Reorganisation of Adaptive Websites using Web Usage Mining Techniques

Dr. Ananthi Sheshasaayee[#1], V.Vidyapriya[#2]
[#1]*Head and Associate Professor, Department of Computer Science, Quaid-E-Millath Government College for Women (A), Chennai*
[#2]*Research Scholar, Department of Computer Science, Quaid-E-Millath* Government College for Women (A), Chennai

Abstract— Web Usage Mining is that area of Web Mining which deals with the extraction of interesting knowledge from logging information produced by Web servers. The motive of mining is to find users' access models automatically and quickly from the vast Web log data, such as frequent access paths, frequent access page groups and user clustering. Through web usage mining, the server log, registration information and other relative information left by user access can be mined with the user access mode which will provide foundation for decision making of organizations. Adaptive web sites are web sites that automatically improve their organization and presentation by learning from their user access patterns. User interaction patterns may be collected directly on the website or may be mined from Web server logs. Through this paper we present the various web usage mining techniques to extract the useful and relevant information on the web for adaptive web sites.

Keywords - Web mining, Web usage mining, Web Log, Adaptive web site, Web site reorganisation.

I. INTRODUCTION

Web mining is a very interesting research topic which combines of the activated research areas: Data Mining and World Wide Web. With the huge amount of information available online, the World Wide Web is a fertile area for data mining research. The Web mining research relates to several research communities, such as database, information retrieval, and AI. The World Wide Web is a popular and interactive medium to disseminate information today. The Web is huge, diverse, and dynamic and thus raises the scalability, multimedia data, and temporal issues respectively.

Web data mining can be defined as the discovery and analysis of useful information from the web log file. Although Web mining puts down the roots deeply in data mining, it is not equivalent to data mining. The unstructured feature of Web data triggers more complexity in the process of Web mining. An exponential growth in on-line information combined with the almost unstructured web data necessitates the development of powerful yet computationally efficient web data mining tools. Web mining is the use of data mining techniques to automatically discover and extract information from Web documents and services [1]. Web mining can be classified into three areas of interest based on which part of the Web to mine: Web content mining, Web structure mining, and Web usage mining as shown in Figure 1.

In practice, the three Web mining tasks above could be used in isolation or combined in an application, especially in Web content and structure mining since the Web documents might also contain links. Web content mining is the process of extracting knowledge from the content of documents or their descriptions. Web document text mining, resource discovery based on concepts indexing or agent; based technology may also fall in this category. Web structure mining is the process of inferring knowledge from the World Wide Web organization and links between references and referents in the Web. Finally, web usage mining, also known as Web Log Mining, is the process of extracting interesting patterns in web access logs [2].

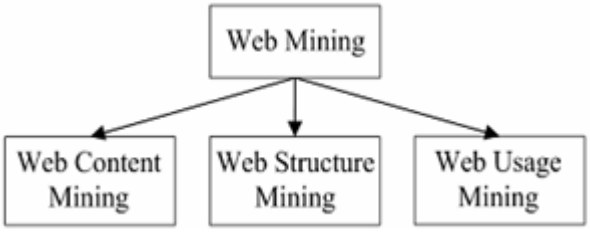

Figure 1: Classification of Web mining

II. ADAPTIVE WEB SITES

An adaptive website adjusts the structure, content, or presentation of information in response to measured user interaction with the site, with the objective of optimizing future user interactions. Adaptive websites are web sites that automatically improve their organization and presentation by learning from their user access patterns. User interaction patterns may be collected directly on the website or may be mined from Web server logs. A model or models are created of user interaction using artificial intelligence and statistical methods. The models are used as the basis for tailoring the website for known and specific patterns of user interaction. The adaptive web site is concerned with mining the log file of a Web site for knowledge about the Web site and its users, and using the knowledge to assist users to navigate and search the Web site effectively and efficiently.

III. ADAPTIVE WEB SITES: USAGE MINING

Web usage mining focuses on techniques that could predict user behaviour during the interaction of the user with the web.

A. Concept of web usage mining

Discovery of meaningful patterns from data generated by client-server transactions on one or more Web servers includes following sources of data:

- Automatically generated data stored in server access logs, referrer logs, agent logs, and Client-side cookies.
- E-commerce and product-oriented user events.
- User profiles and/or user ratings.
- Meta-data, page attributs page content, site structure.

Web usage mining focuses on techniques that could predict user behaviour while the user interacts with the Web. The mined data in this category are the secondary data on the Web as the result of interactions. These data could range very widely but generally we could classify them into the usage data that reside in the Web clients, proxy servers and servers. The Web usage mining process can be regarded as a three-phase process (as shown in Figure 2), consisting of the data preparation or pre-processing, pattern discovery and pattern analysis phases.

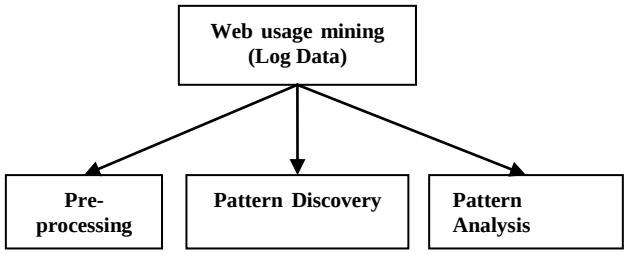

Figure 2: Phases of Web usage mining

In the first phase, Web log data are pre-processed in order to identify users, sessions, page views, and so on. In the second phase, statistical methods, as well as data mining methods (such as association rules, sequential pattern discovery, clustering, and classification) are applied in order to detect interesting patterns. These patterns are stored so that they can be further analysed in the third phase of the Web usage mining process.

B. Web Log Format

A web server log file contains requests made to the web server, recorded in chronological order. The most popular log file formats are the Common Log Format (CLF) and the extended CLF. A common log format file is created by the web server to keep track of the requests that occur on a web site. A standard log file has the following format as shown in Figure 3.

<IP_addr><base_url>.<date><method><file><protocal><code><bytes><referrer><user_agent>

Figure 3: Common Web Log Format

IV. Approaches in Web Usage Mining

The web usage mining generally includes the following several steps: data collection, data pre-treatment or data pre-processing and knowledge discovery and pattern analysis [2].

A. Data Collection

Data collection is the first step of web usage mining, the data authenticity and internality will directly affect the following works smoothly carrying on and the final recommendation of characteristic service's quality.

Therefore it must use scientific, reasonable and advanced technology to gather various data. At present, towards web usage mining technology, the main data origin has three kinds: server data, client data and middle data.

B. Data Pre-processing

Some databases are insufficient, inconsistent and including noise. The data pre-treatment is to carry on an unification transformation to those databases. The result is that the database will to become integrate and consistent, thus establish the database which may mine. In the data pre-treatment work, mainly include data cleaning, user identification, session identification and path completion as shown in Figure 4.

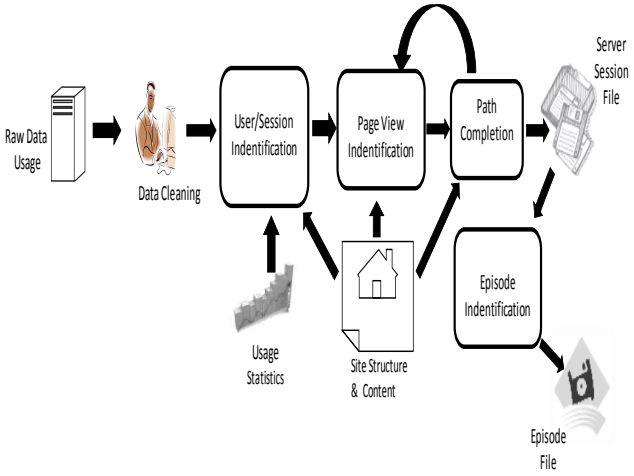

Figure 4: Pre-processing of Web Usage Data

1) Data Cleaning: The purpose of data cleaning is to eliminate irrelevant items, and these kinds of techniques are of importance for any type of web log analysis not only data mining. According to the purposes of different mining applications, irrelevant records in web access log will be eliminated during data cleaning. Since the target of Web Usage Mining is to get the user's travel patterns, following two kinds of records are unnecessary and should be removed:

- The records of graphics, videos and the format information the records have filename suffixes

of GIF, JPEG, CSS, and so on, which can found in the URI field of the every record.

- The records with the failed HTTP status code. By examining the Status field of every record in the web access log, the records with status codes over 299 or fewer than 200 are removed.

2) *User and Session Identification:* The task of user and session identification is find out the different user sessions from the original web access log. User's identification is, to identify who access web site and which pages are accessed. The goal of session identification is to divide the page accesses of each user at a time into individual sessions. A session is a series of web pages user browse in a single access. The difficulties to accomplish this step are introduced by using proxy servers, e.g. different users may have same IP address in the log. A referrer-based method is proposed to solve these problems in this study. The rules adopted to distinguish user sessions can be described as follows:

- The different IP addresses distinguish different users.
- If the IP addresses are same, the different browsers and operation systems indicate different users.
- If all of the IP address, browsers and operating systems are same, the referrer information should be taken into account. The Refer URI field is checked, and a new user session is identified if the URL in the Refer URI field hasn't been accessed previously, or there is a large interval (usually more than 10 seconds) between the accessing time of this record and the previous one if the Refer URI field is empty.
- The session identified by rule 3 may contains more than one visit by the same user at different time, the time oriented heuristics is then used to divide the different visits into different user sessions. After grouping the records in web logs into user sessions, the path completion algorithm should be used for acquiring the complete user access path.

3) *Path completion:* Another critical step in data pre- processing is path completion. There are some reasons that result in path's incompletion, for instance, local cache, agent cache, "post" technique and browser's "back" button can result in some important accesses not recorded in the access log file, and the number of Uniform Resource Locators (URL) recorded in log may be less than the real one. Using the local caching and proxy servers also produces the difficulties for path completion because users can access the pages in the local caching or the proxy servers caching without leaving any record in server's access log.

As a result, the user access paths are incompletely preserved in the web access log. To discover user's travel pattern, the missing pages in the user access path should be appended. The purpose of the path completion is to accomplish this task. The better results of data pre-processing, we will improve the mined patterns' quality and save algorithm's running time. It is especially important to web log files, in respect that the structure of web log files are not the same as the data in database or data warehouse. They are not structured and complete due to various causations. So it is especially necessary to pre-process web log files in web usage mining. Through data pre-processing, web log can be transformed into another data structure, which is easy to be mined.

C. Knowledge Discovery

Use statistical method to carry on the analysis and mine the pre-treated data. We may discover the user or the user community's interests then construct interest model. At present the usually used machine learning methods mainly have clustering, classifying, the relation discovery and the order model discovery. Each method has its own excellence and shortcomings, but the quite effective method mainly is classifying and clustering at the present.

D.Pattern Analysis

A challenge of Pattern Analysis is to filter uninteresting information and to visualize and interpret the interesting patterns to the user. First delete the less significance rules or models from the interested model storehouse; Next use technology of OLAP and so on to carry on the comprehensive mining and analysis; Once more, let discovered data or knowledge be visible; Finally, provide the characteristic service to the electronic commerce website.

V. DATA SOURCES

Web Usage Mining applications are based on data collected from three main sources: (i) Web servers, (ii) proxy servers, and (iii) Web clients [3].

A. Web Servers

Web servers are surely the richest and the most common source of data. They can collect large amounts of information in their log files and in the log files of the databases they use. These logs usually contain basic information e.g.: name and IP of the remote host, date and time of the request, the request line exactly as it came from the client, etc. This information is usually represented in standard format e.g.: Common Log Format, Extended Log Format, Log XML. When getting log information from Web servers, the major issue is the identification of users sessions to clearly identify the paths that users followed during navigation through the web site. This task is usually quite difficult and it depends on the type of information available in log files. The most

common approach is to use cookies to track down the sequence of users_ page requests. If cookies are not available, various heuristics can be employed to reliably identify users_ sessions.

B. Proxy Server

Many Internet Service Providers (ISPs) give to their customer proxy server services to improve navigation speed through caching. In many respects, collecting navigation data at the proxy level is basically the same as collecting data at the server level. The main difference in this case is that a proxy server collects data of groups of users accessing huge groups of web servers. Even in this case, session reconstruction is difficult and not all users_ navigation paths can be identified. When there is no other caching between the proxy server and the clients, the identification of users_ sessions is easier.

C. Web Client

Usage data can be tracked also on the client side by using JavaScript, Java applets, or even modified browsers. These approaches rely heavily on the user's cooperation and raise many issues concerning the privacy laws, which are quite complicate.

VI.APPLICATIONS OF WEB USUAGE MINING

The general goal of Web Usage Mining is to gather interesting information about user access patterns. This information can be exploited later to improve the Web site from the users_ viewpoint. The results produced by the mining of Web logs can used for various purposes [3]: (i) to personalize the delivery of Web content; (ii) to improve the user navigation through pre-fetching and caching; (iii) to improve Web design; or in e-commerce sites (iv) to improve the customer satisfaction.

A. Personalization of Web Content

Web Usage Mining techniques can be used to provide personalized Web user experience. For instance, it is possible to anticipate the user behavior in real time by comparing the current navigation pattern with typical patterns which were extracted from past Web log files. In this area, recommendation systems are the most common application; their aim is to recommend interesting links to products which could be interesting to users. Personalized Site Maps are an example of recommendation system for links proposed an adaptive technique to reorganize the product catalog according to the forecasted user profile [3].

B. Pre-fetching and Caching

The results produced by Web Usage Mining can be exploited to improve the performance of Web servers and Web-based applications. Typically, Web Usage Mining can be used to develop proper pre-fetching and caching strategies so as to reduce the server response time.

C. *Support to Design*

Usability is one of the major issues in the design and implementation of Web sites. The results produced by Web Usage Mining techniques can provide guidelines for improving the adaptive web design.

D. E-commerce

Mining business intelligence from Web usage data is dramatically important for ecommerce Web-based companies. Customer Relationship Management (CRM) can have an effective advantage from the use of Web Usage Mining techniques. In this case, the focus is on business specific issues such as: customer attraction, customer retention, cross sales, and customer departure.

Web mining for usage pattern is the key to discover marketing intelligence in e-commerce. It helps tracking of general access pattern, personalization of web link or web content and customizing adaptive sites. It can disclose the properties and inter-relationship between potential customers, users and markets, so as to improve Web performance, on-line promotion and personalization activities.

There are many popular programs for usage pattern mining. (Table 1)[1]

Table 1: Major Applications of web usage mining & Related Projects

S. No.	Application	Tools/Projects
1.	Usage Characterization	1. Manley 2. Pitkow 3. Almeida
2.	System Improvement	1. Rexford 2. Schecter 3 Agarwall
3.	Business Intelligence	1. SurfAid 2. Buchner 3.Tuzilin 4. Abrahim
4.	Personalization	1. Site Helper 2. Web shifter 3. Mobasher 4. Letizia
5.	Site Modification	1. Etzioni 2. Perkowitz

Web Log Mining uses KDD techniques to understand general access patterns and trends to shed light on better structure and grouping of resource providers. For e.g. Web miner discovers association rules and sequential patterns automatically from server access logs. Commercial software Web Analyst by Megaputer learns the interests of the visitors, based on their interaction with the website. Clementine and DB2 Intelligent Miner for Data are two general-purpose data mining tools, which can be used for web usage mining with suitable data pre-processing.

There are several commercial software tools that could provide Web usage statistics. These statistics could be useful for Web administrators to get a sense of the actual load on the server. However, the statistical data available from the normal Web log data files or even the information provided by Web

trackers could only provide the information explicitly because of the nature and limitations of the methodology itself. The analysis relies on three general sets of information: (1) past usage patterns; (2) degree of shared content; and (3) inter memory associative link structures.

VII.TECHNIQUES USED FOR WEB OPTIMIZATION

Data mining methods have been used to analyse the data on the Web and extract useful knowledge. Web Usage Mining is widely recognized as a valuable source of ideas and solutions for Web optimization. This section explains some of techniques used in Adaptive web sites. Most of the commercial applications of Web Usage Mining exploit consolidated statistical analysis techniques. In contrast, research in this area is mainly focused on the development of knowledge discovery techniques specifically designed for the analysis of Web usage data. Most of this research effort focuses on the following main paradigms: association rules, sequential patterns, and clustering, classification, statistical analysis, dependency modelling [3].

A. Association Rules

It is used to relate pages that are most often referenced together in a single server session. In context of Web Usage Mining, association rules refer to set of pages that are accessed together with a support value more than some specified threshold. Association rule mining using Apriori algorithm may find correlation between users who visited a page containing electronic products to those who access a page about sport related products. When applied to Web Usage Mining, association rules are used to find associations among Web pages that frequently appear together in users_ sessions. The typical result has the form:

"A: html, B: html -> C: html"

which states that if a user has visited page A.html and page B.html, it is very likely that in the same session the same user has also visited page C.html.

B. Sequential Patterns

It is used to find inter-session patterns such that the presence of a set of items is followed by another item in a time-ordered set of sessions or episodes. Web marketers can predict future visit patterns which will be helpful in placing advertisements aimed at certain user groups. Also sequential pattern analysis can be used to find patterns in trend analysis, change point detection or similarity analysis.

C. Clustering

The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. In the Web Usage Mining, there are two types of interesting clusters to be discovered, usage clusters and page clusters. Clustering of users tends to find groups of users showing similar browsing patterns. Such knowledge is useful for inferring user demographics in order to perform market segmentation in E-commerce applications or provide personalized Web content to the users. Also clustering of pages will discover groups of pages having related content. This information is very useful for Internet search engines and Web assistance providers.

D. Classification

It is the task of categorizing data items into one of several predefined classes. It can be done by using supervised learning algorithms such as decision tree classifiers, naïve Bayesian classifiers, k-nearest neighbour classifiers, Support Vector Machines etc. [5]. Weblog information can be integrated with Web content and Web linkage structure mining to help Web page ranking, Web document classification and the construction of multilayered Web information base as well. In a particular discipline, the documents need to be classified based on subject index classification standard such as to classify a set of Web documents automatically, Web linkage information to improve the quality of such classification, use Web usage information to improve the quality of such classification [6].

E. Statistical Techniques

Statistical techniques are most common methods to extract knowledge about visitors to a web site. Many traffic analysis tools produce a periodic report containing statistical information such as the most frequently accessed pages, average view time of a page or average length of a path through a site. This report may include limited low-level error analysis such as detecting unauthorized entry points or finding the most common invalid URI. This type of knowledge can be potentially useful for improving the system performance, enhancing the security of the system; facilitating the site modification task and providing support for marketing decisions [5].Weblog records provide rich web usage information for data mining. Mining Weblog access sequence may help pre fetch certain Web pages into a Web server buffer, such as those pages that are likely to be requested in the next several clicks [6].

F. Dependency Modelling

It is one of the useful pattern discovery tasks in Web Mining. Here a model is to be developed capable of representing significant dependencies among the various variables in the Web domain. For example, one may build model representing the different steps a visitor follows while shopping online or visiting a job portal and registering to job portal, or performing online billing of premiums, transactions done etc . Here the techniques need to be used to model the browsing behaviour of users. The techniques include Hidden Markov Model and Bayesian Belief Networks. Modelling of Web usage patterns will provide theoretical framework for analysing the behaviour of users and also useful for predicting future Web resource consumption. Such information may help in

developing strategies to increase the sales of products or improve the navigational convenience of users.

VIII.CHALLENGES AND FUTURE TRENDS

With explosive growth of the information sources available on the World Wide Web, it has become increasingly necessary for users to utilize automated tool in order to find the required information resources, and to track and analyse their usage patterns. These factors give rise to the necessity of creating server side and client side intelligent systems that can effectively mine for knowledge. The analysis of large web log files is a complex task not fully addressed by existing web access analysers. However, it is hard to find appropriate tools for analysing raw web log data to retrieve significant and useful information. The existing techniques for analysing web usages have different drawbacks, i.e., either huge storage requirements, excessive I/O cost, or scalability problems when additional information is introduced into the analysis. Most of the currently available Web server analysis tools provide only explicitly and statistical information without real useful knowledge for Web managers. The task of mining useful information becomes more challenging when the Web traffic volume is enormous and keeps on growing. The sharpening on the mining tools in many different aspects is important for the future development in this area:

- Web usage mining must handle the integration of offline data with e-business analytic tools, RDBMS, catalogues of products and services and other applications.
- Some new variables or logs should be sought that can be used for finding more natural, meaningful and useful patterns.
- New tools are needed which will not use up too much resources or process time during the web mining process.
- There will always be a need to have benchmark tests to improve the performance of mining algorithms, as the efficiency and effectiveness of a mining algorithm can be measured and a better tool for web data mining can be derived.
- It is important to improve visualization, as much of the data is unorganized and difficult for the user to understand.

IX. CONCLUSION

Web Usage Mining has been an important area in data mining research in recent years from the standpoint of supporting human-centred discovery of knowledge. The present day model of web mining suffers from a number of short comings. As services over the web continue to grow, there will be a continuing need to make them robust, scalable and efficient. Web usage mining can be applied to better understand the behaviour of these services, and the knowledge extracted can be useful for various indices of optimizations. Web usage mining (WUM) can be used to determine if the Information architecture of a web site is structured correctly. Existing WUM tools however, do not indicate which data mining algorithms are being used or provide effective graphical visualizations of the results obtained.

In future, web usage mining research promises lot of space for advancements in the techniques and tools that can make some improvements in web sites specifically by focusing on the visualization of user navigation pattern by using combination technology of knowledge-based system and web-mining method. The purpose of web usage mining is to make contributions in improving the overall quality of Information Systems, to support designers during the design process with using existing data structures to ensure ease and quick access for end users.

REFERENCES

[1] Chhavi Rana, "A Study of Web Usage Mining Research Tools", *Int. J. Advanced Networking and Applications*, Volume: 03 Issue: 06 Pages: 1422-1429 (2012) ISSN: 0975-0290.

[2] Rajni Pamnani, Pramila Chawan, "Web Usage Mining: A Research Area in Web Mining".

[3] Federico Michele Facca, Pier Luca Lanzi, "Mining interesting knowledge from weblogs: a survey", *Data & Knowledge Engineering* 53 (2005) 225–241.

[4] H.-Y. Paik, B. Benatallah, R. Hamadi, "Dynamic restructuring of e-catalog communities based on user interaction patterns", *World Wide Web* 5 (4) (2002) 325–366.

[5] Srivastava J., Cooley R., Deshpande, M. and Tan, P. N., "Web Usage Mining: Discovery and applications of Usage Patterns from web data", *SIGKDD Explorations* 2000.

[6] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, 2nd Edition, Elsevier

[7] H.K. Dai, B. Mobasher, "Using ontologies to discover domain-level web usage profiles" , *Proceedings of the 2nd Semantic Web Mining Workshop at ECML/PKDD* 2002, Helsinki, Finland, August 2002.

[8] M. Eirinaki, M. Vazirgiannis, I. Varlamis,"SEWeP: Using Site Semantics and Taxonomy to Enhance the Web Personalization Process", *SIGKDD* '03, August 24-27, 2003.

[9] B. Masand, M. Spiliopoulou, J. Srivastava, and O. R. Zaiane, "Web Mining for Usage Patterns & Profiles", *WEBKDD02, SIGKDD Explorations*, 4(2), 2002

[10] H. Dai, and B. Mobasher, "Integrating Semantic Knowledge with Web Usage Mining for Personalization", *Information Systems Journal*, 2009, 1-28.