WEB LOG MINING – CASE STUDY

1Ramya.R, 2Lalitha Devi.K
Research scholar, Bharathiyar University

Abstract—The world of web is very diverse and contains huge amount of information. To obtain relevant information from the www is a challenging task today. To ensure the retrieval of pertinent information, suitable web mining techniques can be applied. All most all the application of data mining techniques to the World Wide Web referred as Web mining where this term has been used in three distinct ways; Web Structure Mining, Web Usage Mining and Web Content Mining. E Learning is the Web based application where it will facing with large amount of data. In order to produce the server logs patterns and user performance, this work implements the high level process of Web Log Mining using some modification in basic Association Rules algorithm call modified. Web Log Mining consists of three main phases, namely Preprocessing of Log Record, Discovering of Pattern and Analysis of Pattern. Server log files become a set of raw data where it’s must go through with all the Web Log Mining phases to producing the final results. Here, Web Log Mining, approach has been combining with the basic Association Rules, with modified to optimize the content of the server log record.

Keywords—Web Mining, Apriori Algorithm, NASA Classification, Web Query Logs.

I. INTRODUCTION

Web mining is a suitable technique to explore the world of web and fetch the desired information. Web logs are important structures maintained by different web servers to capture the important information which may contain the IP address, requested URL, timestamp etc. [1] Web based logs are useful for the web designers as well as for the design theorists for effectively studying the inferences generated that helps in the web site building, testing and modification over time. [2] popular Web sites attract millions of visitors. These visitors leave behind vast amount of Web site traversal information in the form of Web server and query logs. By analyzing these logs, it is possible to discover various kinds of knowledge, which can be applied to improve the performance of Web services. A particularly useful kind of knowledge is knowledge that can be immediately applied to the operation of the Web site; we call this type of knowledge the actionable knowledge. In this paper, we present three examples of actionable Web log mining. The first method is to mine a Web log for Markov models that can be used for improving caching and prefetching of Web objects. A second method is to use the mined knowledge for building better, adaptive user interfaces. The new user interface can adjust as the user behavior changes with time.

A. Classification Of Web-Log Analysis

Web log analysis Web mining is a suitable technique to discover and extract interesting knowledge/patterns from Web. The shortcoming of different search engines provides a path for the development of suitable web mining techniques to improve the quality of the data retrieval process.

Web huge volumes of data being collected create new challenges for such techniques as organizations look for ways to make use of the stored information to gain an edge over competitors. It is reasonable to believe that data collected over an extended period contains hidden knowledge about the business or patterns characterizing customer profile and behavior. With the rapid growth of the World Wide Web, the study of knowledge discovery in web information, modeling information and predicting the user’s access on web site information has become very important. From the application point of view, business and administration, knowledge obtained from the Web usage patterns could be directly applied to efficiently manage activities related to e-CRM, e-Business, e-Services, e-Newspapers, e-Education, Digital Libraries e-Government, and so on [3].

Data collection

The first step is to collect relevant data from the different sources. There are three main sources for data in usage mining: server side data, client side data and intermediary Server data Server data are data that are collected from web servers; it includes log files, cookies and explicit user input.

i) Server log files

Servers contain different types of logs, which are considered to be the main date resource for web usage mining. The most popular logs are:

Common Log Format (CLF): created to keep track of requests that occur on a website in chronological order. It contains the IP address of the client, hostname, username, time stamp, file name and file size. CLF has the following elements:
Remote host: the IP address or domain name of the client.
Base URL: the URL of the user request.
Date: the date and time of the request.
Method: the method used by the client, such as GET, HEAD or POST.
File: the file requested by the client.
Protocol: the protocol used.
Code: the status code of the three requests; it consists of 3 Digits.
Bytes: the number of bytes returned to the client
Referrer: the URL from the referring server
User agent: the operating system type and version data.

ii) Cookies
Cookies are strings that are sent from the web server to the Client’s browser. The browser saves the cookie in a text file And resends it to the server each time the user visits the site. This way, the server stores information about all visitors, the Pages visited, any products purchased, etc. in the cookie log within the client’s machine. The main advantage of cookies is their small size. However, cookies, like log files, may contain unreliable user.

II. RELATED WORK

Web usage mining is the “Applying data mining techniques to web data repositories to extract patterns “.Data mining techniques that are commonly used includes association rules, sequential pattern, clustering, and classification. Association rules are used to find the relationship between attributes from the item set. In web usage mining item set is set of pages. Rules are applied to discern pages which are often looked together In order to reveal associations between groups of users with specific interests. The resulted guidelines to web designers for reorganizing Websites. [3] Used association rules to decide the next likely web page requests based on significant statistical correlations.

Clustering is a technique to group together items that have similar features. In Web usage domain, there are two clustering groups, user clusters and page clusters. Page clustering generates the group of pages that are considered to be related according to user view. In user clustering the goal is to group users which have same browsing patterns. Such understanding can be used in business to perform market segmentation and Web site personalization.

Created a model by applying clustering algorithm, and then the model is adjusted by statistical approach based on the change of Behavior of users or data domain of website periodically.[12] Proposed to integrate Markov model based sequential pattern mining with clustering.[8] experimented for many of the tunable parameters, such as the time delta involved in session zing logs, confidence and support for associations, initializing of the melodies in clustering. Classification is a method that maps a data item into one of several predefined classes. In Web usages mining the users are in different classes according to their profiles.

III. RESEARCH METHODOLOGY

A. Web Query Logs:

In this thesis, now describe an application whereby we apply data mining to an Internet search engine in this paper. In particular, we apply data mining to discover useful and implicit knowledge from user logs of the Microsoft Encarta search engine (http://encarta.msn.com) to improve its performance. The user logs keep traces of users behaviors when they use the Encarta search engine. The purpose of the weblog mining is to improve Encarta search engine’s performance (which is defined precisely later in the paper) by utilizing the mined knowledge as cache. Indeed, data mining is a promising approach since popular search engines like Encarta get hundreds of thousands of hits each single day, and therefore, it would be infeasible for traditional methods or human to analyze such logs.

Figure 3. 1. Prediction to frequency weight calculation

Simultaneously exist a number of sessions on a web server. Based on their access sequences, our prediction model can predict future requests for each particular session. Different sessions will give different predictions to future objects. Since our prediction of an object comes with a probability of its arrival, we can combine these predictions to calculate the future occurrence frequency of an object.

Let Oi denote a web object on the server. Sj be a session on a web server, Pi,j be the probability predicted by a session Sj for object Oi. If Pi,j=0, it indicates that object Oi is not predicted by session Sj. Let Wi be the future frequency of requests to object Oi. A weight can be computed according to the following equation:

\[
P_{i,j} = \sum_{k=1}^{n} P_{i,k} \times W_{i,k}
\]

where W_{i,k} is the weight of object Oi in session S_k.
which lacks an integrated structure or schema, it becomes much more difficult for users to access relevant information efficiently. Meanwhile, the substantial increase in the number of websites presents a challenging task for web masters to organize the contents of websites to cater to the need of user’s. Analyzing and modeling web navigation behavior is helpful in understanding demands of online users. Following that, the analyzed results can be seen as knowledge to be used in intelligent online applications, refining website maps, and web based personalization system and improving searching accuracy when seeking information. Nevertheless, an online navigation behavior grows each passing day, thus extracting information intelligently from it is a difficult issue.

IV. WEB QUERY LOG MINING

4.1. Web miner classifier

NASA’s based on these cost functions, we can find a minimal value $M$ for the Overall Cost and its corresponding number of index pages to build. This optimal index-page construction process represents another major contribution of our work. What we do is to analyze the overall cost as a function of the number of index pages $M$ to construct, based on a fixed value of $L$. We can then find empirically the best value for $M$ so as to minimize the overall cost of user browsing effort. The overall cost function calculated from the combination of the web page switching cost and the index page cost with the NASA data. As seen in the figure, the overall cost has a minimum value at around 0.6. This is an indication that when considering all factors, it is the best to include around 3 to 4 index pages, where each page contains $L$ hyperlinks. The optimal number in this example is computed automatically from different statistics, rather than decided.

Fig. 4.2 Cost of page construction using the clustering results on NASA’s log.

Multi-layer perception training uses the back-propagation learning algorithm to approximate the optimal discriminate function defined by Bayesian theory. The output of the MLP approximates the posteriori probability functions of the classes being trained. The Sigmoidial activation function is used for learning the input weight vectors in the training phase.

Figure 4.2 NASA log file example
We now describe how to compose index pages in our framework. In their algorithm, Jerkewitz and E-zine first computes clusters from the web logs and then put all clusters in index pages, so that each cluster will correspond to one index page. In our experience, we have found that often each cluster will contain a large number of index pages. When hundreds of hyperlinks are included in an index page, it is very difficult for a user to find the information he/she is looking for. In addition, we feel that there should be a limited number of index pages; if the user is required to read a huge number of index pages then it might defeat the purpose of including the index pages in the first place. Therefore, in index page construction, we will include two parameters. Let the $L$ be the number of hyperlinks we would like to include in each index page, and let $M$ be the number of index pages we wish to build. Algorithm ConstructIndexPages takes the parameters $M$ and $L$ and the clusters.

**Apriori Algorithm**

The algorithm is based on the large itemset property which states: Any subset of a large itemset is large and any subset of frequent item set must be frequent. The FP-growth method is efficient and scalable for mining both long and short frequent patterns and is about an order of magnitude faster than the Apriori algorithm and also faster than some recently reported new frequent-pattern mining methods. The Apriori heuristic achieves good performance gained by (possibly significantly) reducing the size of candidate sets. However, in situations with a large number of frequent patterns, long patterns, an Apriori algorithm suffer [4].

**Execution Time Comparisons**
The execution time comparison experiment is performed on datasets with 80K, 50K and 30K records; In these graphs the response times of both the algorithms increases as the support threshold is reduced.

**Figure 4.3 Apriority Algorithm of Clusters.**

### Access log of a machine

```
2013-04-06 10:58:58 192.168.30.121(50862) -> 206.190.36.45(80) in.yahoo.com/jsa
2013-04-06 10:58:57 192.168.30.121(50860) -> 206.190.36.45(80) in.yahoo.com/p.gif;
2013-04-06 10:58:57 192.168.30.121(50851) -> 216.115.100.103(80).
```

**V. Conclusions and Future work**

The purpose of this chapter is to advocate the discovery of actionable knowledge from Web logs. Actionable knowledge is particularly attractive for Web applications because they can be consumed by machines rather than human developers. Furthermore, the effectiveness of the knowledge can be immediately put to test, making the merits of the type of knowledge and methods for discovering the knowledge under more objective scrutiny than before. In this chapter, we presented two examples of actionable Web log mining. The first method is to mine a Web log for Markov models that can be used for improving caching and prefetching of Web objects. A second method is to use the mined knowledge for building better, adaptive user interfaces. A third application is to use the mined knowledge from a query web log to improve the search performance of an Internet Search. In our future work, we will further explore other types of actionable knowledge in Web applications, including the extraction of content knowledge and knowledge integration from multiple Web sites.

**References**

1. F. Yongjian and M.Yi Shih , A Framework for Personal Web Usage Mining, Department of Computer Science University of Missouri-Rolla, 2005.
Sistemistica University di Napoli Federico II via Claudio, 2180125 Napoli, Italy 2007.


[10] Sharma, Web Usage Mining Using Neural Network Associate Prof (Comp.Sc), Peoples University, Bhopal 2012.