Enhancing Web Navigation Usability By Validating Usage Pattern

B.Venkatesh
Assistant Professor
Dept of Information Technology
Anjalai Ammal Mahalingam Engineering College
THIRUVARUR-Dist

R.Dharani, S.Mhonisha, B.Srileka
Dept of Information Technology
Anjalai Ammal Mahalingam Engineering College
THIRUVARUR-Dist

***ABSTRACT*:** *For the past few years, the Internet users are facing the problem of navigating through the web pages. Hence, we try to find the resolve the problem of users by understanding their perspective in their surfing. The knowledge of the users in moving towards their desired product is understood by analyzing the web server log using usage mining. Usage mining which is one of the concept of data mining which helps in the identification of user pattern and rectifying their problem in usage of the web sites or any web applications. This project depicts the idea of enhancing the web navigation by using the usage mining. The Enhancement is done by comparing the actual pattern and anticipated pattern with the threshold value to identify the deviation of the user activities to find their product .The input for this project is from the user log from which we can identify the user's navigation in the project.*

Keywords— Actual pattern, Anticipated pattern, Access log, sitemap

I. INTRODUCTION

As the internet and mobile has become the vital part in our daily life and all the data and work has become digitized and all users has turned to application that can work with the single touch. So, the enhancement of the web navigation is done to make the user of the system to easily navigate to find their desired item. The motive for this enhancement is that the user's idea of moving from one page to another page may mismatch with the designers way of design to navigate in the website. Hence, the user's idea of moving through the web pages is understood by making analysis in web log and making arrangements based on their needs.

The input for the project is the user log the contains the user activity such as users entry from which IP address, their navigation through the system, time of login, path traced by them etc., These entries in the log data are collected and certain manipulations were made to find the deviation of the users path with the designers perspective by comparing with the threshold value. Two calculations were made which are temporal deviation calculation and logical deviation calculation.

For this enhancement the algorithm known as Trial tree algorithm is used which helps the developer to find the deviation of the path of the user from the developer's point of view. This algorithm works by making the null root node and the branching them based on the user movement in using the system, the common path traced by the user are marked as the parent node and then any change in the path is marked in the proceeding nodes. From this we can identify the path were the user get struck and then the developer can find the solution to it to make the user correctly navigate to the right location of the product.

II. RELATED WORK

A. Anticipated Pattern

Analyzing the web resources find possible paths and patterns. Sitemap of website can be extracted from web site and create a pattern from sitemap. The Sitemaps protocol allows a webmaster to inform search engines about URLs on a website that are available for crawling. A Sitemap is an XML file that lists the URLs for a site. It allows webmasters to include additional information about each URL when it was last updated, how often it changes, and how important it is in relation to other URLs in the site. This allows search engines to crawl the site more intelligently.

B. Actual Usage Pattern

Web server logs are the data source. Each entry in a log contains the IP address of the originating host, the timestamp, the requested Web page, the referrer, the user agent and other data. Typically, the raw data need to be preprocessed and converted into user sessions and transactions to extract usage patterns. Web server log repositories are great source of knowledge, which keeps the record of web usage patterns of different web users. The actual usage patterns can be extracted from Web server logs routinely recorded for operational websites by first processing the log data to identify users, user

sessions, and user task-oriented transactions, and then applying an usage mining algorithm to discover patterns among actual usage paths. Sequence of related operation rules can be specified for a series of transitions. This model specifies both the path and the benchmark interactive time for some specific states.

III. ARCHITECTURE OF THE METHOD

The entire project can be divided into the following three modules:

1. Anticipated Pattern Design.
2. Actual Usage Pattern
3. Deviation Computation

A. *Anticipated Pattern Design*

Anticipated design pattern are expected usage pattern of a website. It can be developed using resources available in the website. Sitemap gives all available URL of the website. Based on the availability of resources in the web site, design the patterns. Analyzing the web resources find possible paths, Pattern design, Sitemap of website can be extracted from web site and create a pattern from sitemap. The Sitemaps protocol allows a webmaster to inform search engines about URLs on a website that are available for crawling. A Sitemap is an XML file that lists the URLs for a site. It allows webmasters to include additional information about each URL, when it was last updated, how often it changes, and how important it is in relation to other URLs in the site. This allows search engines to crawl the site more intelligently.

B. *Actual Usage Pattern*

Each entry in a log contains the IP address of the originating host, the timestamp, the requested Web page, the referrer, the user agent and other data. Typically, the raw data need to be preprocessed and converted into user sessions and transactions to extract usage patterns. Web usage mining or web log mining process can be regarded as a five phase process. Web mining is an appliance of data mining techniques to large web log data repositories. Web mining is an appliance of data mining techniques to large web log data repositories. 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. 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 user's with the web.

C. *Deviation Computation*

The actual user's navigation trails are extracted from the aggregated trail tree is compared against corresponding anticipated models automatically. This comparison will yield a set of deviations between actual and anticipated pattern. Some common problems of actual user's interaction with the web application by focusing on deviations that occur frequently are identified. Combined with expertise in product internal and contextual information, results can also help identify the root causes of some usability problems existing in the web design. Based on logical choices made and time spent by user's at each page, logical deviation will be calculated. When the path choice anticipated by the model available in the pattern but not selected, a single deviation is counted. Sum up all the above deviations over all the selected user transactions for each page.

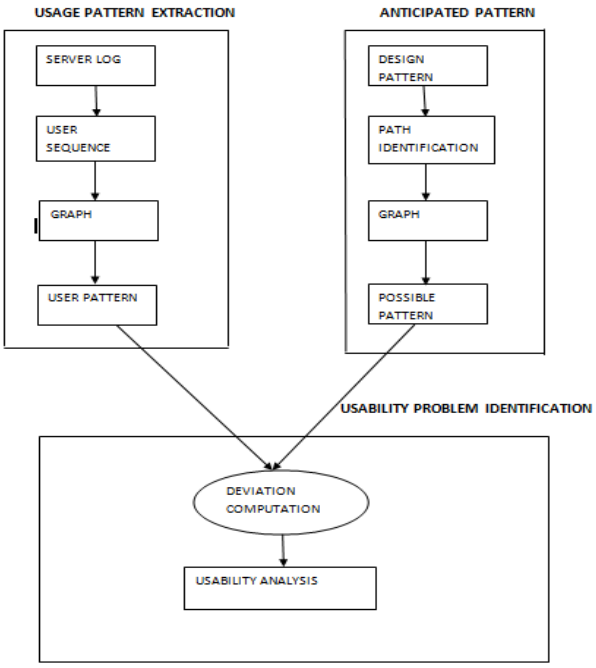

Fig. 1. System Architecture

IV. PATTERN EXTRACTION

A. *Data Preparation and Preprocessing*

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. Figure 3.1 shows the processing log file. 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 user's 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.

The data present in the log file cannot be used as it is for the mining process. Therefore the contents of the log file should be cleaned in this preprocessing step. The unwanted data are removed and a minimized lo file is obtained.

B. 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. Figure 3.2 depicts the discovery of pattern. After the conversion of the data in the log file into a formatted data the pattern discovery process is under gone. With the existing data of the log files many useful patterns are discovered either with user id's, session details, time outs etc.

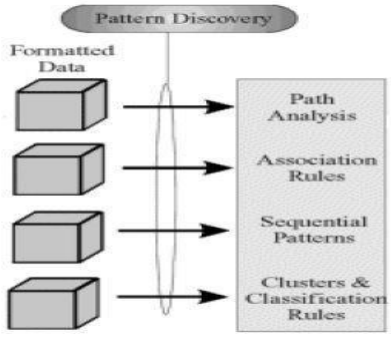

Fig.2. Pattern Discovery

V. WEB USAGE PATTERN ANALYSIS

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.

Dilip Singh Sisodia and Shrish Verma proposed a web usage pattern analysis through web logs. The log files are preprocessed and converted into required formats so web usage mining techniques can apply on these web logs. It is the process of discovering useful patterns from the web server log files. 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.

R. Cooley, M Deshpande, and J. Srivastava proposed web usage mining process.web usage mining is the application of data mining techniques to discover usage patterns from web data, in order to understand and better serve the needs of web-based applications.web usage mining has seen a rapid increase in interest, from both the research and practice communities.

A. CONTENTS OF A LOG FILE:

L.K. Joshila Grace, V.Maheswari, Dhinaharan Nagamalai analyzing the web logs and web user in web mining. Log files are files that list the actions that have been occurred. These log files reside in the web server. Computers that deliver the web pages are called as web servers. The Web server stores all of the files necessary to display the Web pages on the user's computer. All the individual web pages combines together to form the completeness of a Web site.

The Log files in different web servers maintain different types of information. The basic information present in the log file are:

1. User name: This identifies who had visited the web site. The identification of the user mostly would be the IP address that is assigned by the Internet Service provider (ISP). This may be a temporary address that has been assigned. Therefore here the unique identification of the user is lagging. In some web sites the user identification is made by getting the user profile and allows them to access the web site by using a user name and password. In this kind of access the user is being identified uniquely so that the revisit of the user can also be identified. International Journal of Network Security & Its Applications.
2. Visiting Path: The path taken by the user while visiting the web site. This may be by using the URL directly or by clicking on a link or trough a search engine.
3. Path Traversed: This identifies the path taken by the user within the web site using the various links.
4. Time stamp: The time spent by the user in each web page while surfing through the web site. This is identified as the session.
5. Page last visited: The page that was visited by the user before he or she leaves the web site.
6. Success rate: The success rate of the web site can be determined by the number of downloads made and the number copying activity under gone by the user. If any purchase of things or software made, this would also add up the success rate.
7. User Agent: This is nothing but the browser from where the user sends the request to the web server. It's just a string describing the type and version of browser software being used.
8. URL: The resource accessed by the user. It may be an HTML page, a CGI program, or a script.
9. Request type: The method used for information transfer is noted. The methods like GET, POST.

These are the contents present in the log file. This log file details are used in case of web usage mining process. According to web usage mining it mines the highly utilized web site. Table 3.1 shows the content of log file. The utilization would be the frequently visited web site or the web site being utilized for longer time duration. Therefore the quantitative usage of the web site can be analyzed if the log file is analyzed.

B. LOCATION OF LOG FILE

A Web log is a file to which the Web server writes information each time a user requests a web site from that particular server. A log file can be located in three different places. They are Web Servers, Web proxy Servers and Client browsers.

1) Web server log file

The log file that resides in the web server notes the activity of the client who accesses the web server for a web site through the browser. In the server which collects the personal information of the user must have a secured transfer.

2) Web proxy server log files

A Proxy server is said to be an intermediate server that exist between the client and the Web server. Therefore if the Web server gets a request of the client via the proxy server then the entries to the log file will be the information of the proxy server and not of the original user. These web proxy servers maintain a separate log file for gathering the information of the user.

3) Client browser log files

This kind of log files can be made to reside in the client's browser window itself. Special types of software exist which can be downloaded by the user to their browser window. Even though the log file is present in the client's browser window the entries to the log file is done only by the Web server.

4) Access log files

The server access log records all requests that are processed by the server. The location and content of the access log are controlled by the Custom Log directive. The Custom Log directive is used to log requests to the server. A log format is specified, and the logging can optionally be made conditional on request characteristics using environment variables. The Log Format directive can be used to simplify the selection of the contents of the logs. This section describes how to configure the server to record information in the access log.

5) Sitemap

Sitemap is used for the graphical representation of the architecture of a web site. The Sitemaps protocol allows a webmaster to inform search engines about URLs on a website that are available for crawling. A Sitemap is an XML file that lists the URLs for a site. It allows webmasters to include additional information about each URL: when it was last updated, how often it changes, and how important it is in relation to other URLs in the site. This allows search engines to crawl the site more intelligently. Sitemaps are a URL inclusion protocol and complement exclusion protocol. The Potts model is frequently used to describe the behavior of image classes, since it allows to incorporate contextual information linking neighboring pixels in a simple way. Data are extracted from the web server log. A mining process performed on the data collected from the server log. A preprocessing involves in the data is transformed into different states.

Sitemap are particularly beneficial on websites where Some areas of the website are not available through the browser interface. Webmasters use rich Ajax or Flash content that is not normally processed by search engines. The site is very large and there is a chance for the web crawlers to overlook some of the new or recently updated content. When websites have a huge amount of pages that are isolated or not well linked together when a website has few external links.

ACKNOWLEDGEMENT

The authors would like to all the persons those who helped us in doing this project.

REFERENCE

[1] R.Geng and J.Tian, "Improving Web Navigation Usability by Comparing Actual and Anticipated Usage" IEEE Trans. In Human-machine systems, vol.45,NO.1.

[2] J.Srivastava, R.Cooley, M.Deshpande and Pang-Ning "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data" SIGKDD Explorations, volume 1, issue 1.

[3] J.Grace, V. Maheshwari and D.Nagamalai "ANALYSIS OF WEB LOGS AND WEB USER IN WEB MINING" International Journal of Network Security & Its Applications (IJNSA), Vol.3, No.1, January 2011.

[4] K. R. Suneetha and R. Krishnamoorthi "Identifying User Behavior by Analyzing Web Server Access Log File" IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.4, April 2009.

[5] R. Cooley, B. Mobasher, J.Srivastava "Web Mining: Information and pattern discovery on the World Wide Web"

[6] T. Arce, P. E. Rom´an, J. D. Vel ´ asquez, and V. Parada, "Identifying web sessions with simulated annealing," Expert Syst. Appl., vol. 41, no. 4, pp. 1593–1600, 2014.

[7] T.Carta, F.Paterno, and V.Figuerêdo de Santana "Web Usability Probe: A Tool for Supporting Remote Usability Evaluation of Web Sites" P. Campos et al. (Eds.): INTERACT 2011, Part IV, LNCS 6949, pp. 349–357, 2011.

[8] M. Alphy and A. Sharma "Study on online community user motif using web usage mining" ScieTech 2016 Journal of Physics: Conference Series 710 (2016) 012015.

[9] Laila Paganelli and Fabio Paterno "Tools for remote usability evaluation of Web applications through browser logs and task models" Behavior Research Methods, Instruments, & Computers 2003, 35 (3), 369-378.

[10] S. S. Patil and H.P.Khandagale "Survey Paper on Enhancing Web Navigation Usability Using Web Usage Mining Techniques" International Journal of Modern Trends in Engineering and Research (IJMTER) Volume 03, Issue 02, [February – 2016].

[11] Martin F. Arlitt and Carey L. Wfizpnson "Internet Web Servers: Workload Characterization and Performance Implications" IEEE/ACM Transaction on networking.Vol.5, No.5.October 1997.

[12] F. E. Ritter, A. R. Freed, and O. L. Haskett, "Discovering user information needs: The case of university department web sites," *ACM Interactions*, vol. 12, no. 5, pp. 19–27, 2005.

[13] M. Rauterberg, "AMME: An automatic mental model evaluation to analyze user behavior traced in a finite, discrete state space," *Ergonomics*, vol. 36, no.11, pp.1369–1380, 1993.

[14] D. Peebles and A. L. Cox, "Modeling interactive behaviour with a rational cognitive architecture," in *Human Computer Interaction: Concepts, Methodologies, Tools, and Applications*, C. S. Ang and P. Zaphiris, Eds. Hershey, PA, USA: Inf. Sci. Ref., 2008, pp. 1154–1172.

[15] J. H. Morgan, C.-Y. Cheng, C. Pike, and F. E. Ritter, "A design, tests and Considerations for improving keystroke and mouse loggers," *Interacting Compute.*, vol. 25, no. 3, pp. 242–258, 2013.

[16] A. McDonald and R. Welland, "Web engineering in practice," in
Proc.10th Int. World Wide Web Conf., May 2001, pp. 21–
30.

[17] D. E. Kieras and D. E. Meyer, "An overview of the EPIC architecture for Cognition and performance with application to human- computer interaction," *Human-Comput. Interaction*, vol. 12, no. 4, pp.
391–438,
1997.

[18] M. D. Byrne, "ACT-R/PM and menu selection: Applying a cognitive architecture to HCI," *Int. J. Human-Comput. Stud.*, vol. 55, no. 1, pp. 41–84, 2001.