Reorganisation of Adaptive Websites Using Web Usage Mining Techniques

Dr. Ananthi Sheshasayye¹, V. Vidyapriya²
¹Head and Associate Professor, Department of Computer Science, Quaid-E-Millath Government College for Women (A), Chennai,
²Research Scholar, Department of Computer Science, Quaid-E-Millath Government College for Women (A), Chennai,

Abstract—Web Usage Mining is the part of Web mining for extraction of interesting knowledge from log files produced by Web Servers. It focuses on techniques that predict the user behaviours based on the previous user access patterns. Adaptive Web Sites are web sites that automatically enhance their organisation and presentation from the knowledge obtained from their user access patterns. These patterns can be collected directly on the website or web server log may be mined. 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, Adaptive web site, Web usage mining, Web Log, Web site reorganisation.

I. INTRODUCTION

Web mining is a very important research topic that combines Data Mining and World Wide Web. With the huge amount of information available in online, the Web is a key medium for data mining research. The Web mining research uses the data base, information retrieval and AI. The web is huge and interest in analysis of user behaviour on the web is also increased.

Web data mining is defined as the discovery, extraction and analysis of useful facts from the web log file. The unstructured feature of Web data generates more complexity in the Web mining process. Web mining is the use of data mining techniques to automatically discover and extract useful patterns from Web documents [1]. The three classifications of Web mining are Web content mining, Web structure mining, and Web usage mining based on which part of the Web to mine as shown in Figure 1.

In practice, these three Web mining tasks can be used in isolation or combined in an application. Web content mining is focusing on the process of extracting knowledge from the content of documents or their descriptions. Web structure mining adapts the process of inferring knowledge from the World Wide Web and links between references and referents in the Web. The web usage mining, also known as Web Log Mining, is the process of extracting interesting patterns in web access logs [2].

II. ADAPTIVE WEB SITES

An adaptive website has the property that it automatically adjusts the structure, content as well as the presentation of information in response to measured user interaction with the web site to optimize future user interactions. Adaptive websites can be defined as web sites that automatically improve their organization and appearance by learning from the user access patterns [3]. User interaction patterns can be collected either directly on the website or mining from Web server logs. Then a model or models are created using artificial intelligence and statistical methods. These models are used as the basis for improving the website for known and specific patterns of user interaction. The adaptive web site deals with mining the log file of a Web site for knowledge about the Web site and its users, and using the knowledge obtained to support users to navigate and search the Web site effectively and efficiently [3].

III. ADAPTIVE WEB SITES: USAGE MINING

Web usage mining emphasizes on techniques which are used to analysis the user behaviour during interaction of the user with the web pages.

A. Notion of web usage mining

The meaningful patterns from data generated by client-server transactions on one or more Web servers includes following sources of data [2]:
• data that are stored in server access logs, referrel logs, agent logs, and Client-side cookies.
• product-oriented user events such as e-commerce
• User profiles and user ratings.
• data from web page access, page content and web site structure.

The above mined data are the secondary data on the Web as the result of user interactions on the web. These data could come from the usage data that reside in the Web clients, proxy servers and servers. The Web usage mining process can be done by following three steps as shown in Figure 2, consisting of the data preparation or pre-processing, pattern discovery and pattern analysis phases [1].

![Figure 2: Phases of Web usage mining](image)

In the first step, Web log data’s are pre-processed in order to identify users, sessions, page views. In the second step, statistical methods, dependency modelling as well as data mining methods such as association rules, sequential patterns, 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 step of the Web usage mining process to optimize the web page access [1].

B. Format of Web Log File

A web server log file contains requests made to the web server, recorded in sequential order. The Common Log Format (CLF) and the extended CLF are the most popular log file formats. A common log format file is created by the web server to keep track of the requests made for a web site. A standard log file has the following format as shown in Figure 3[2].

![Figure 3: Web Log Format](image)

|IP_addr|user|date|method|file|protocol|code|bytes|referer|user_agent|

IV. PHASES OF WEB USAGE MINING

The web usage mining can be done in the following 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, because the data authenticity will affect the following works as well as the final recommendation for user access behaviour. So the data collection must be done using scientific, reasonable and advanced technology to gather various data. At present, the web usage mining technology has three main data origins: server data, client data and middle data like proxy server data.

B. Data Pre-processing

Some databases may have insufficient, inconsistent data including noise. The data pre-processing is done by transformation to those databases for unification. Then the database is become integrate and consistent and thus establish the database which may mine. The data pre-processing includes data cleaning, user identification, session identification and path completion as shown in Figure 4 [2].

![Figure 4: Pre-processing of Log file](image)

1) Data Cleaning: The purpose of data cleaning is to remove inappropriate items. As in the need of different mining applications, irrelevant records in web access log file are eliminated during data cleaning. Since the Web Usage Mining is used to get the user’s travel patterns, following two kinds of records should be removed from the log files:

• The records of graphics, videos and the format information from the log file. These records have filename with suffixes like GIF, JPEG and CSS.
The records with the failed HTTP status code. By examining the Status field, records with status codes over 299 or fewer than 200 are removed from web access log [2].

2) User and Session Identification: The task of user and session identification is to find the different user sessions from the available information in the log file. The purpose of user’s identification is to identify users those who access web site and pages that are accessed. The need of session identification is to divide the page accesses of each user at a time based on individual sessions. The rules adopted to distinguish user sessions can be described as follows[2]:

- The different IP addresses represent different users.
- If the IP addresses are same, the different browsers and operation systems show different users.
- If all of the IP address, browsers and operating systems are same, the referer 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 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, then time oriented heuristics is used to divide the different visits into different user sessions. After grouping the records in web logs into user sessions, the path completion algorithm can be used for acquiring the complete user access path [2].

3) Path completion: The path completion is the critical step in data pre-processing. The local cache, agent cache, “post” technique and browser’s “back” button may result in some of the important accesses are not recorded in the access log file, and hence the number of Uniform Resource Locators (URL) recorded in log file may be less than the real number of accesses. The usage of local cache and proxy servers produce the difficulties in 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. The missing pages in the user access path should be appended to discover user’s travel pattern. This task is achieved by the path completion. Through data pre-processing, web log can be transformed into another data structure that can be mined easily for knowledge discovery.

C. Knowledge Discovery

The statistical methods are used to analyse and mine the pre-processed data. At present, the machine learning methods used are clustering, classification, the relation discovery and the order model discovery. Each method has its own advantages and disadvantages; the quite effective methods are classifying and clustering [2].

D. Pattern Analysis

It is the last step in the web usage mining. The purpose of pattern analysis is to filter the uninteresting information and to visualize and interpret the interesting patterns to the user. It is accomplished by first deleting the less significance rules or models from the interested model. Then use technology such as OLAP to carry the comprehensive mining and analysis. While performing the above task, let discovered data or knowledge be visible. These characteristic services are provided to the electronic commerce website [2].

V. WEB USAGE: 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 [4].

A. Web Servers

Web servers are the most common source of data. They have large amount of information in their log files. These logs contain basic information such as name and IP of the remote host, date and time of the request, the request line that came from the client. This information is represented in standard format using Common Log Format, Extended Log Format and Log XML. The identification of users’ session is the major issue in getting log information from Web servers, to identify the paths that users followed during navigation through the web site. The most appropriate approach is to use cookies to track down the sequence of users' page requests. If cookies are not available, other experimental methods can be used to identify users’ sessions [4].

B. Proxy Server

The proxy server services are given by Many Internet Service Providers (ISPs) to their customer to improve navigation speed through caching. Collecting navigation data through the proxy server is basically the same as collecting data at the server level. But a proxy server collects data of groups of users accessing huge groups of web servers. The identification of users’ sessions is easier when there is no other caching between the proxy server and the clients [4].

C. Web Client

By using JavaScript, Java applets or even modified browsers, usage data can be tracked on the client side. These approaches rely heavily on the user’s cooperation and raise many issues that depends on privacy laws, which are quite complicate [5].

VI. APPLICATIONS OF WEB USAGE MINING

The general goal of Web Usage Mining is to collect interesting information about user access patterns. The outcomes from the mining of Web logs can be used for various purposes [4]: (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 create personalized web site as per web user experience. By comparing the current navigation pattern with typical patterns...
which were extracted from previous entries from Web log files it is possible to anticipate the user behavior in real time. The most common application is the recommendation system; its aim is to recommend interesting links to products which could be interesting to users. An example of recommendation system as the links proposed for adaptive technique to reorganize the product catalog according to the forecasted user profile is Personalized Site Maps [4].

B. Pre-fetching and Caching

The results produced by Web Usage Mining can be used to improve the performance of Web servers and Web-based applications. The Web Usage Mining can be used to develop proper pre-fetching and caching strategies to reduce the server response time [4].

C. Support to Design

The major issue in the design and implementation of Web site is usability. The results produced by Web Usage Mining techniques can provide strategies for improving the adaptive web design [3].

D. E-commerce

Mining business intelligence from Web usage data is important for e-commerce Web-based companies. Customer Relationship Management (CRM) adapts the effective advantage from the use of Web Usage Mining techniques. Here customer attraction, customer retention, cross sales, and customer departure issues are focused [4].

To discover the marketing intelligence, web mining for usage pattern is done for E-commerce. It helps to keep track of the general access pattern, personalization of web link or web content and customizing adaptive sites. It can also be used to get the properties and inter-relationship between potential customers, users and markets to improve web performance and on-line promotion [1].

The various popular programs for usage pattern mining is shown in Table 1[1].

<table>
<thead>
<tr>
<th>S. No.</th>
<th>Application</th>
<th>Tools/Projects</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Usage Characterization</td>
<td>Manley, Pitkow, Almeida</td>
</tr>
<tr>
<td>2.</td>
<td>System Improvement</td>
<td>Rexford, Schecter, Agarwali</td>
</tr>
<tr>
<td>3.</td>
<td>Business Intelligence</td>
<td>Surf Aid, Buchner, Tuzilin, Abraham</td>
</tr>
<tr>
<td>4.</td>
<td>Personalization</td>
<td>Site Helper, Web shifter, Mobasher, Letizita</td>
</tr>
<tr>
<td>5.</td>
<td>Site Modification</td>
<td>Etzioni, Perkowitz</td>
</tr>
</tbody>
</table>

Table 1: Applications of web usage mining & Related Projects

In Web Log Mining, the Knowledge Discovery techniques are used to understand general access patterns and methods to improve the web site structure and grouping of resources providers. The Web miner tool determines association rules and sequential patterns automatically from server access logs. The Commercial software tool Web Analyst learns the interests of the visitors based on their interaction with the website. The two general-purpose data mining tools that can be used for web usage mining with suitable data preprocessing are Clementine and DB2 Intelligent Miner [1].

The statistics provided by the commercial software tools are used by web administrators to get a sense of the actual load on the server. The statistical data provided by the normal Web log data files are partial because of the nature and limitations of the methodology itself. The analysis relies on three general sets of information: past usage patterns, degree of shared content and inter memory associative link structures.

VII. TECHNIQUES USED FOR WEB OPTIMIZATION

Data mining methods are used to analyse the data on the Web and extract useful knowledge. This section explains some of the techniques used in Adaptive web sites for Web optimization. The commercial applications of Web Usage Mining uses consolidated statistical analysis techniques. The research in this area is mainly focused on the development of knowledge discovery techniques designed for the analysis of Web usage data. Most of this research effort based on the following main paradigms: association rules, sequential patterns, and clustering, classification, statistical analysis, dependency modelling [4] [6].

A. Association Rules

It is used to relate pages that are referenced together in a single server session. In 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 A priori algorithm may find correlation between users of different interest. 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 [4].

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. Web marketers can predict a future visit pattern that helps to place advertisements aimed at certain user groups. The sequential pattern analysis can also be used to find patterns in trend analysis and similarity analysis [4].

C. Clustering

The clustering techniques look for groups of similar items in the large data based on a distance function that computes the similarity between groups. 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. The two types of interesting clusters to be discovered in web usage mining are usage clusters and page clusters. Clustering of users tends to find groups of users showing similar browsing patterns. It is used to perform market segmentation in E-commerce applications or provide personalized Web content to the users. The clustering of pages will discover groups of pages having related content. It is very useful for Internet search engines and Web assistance providers [6].

D. Classification

With the Classification, the data items are categorized into one of several predefined classes. It is done by using supervised learning algorithms such as decision tree classifiers, naive Bayesian classifiers, k-nearest neighbour classifiers, Support Vector Machines [7]. Weblog information can be used with Web content and Web linkage structure mining to help Web page ranking, Web document classification. In a particular discipline, the documents need to be classified based on subject index classification standard. Then use Web linkage information to improve the quality of such classification. Finally use Web usage information to improve the quality of such classification [8].

E. Statistical Techniques

The Statistical techniques are used to extract knowledge about visitors to a web site. Many web traffic analysis tools produce a periodic report containing statistical information like the most frequently accessed pages, average view time of a page or average length of a path through a site. It may include limited low-level error analysis such as detecting unauthorized entry points and finding the most common invalid URI. This type of knowledge is used for improving the system performance, enhancing the security of the system [7]. The Weblog records provide web usage information for data mining. The pages that are likely to be needed in the next upcoming clicks may be pre fetched into web server buffer using the above log access sequence [6][8].

F. Dependency Modelling

A model needs to be developed which will represent major dependencies between the various variables in the Web domain. Some of the techniques used to model the browsing behaviour of users are Hidden Markov Model and Bayesian Belief Networks. The models created will provide theoretical framework for analysing the behaviour of users and also useful for predicting future Web resource consumption. This information helps to develop strategies to increase the sales of products and improve the navigational convenience of users [6].

VIII. CHALLENGES AND FUTURE TRENDS

As the information sources available on the Web increases, it is necessary to utilize the automated tools to find the information sources as well as their usage patterns. It is necessary to create intelligent systems both on server side and client side to mine for the knowledge on usage patterns. The existing techniques for analysing web log files have various drawbacks such as large storage requirements, expensive I/O cost and scalability. Many server analysis tools provide only statistical information about web access. As the web complexity increases, the task of mining useful information is more challenging.

To improve the efficiency of web usage mining, the mining tools can be integrated with the offline data with e-business analytic tools and other applications; some new variables can be introduced to find meaningful and useful patterns. The visualisation of usage patterns can also be improved using knowledge mining.

IX. CONCLUSION

Web Usage Mining is based on mining server log files. It provides support to design a web site, to design personalisation of server access, and also for other business decision making. As the utilities of the web continue to grow, there is a need to make them as scalable and efficient. The web usage mining can be used for understanding the behaviour of those services and also the knowledge extracted from various usage patterns can be useful for optimization of web sites.

With the existing techniques, it is hard to differentiate between the new visitors and visitors who already navigate between the various pages of a website. By focusing on the visualisation of user navigation patterns by using knowledge extraction through various data structures, the adaptive web sites can be reorganised for effective utilisation of Web.

REFERENCES