Enhancing the Website Structure by Reconciling Website

Joy Shalom Sona[1], Prof. Asha Ambhaikar[2]

[1](Research Scholar,Dept. of Comp. Science and EngineeringRCET, Bhilai, India)
[2](Dept. of computer science and engineering RCET, Bhilai, India)

Abstract —In this Modern world www grows tremendously. It increases the complexity of web applications and web navigation. Existing website systems are not easier for user to extract information and having some shortcomings. To enhance these shortcomings we propose a new reconciling website system. Recommendations play an important role towards this direction. Our Recommendation is based on user Browsing patterns. Our approach presents a comprehensive overview of web mining methods and techniques used for the evaluation of reconciling systems to achieve better web navigation efficiency in order to improve the efficiency of web site. It integrates and coordinates among different reasons for making recommendations including frequency of access, and patterns of access by user to the web site. It is new way to increase the efficiency of web site system using web mining techniques We are not argue the structure or content of the web site but we recommended to web site developer. Our proposed techniques are achieved better web navigation efficiency and it is highly effective from existing one.*

Keywords—Web Structure Mining, Web Content Mining, Web Usage Mining, Reconciling Website System, Browsing Efficiency, Recommendation

I.   INTRODUCTION

The most of the people browsing the internet for retrieving information. But most of the time, they gets lots of insignificant and irrelevant document even after navigating several links. Factors for web designers when considering the design of a new website include the attractiveness of the design, an effective structure to the web page to deliver information quickly, and user satisfaction among a growing and diverse set of users faced with ever increasing web contents. However, with the development of more and more web-based technologies and the growth in web content, the structure of a website becomes more complex and web navigation becomes a critical issue to both web designers and users.

1.1 Web Mining Overview

Web mining is an application of the data mining techniques to automatically discover and extract knowledge from the Web. According to Kosala et al [2], Web mining consists of the following tasks:

1.1.1   *Resource finding*
The task of retrieving intended Web documents.
1.1.2   *Information selection and pre-processing*
Automatically selecting and pre-processing specific information from retrieved Web resources.
1.1.3   *Generalization*
Automatically discovers general patterns at individual Web sites as well as across multiple sites.
1.1.4   *Analysis*
Validation and/or interpretation of the mined patterns.

There are three areas of Web mining according to the usage of the Web data used as input in the data mining process, namely, Web Content Mining (WCM), Web Usage Mining (WUM) and Web Structure Mining (WSM).
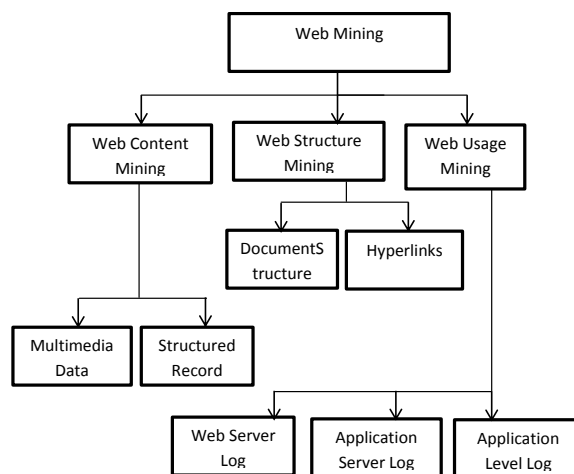
Fig.1 Web Mining Classification

1.2 Web Content Mining (WCM)

Web Content Mining is the process of extracting useful information from the contents of web documents. The web documents may consists of text, images, audio, video or structured records like tables and lists. Mining can be applied on the web documents as well the results pages produced from a search engine. There are two types of approach in content mining called agent based approach and database based approach. The agent based approach concentrate on searching relevant information using the characteristics of a particular domain to interpret and organize the collected information. The database approach is used for retrieving the semi-structure data from the web.

1.3 Web Usage Mining (WUM)

Web Usage Mining is the process of extracting useful information from the secondary data derived from the interactions of the user while surfing on the Web. It extracts data stored in server access logs, referrer logs, agent logs, client-side cookies, user profile and meta data.

1.4 Web Structure Mining (WSM)

The goal of the Web Structure Mining is to generate the structural summary about the Web site and Web page. It tries to discover the link structure of the hyperlinks at the inter-document level. Based on the topology of the hyperlinks, Web Structure mining will categorize the Web pages and generate the information like similarity and relationship between different Web sites. This type of mining can be performed at the document level (intra-page) or at the hyperlink level (inter-page). It is important to understand the Web data structure for Information Retrieval.

II.   RELATED WORK

2.1 Web Mining

Web mining has emerged as a specialized field during the last few years and refers to the application of knowledge discovery techniques specifically to web data. Web content and web structure mining, respectively, refer to the analysis of the content of web pages and the structure of links between them. Web usage mining, on the other hand, is the process of applying data mining techniques to the discovery of patterns in web data [5]. Web usage mining involves four steps: user identification, data pre-processing, pattern discovery and analysis. User access patterns are models of user browsing activity. In most cases these are deduced from web server access logs. An alternative method includes client-side logging, using techniques such as cookies. This is referred to as web-log mining [4]. Mining activities help us to know the data patterns. User patterns, extracted from Web data, have been applied to a wide range of applications. Projects by Spiliopoulou and Faulstich (1998), Wu et al. (1998), Zaiane et al. (1998), Shahabi et al. (1998) have focused on Web Usage Mining in general, without extensive tailoring of the process towards one of the various sub-categories. The WebSIFT project is designed to perform Web Usage Mining from server logs in the extended NSCA format. Chen et al. (1996) introduce the concept of maximal forward reference to characterize user episodes for the mining of traversal patterns. A maximal forward reference is the sequence of pages requested by a user up to the last page before backtracking occurs during a particular server session. The Speed Tracer project [Wu et al., 1998] from IBM Watson is built upon work originally reported in Chen et al. (1996). In addition to episode identification,

SpeedTracer makes use of referrer and agent information in the preprocessing routines to identify users and server sessions in the absence of additional client side information. The Web Utilization Miner (WUM) system [Spiliopoulou and Faulstich, 1998] provides a robust mining language in order to specify characteristics of discovered frequent paths that are interesting to the analyst. Zaiane et al. (1998) have loaded Web server logs into a data cube structure in order to perform data mining as well as On-Line Analytical Processing (OLAP) activities such as roll-up and drill-down of the data. Their WebLogMiner system has been used to discover association rules, perform classification and time-series analysis. Shahabi et al. (1997) and Zarkesh et al. (1997) have one of the few Web Usage mining systems that rely on client side data collection. The client side agent sends back page request and time information to the server every time a page containing the Java applet is loaded or destroyed [5].

2.2 Adaptive Website

Users interact with a website in multiple ways, while their mental model about a particular subject can obviously differ from those of other users and the web developer. Consequently, improving the interaction between users and websites is of importance. Raskin [6] introduces various ways of quantification in measuring interface design in his book. Especially, he mentions information-theoretic efficiency, which is defined similarly to the way efficiency is defined in thermodynamics; in thermodynamics we calculate efficiency by dividing the power coming out of a process by the power going into the process. If, during a certain time interval, an electrical generator is producing 820 watts while it is driven by an engine that has an output of 1000 W, it has an efficiency 820/1000, or 0.82. Efficiency is also often expressed as a percentage; in this case, the generator has an efficiency of 82%. This calculation can be applied to calculate the information efficiency. Srikant and Yang [7] propose an algorithm to automatically find pages in a website whose location is different from where visitors expect to find them. The key insight is that visitors will backtrack if they do not find the information where they expect it: the point from where they backtrack is the expected location for the page. They also use a time threshold to distinguish whether a page is target page or not. Nakayama et al. (2000) proposes a technique that discovers the gap between website designers' expectations and users' behavior. The former are assessed by measuring the inter-page conceptual relevance and the latter by measuring the inter-page access co-occurrence. They also suggest how to apply quantitative data obtained through a multiple regression analysis that predicts hyperlink traversal frequency from page layout features. Most adaptive systems include a procedure on mining web log to understand user behaviors and patterns and to improve their website automatically and efficiently. However, none of them try to calculate the efficiency to improve the web structure. We want to apply the efficiency concept from [6] and develop the efficiency calculation function.

III.   METHODOLOGY

Our proposed techniques includes following steps:
   3.1 Mining the web architecture
   3.2 Determining user log
   3.3 Obtaining website browsing efficiency.

3.1 Mining the web architecture

A website consists of Web pages, which connect to each other through hyperlinks. The website can be modeled as a graph, G=(V, E). Vertices V={v1, v2,…vn }, where vi(i=1,2,…n) denotes a page. Edges or arcs E={eij | the hyperlink from the source page i to the destination page j}. P is the set of ordered pares (i,j) such that there is a path from i to j, where each node is visited once. R is the set of ordered pares (i,j) such that there is a route user navigate from i to j.

Fig 2: Graph Form of a Website
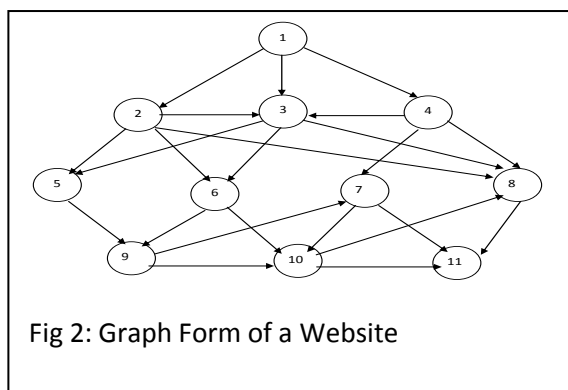
The web structure mining program proposes to grab the website structure and save it. When a designer types the IP address of a website to the program, the program automatically starts to mine the entire website architecture. Firstly, the program downloads the page and analyzes its HTML code. Next, the system seeks out the hyperlinks in the page and repeats the actions until the page does not belong to the domain the designer inputs. Finally, the program obtains the entire website architecture and saves it.

3.2 Determining user log

This involves following four task: *user identification*, *data pre-processing*, *pattern discovery* and *pattern analysis*.User access patterns are models of users' browsing activity. In most cases these are deduced from web server access logs. A web server access log is a complete review of access of a server from a client. User browsing records can be collected from three different sources: the web server log file, proxy server log file, and browser cookies. A web server log file records all user access activities on that server. As an example we note here that a log consists of the following elements: client's IP address, user id, access time, request method (get or post), URL, protocol error code, number of bytes transmitted**.**

3.3 Obtaining Website Browsing Efficiency

The browsing efficiency of a website can be calculated by Eq.(1).

$$Efficiency = (shortest\ path\ from\ start\ page\ to\ target\ page)/operating\ cost$$

User operating behavior and shortcomings of the website determine by the help of user browsing behavior patterns. For calculating the efficiency of a website, there is need's to determined the user's operating route that is start page and target page. The term operating cost refers to the number of pages visited between the begin page and the target (end) page.

IV.   CONCLUSION

In this paper we proposed a Reconciling Website System which improves the web navigation efficiency and suggests the reorganization of the web site. Reconciling Websites can make popular pages more accessible, highlight interesting links, connected related pages. Reconciling web sites can advice to a Website's developer summarizing access information and making suggestions for that particular website. These suggestion based on the user browsing behavior which increase the efficiency by reorganizing Web Structure. This gives the beneficial information to the web developer for providing easier navigation to the Website.

REFERENCES

[1] Ji-Hyun Lee, Wei-Kun Shiu: An adaptive website system to improve efficiency with web mining techniques. Advanced EngineeringInformatics 18(3): 129-142 (2004)

[2] R. Kosala, H. Blockeel, "Web Mining Research: A Survey", SIGKDD Explorations, Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining Vol. 2, No. 1 pp 1-15, 2000.

[3] Perkowitz M, Etzioni O. Adaptive web site: an AI challenge. IJCAI- 97 1997.

[4] Koutri M, Daskalaki S, Avouris N. Adaptive interaction with web site: an overview of methods and techniques. Computer Science and Information Technologies, CSIT 2002.

[5] Srivastava J, Cooley R, Deshpande M, Tan P-N. Web usage mining: discovery and applications of usage patterns from web data. ACM SIGKDD 2000.

[6] Raskin J. The human interface, first ed. Menlo, CA: Stratford Publishing, Inc.; 2000.

[7] Srikant R, Yang Y. Mining web logs to improve website organization. ACM 2001.

[8] Spiliopoulou M, Faulstich L. Wum: A web utilization miner. EDBT Workshop WebDB 98. Spain: Valencia; 1998.

[9] Wu K-L, Yu P-S, Ballman A. Speed-tracer: A web usage mining and analysis tool. IBM Systems Journal 1998;37(1).

[10] Zaiane O, Xin M, Han J. Discovering web access patterns and trends by applying olap and data mining technology on web logs. In: Advances in Digital Libraries. CA: Santa Barbara; 1998. p.19–29.

[11] Shahabi C, Zarkesh A, Adibi J, Shah V. Knowledge discovery from users web-page navigation Workshop on Research Issues in Data Engineering. England: Birmingham; 1997.

[12] Chen M-S, Park J-S, Yu P-S. Data mining for path traversal patterns in a web environment. 16th International Conference onDistributed Computing Systems 1996;385–92.

[13] Zarkesh A, Adibi J, Shahabi C, Sadri R, Shah V. Analysis and design of server informative www-sites 6th International Conference on Information and Knowledge Management. Nevada: Las Vegas; 1997.

[14] Nakayama T, Kato H, Yamane Y. Discovering the gap between web site designers' expectations and users' behavior. Computer Networks 2000;33(1-6):811–22.

[15] Nielsen//NetRatings_Global. Global Internet Index; March 2001

[16] Wang, May and Yen, Benjamin, "Web Structure Reorganization to Improve Web Navigation Efficiency" (2007). *PACIS 2007 Proceedings.* Paper 46.

[17] M. Kilfoil , A. Ghorbani , W. Xing , Z. Lei , J. Lu , J. Zhang , X. Xu," Toward An Adaptive Web: The State of the Art and Science (2003),CNSR 2003.