-processed web logs. The last gives knowledge of how the web site was used during a specific time interval specified by pre-processed logs. The knowledge about the web site usage is retrieved by forming statistics of viewed web pages, errors displayed, time spend on the web site by summarizing intervals between user requests. One should keep in mind that if using log files for any assessment purpose, we base the result on the sample of data, depending on the sample size the results can differ.

5. DATA PREPROCESSING

On completion of preprocessing, pattern mining can be performed in order to find the useful patterns that can be used to improve the sites. Pattern mining will return several findings such as:-

5.1. GENERAL STATISTICS

General statistics are the summary of the whole log file. Usually it provides the Total Hits, Page Views, and Total Visitor.

5.2. ACCESS STATISTICS

Access statistics provides information such as Most Popular Access Page and Most Downloaded Files.

5.3. VISITORS INFORMATION

Visitor's information will provide the information such as the most active country which accesses the website.

5.5. ERROR

Error is important for the system administrator's website in order to improve the site as well as to reduce the error such as "404 file not found

5.4. REFERRER

Referrer will provide information such as the most used search engines and phrases, and keyword used.

Field number	Example data	Field name	Meaning
1	209.240.221.71	Remote host	Remote host, IP or DNS host name
2	-	rfc931	Remote log identification name, in most cases filed take value of "-"
3	user sdftre	authuser	Authentication id of the user, can also be a password required to access
4	Thu July 1712:38:091999	date	Date time (in Greenwich mean time format)
5	"GET" index.hml/products.htm	request	Request or transaction
6	200	status	HTTP status code returned to the client, 200 equals to success
7	3234	bytes	Size of the document or the transaction transferred to the client

The table above shows the multiple fields that can be read from the any web log.

6. USAGE OF DATA MINING TECHNIQUES

There are multiple knowledge discovery techniques for the website improvement that has been conducted by other researchers, some are listed below:

6.1. NAVIGATIONALSTRUCTURE MINING

Chui and Li proposed a hyperlink frequent items extraction algorithm which allows the automatic extraction of navigational structures without performing textual analysis. The process of extracting structure from web site pages was called by them structure mining. Results of their study revealed that organization of the navigation structure can be used as predictor for the user performance in using the web site.

6.2. NEGATIVE ASSOCIATION MINING FOR THE WEBSITE

Pilarczyk presents the usage of the negative association rules mining for the web pages on the selected web site. His method is based on the mining of association rules derived from HTTP server logs. In his research Pilarczyk discovered both positive and negative association between pages and tried to evaluate the hyperlink usability.

7. DISCUSSION

The usability of estimating utility of the hyperlinks by using data mining techniques depends on the web expert willingness for the changes in the web site structure. During the meeting, in average, the web expert validated the selected web pages lower than the method. This can indicate criticism on the pages design from the web expert point of view

8. CONCLUSION

After validating the hyperlink utility values with the web expert, the noticeable fact was that the web expert was careful for accepting the highly negative values of the hyperlinks. This can be due to th possible losses in the web site's connectivity. Although the validation of the method does not show strong relationship, one should remember that there is always gap between user behavior and the web expert visio

9. REFERENCES

[1]. Goker, A., He, D. and Harper, D. (2002). Combining evidence for automatic web session identification, Information Processing and Management 38(5): 727– 742.

[2]. Good, N., Schafer, B., Konstan, J., Borchers, A., Sarwar, B., Herlocker, J. and Riedl, J. (1999). Combining collaborative filtering with personal agents for better recommendations, Proceedings of the 16th National Conference on Artificial Intelligence, American Association for Artificial Intelligence, Menlo Park, CA, pp. 439–446.

[3] Jiang, X.-M., Song, W.-G. and Zeng, H.-J. (2005). Applying associative relationship on the click through data to improve web search, Advances in Information Retrieval, Vol. 3408/2005, Springer, pp. 475–48