Mining Web Logs File to Provide Unimpaired Websites Using Fault Analyzing Technique

C.Kanagalakshmi
Department of Computer Science and Engineering
A.V.C. College of Engineering

K. TamilSelvan
Department of Computer Science and Engineering
A.V.C. College of Engineering

***Abstract* - Providing a flawless website is the only thing to increase the customer redundancy rate in most of the commercial multipage websites. The unavailability of resource or page and fault in process may make the user to switch over to other websites. This may cause loss of visitor rate for the specific website. In order to maintain such kind of websites, the web administrator has to go through that website by testing hundreds of pages. To do this the better method is to analyze the log files of the website to identify faults and error and also finding out user's usability problem and usage pattern is one of the big tasks. In order to overcome the above needs a new methodology is adopted to run along with the web sites and to analyze web usability errors.**

***Index Terms*- User navigation, web sites, access tracking, interactive visualization, web usage mining**

I. Introduction

The World Wide Web (WWW) continues to grow at an astounding rate in both the sheer volume of traffic and the size and complexity of Web sites. The complexity oft asks such as Web site design, Web server design, and of simply navigating through a Web site has increased along with this growth. An important input to these design tasks is analysis of how a Web site is being used. Usage analysis includes straightforward statistics, such as page access frequency, as well as more sophisticated forms of analysis, such as finding the common traversal paths through a Web site. Usage information can be used to restructure a Website in order to better serve the needs of users of a site. Long convoluted traversal paths or low usage of a page with important site information could suggest that the site links and information are not laid out in an intuitive manner. The design of a physical data layout or caching scheme for a distributed or parallel Web Server can be enhanced by knowledge of how users typically navigate through the site. Usage information can also be used to directly aide site navigation by providing a list of "popular" destinations from a particular Web page.

Web Usage Mining is the application of data mining techniques to large Web data repositories in order to produce results that can be used in the design tasks mentioned above. Some of the data mining algorithms that are commonly used in Web Usage Mining are association rule generation, sequential pattern generation, and clustering. Association Rule mining techniques discover unordered correlations between items found in a database of transactions. In the context of Web Usage Mining a transaction is a group of Web page accesses, with an item being a single page access. The percentages reported in the examples above are referred to as *confidence*. Confidence is the number of transactions containing all of the items in a rule, divided by the number of transactions containing the rule.

It is proposed a data mining model that captures the user navigation behaviour patterns. The user navigation sessions are modelled as a hypertext probabilistic grammar whose higher probability strings correspond to the user's preferred trails. An algorithm to efficiently mine such trails is given. We make use of the N-gram model which assumes that the last N pages browsed affect the probability of the next page to be visited. The model is based on the theory of probabilistic grammars providing it with a sound theoretical foundation for future enhancements.

Index Terms -List User navigation, web sites, access tra

Moreover we propose the use of entropy as an estimator of the grammars statistical properties Extensive experiments were conducted and the results show that the algorithm runs in linear time the grammars entropy is a good estimator of the number of mined trails and the real data rules confirm the effectiveness of the model.

Data Mining and Knowledge Discovery is an active research discipline involving the study of techniques which search for patterns in large collections of data. Meanwhile the explosive growth of the World Wide Web known as the web mining in recent years has turned it into the largest source of available online data mining. Therefore mining the application of data mining techniques to the web mining called web data mining was the natural subsequent step and it is now the focus of an increasing number of researchers. In web data mining there are currently three main research directions mining for information mining the web link structure and mining for user behaviour patterns Mining for information focuses on the development of techniques to assist users in processing the large amounts of data they face during navigation and to help them and the information they are looking for see for example, Mining the link structure aims at developing techniques to take advantage of the collective judgment of web page quality in the form of hyperlinks which can be viewed as a mechanism of implicit endorsement see.

The aim is to identify for a given subject the authoritative and the hub pages Authoritative pages are those which were conferred authority by the existing links to it and hubs are pages that contain a collection of links to related authorities Finally, the other research direction, which is being followed by an increasing number of researchers is mining for user navigation patterns This research focuses on techniques which study the user behaviour when navigating within a web site. Understanding the visitor's navigation preferences is an essential step in improving the quality of electronic commerce services. In fact the understanding of the most likely access patterns of users allows the service provider to customize and adapt the site's interface for the individual user and to improve the site's static structure within the underlying hypertext system.

II. Proposed Models

A. Web Client Session And Access Management

It is the module provided for the test clients. The clients are allowed to browse access and download files from the web pages. The unique session will be created for every single client who accessing the web page.This module fully concentrated on providing usability to the clients and records their action on web page. An important aspect of correctly managing state information through session IDs relates directly to authentication processes. While it is possible to insist that a client using an organizations web application provide authentication information for each "restricted" page or data submission, it would soon become tedious and untenable. Thus session IDs are not only used to follow clients throughout the web application, they are also used to uniquely identify an authenticated user thereby indirectly regulating access to site content or information. The users will have two options they can create own account or they may access the server's guest access.

B. Client User's Access and Session recording Process

In order to effectively manage the web pages, it is necessary to get feedback about the activity and performance of the website and if any problem is occurring while the user interacting with the websites. It is must to provide log capabilities for a professional website. In our website a specific error log mechanism is implemented for logging every user activity and the session activity of the client in our website from the initial request to the URL mapping to the final resolution of the connection. It includes user activity and errors occurred during entire session. This error log record will be holding the valuable information like problems encountered while on user request, pages accessed frequently, time taken for the user to switch over to the next page, time when the page is accessed, client's information like browser, user ID, clients host address, hostname. These all information will be recorded for each and every request and access per page.

C. Culpability Detection and Analysing Of Usage Patterns

The volumes of click stream and user data collected by Web-based organizations in their daily operations have reached astronomical proportions. Analysing such data can help these organizations determine the life-time value of clients, design cross-marketing strategies across products and services, evaluate the effectiveness of pro-motional campaigns, optimize the functionality of Web-based applications, provide more personalized content to visitors, and find the most effective logical structure for their Web space. This type of analysis involves the automatic discovery of meaningful patterns and relationships from a large collection of primarily semi-structured data, often stored in Web and applications server access logs, as well as in related operational data sources.The modification allows putting thedata in correct order by using User-ID and time-stamp sort.The Patterns reorganizations system is to study the client behavior from the web data towardsproviding quality service. Client Behavior Pattern Recognition System is described that reflects the fivehierarchical structures in the recognition system.It consists of data mining and online analysis process, used for system log data analysis by logmining online analysis technique. Generating interest at the framework level and variousapplication parameters of the framework prompted to present the following architecture on which algorithm would beapplied. The following system based on web log mining architecture for proposing the basic interactiveelements for miner.

D. Generation of The Report File Usage Pattern

The overall Web us-age mining process can be divided into three inter-dependent stages: data collection and pre-processing, pattern discovery, and pattern analysis. In the pre-processing stage, the click stream data is cleaned and partitioned into a set of user transactions representing the activities of each user during different visits to the site. Other sources of knowledge such as the site con-tent or structure, as well as semantic domain knowledge from site ontologiessuch as product catalogs or concept hierarchies, may also be used in pre-processing or to enhance user transaction data. In the pattern discovery stage, statistical, database, and machine learning operations are per-formed to obtain hidden patterns reflecting the typical behavior of users, as well as summary statistics on Web resources, sessions, and users.Identification of page views is heavily dependent on the intra-page structure of the site, as well as on the page contents and the underlying site do-main knowledge. Recall that, conceptually, each page view can be viewed as a collection of Web objects or resources representing a specific "user event," e.g., clicking on a link, viewing a product page, adding a product to the shopping cart.

E. Web Admin Control Panel for Report Analysis

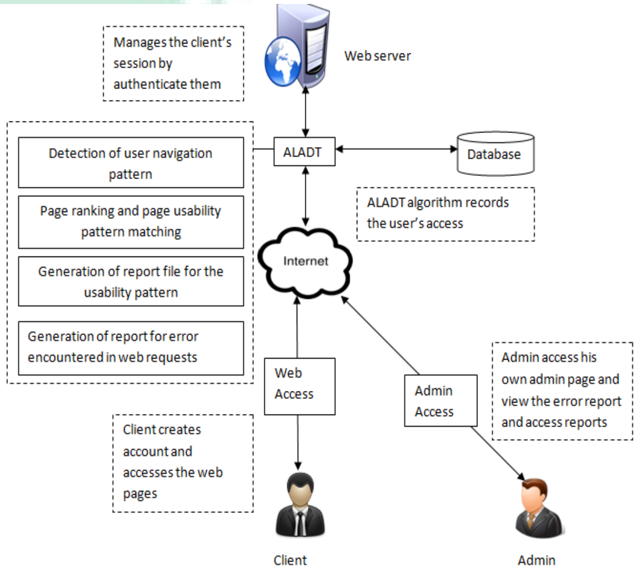
In collaborative filtering applications which rely on the profiles of similar users to make recommendations to the current user, weights may be based on user ratings of items. In most Web usage mining tasks the weights are either binary, representing the existence or non-existence of a page view in the transaction; or they can be a function of the duration of the page view in the user's session. In the case of time durations, it should be noted that usually the time spent by a user on the last page view in the session is not available. One commonly used option is to set the weight for the last page-view to be the mean time duration for the page taken across all sessions in which the page view does not occur as the last one. In practice, it is common to use a normalized value of page duration instead of raw time duration in order to account for user variance.

IV. Algorithm

```
repeat
foreach record begin
for j ← 1 to n
Increment support of value(Ej) by bj .
end
Sort pages by support.
P := Page with highest support (break ties at
random).
if support(P)  Sb begin
Add hP, support(P)i to list of recommended pages.
foreach record begin
for k := 1 to n begin
if value( Ek) = P
Set  $E_k$; $_{Ek+1}$ ,E $_n$ to null;
end
end
end
until (support(P) <Sb
```

V. Conclusion

III. Architecture Diagram


Manages the client's session by authenticate them
Web server
Detection of user navigation pattern
Page ranking and page usability pattern matching
Generation of report file for the usability pattern
Generation of report for error encountered in web requests
ALADT
Database
ALADT algorithm records the user's access
Internet
Web Access
Admin Access
Admin access his own admin page and view the error report and access reports
Client creates account and accesses the web pages
Client
Admin

Web usage mining, the analysis of user navigation paths through web sites, is a common technique for evaluating site designs or adaptive hypermedia techniques. However, often it is hard to relate aggregated clusters or measures tactual user navigation behaviour. By contrast, basic graph based visualizations of user navigation paths are easier to interpret, but it is difficult to find effective views that convey all the required information. In this paper it has been present the Navigation pattern analyser, a web usage analysis tool that combines the two approaches. The Navigation pattern makes use of the rich data set that is collected by the Scone proxy-based web enhancement framework and facilitates dynamic selection of the data and interactive exploration with various layout mechanisms. Several aggregated measures can be calculated and exported to statistical and web mining packages.

VI. Future Work

An interesting problem for future research is that in websiteswithout a clear separation of content and navigation, it can be hardto differentiate between visitors who backtrack because they arebrowsing a set of target pages, and visitors who backtrack because they are searching for a single target page. While we have proposed using a time threshold to distinguish between the two activities, it will be interesting to explore if there are better approaches to solvethis problem. Our key insight is that visitors will backtrack if they do not find information where they expect it. The point from where they backtracks the expected locations for the page.

REFERENCES

[1] A. Agarwal and M. Prabaker, “Building on the usability study: Two explorationson how to better understand an interface,” in *Human-ComputerInteraction. New Trends*, J. Jacko, Ed. New York, NY, USA: Springer,2009, pp. 385–394.

[2] J. R. Anderson, D. Bothell, M. D. Byrne, S. Douglass, C. Lebiere, andY. Qin, “An integrated theory of the mind,” *Psychol. Rev.*, vol. 111,pp. 1036–1060, 2004.

[3] T. Arce, P. E. Rom´an, J. D. Vel´asquez, and V. Parada, “Identifying websessions with simulated annealing,” *Expert Syst. Appl.*, vol. 41, no. 4,pp. 1593–1600, 2014.

[4] M. F. Arlitt and C. L. Williamson, “Internet Web servers: Workloadcharacterization and performance implications,” *IEEE/ACMTrans. Netw.*,vol. 5, no. 5, pp. 631–645, Oct. 1997.

[5] C. M. Barnum and S. Dragga, *Usability Testing and Research*. WhitePlains, NY, USA: Longman, Oct. 2001.

[6] B. Beizer, *Software Testing Technique*. Boston, MA, USA: Int. ThomsonComput. Press, 1990.

[7] J. L. Belden, R. Grayson, and J. Barnes, “Defining and testing EMRusability:” *Healthcare Information and Management SystemsSociety*, Chicago, IL, USA, Tech. Rep., (2009).

[8] M. C. Burton and J. B. Walther, “The value of Web log data in use-baseddesign and testing,” *J. Comput.-Mediated Commun.*, vol. 6, no. 3, p. 0,2001.

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

[10] T. Carta, F. Patern`o, and V. F. D. Santana, “Web usability probe: A tool forsupporting remote usability evaluation of web sites,” in *Human-ComputerInteraction—INTERACT 2011*. New York, NY, USA: Springer, 2011,