Web Usage Mining: A Survey on Pattern Extraction from Web Logs

Ankita Kusmakar
M.Tech Scholar Computer Science, LNCT BHOPAL, INDIA

Sadhna Mishra
Professor Computer Science, LNCT BHOPAL, INDIA

Abstract—Today, the Web has evolved by huge development in number of websites and number of accessing users. So in order to provide better service along with enhancing the quality of websites, it has become very important for the website owner to better understand their customers. This is done by mining web access log files to extract interesting patterns. This paper presents an overview of web usage mining, its techniques & also provides a summary of the pattern extraction algorithm used for Web Usage Mining.

Keywords—Web Usage Mining, World Wide Web, Pattern Discovery, Data Preprocessing, Web Mining techniques

I. INTRODUCTION

In the present world, the internet has become a vital part of our everyday life. The World Wide Web (WWW) has been influencing both its user & web sites owners. In the past decade, the growth in number of websites & visitors has increased remarkably. As a result large quantity of web data has been generated. Data mining techniques are applicable to mine interesting data. But we cannot apply the data mining techniques directly to the web data since the web data is unstructured or semi-structured. Thus we use web mining which can be applied to web data. Web usage mining [1,2] extracts meaningful & interesting patterns of usage in web site which can be uses in a various ways like improvement of web sites, checking of fraudulent element, better understanding of user behaviour etc.

Categories of web mining: Web mining is categorised under three categories as shown in figure 1:
1) Web Content Mining: Web content mining is a process of extracting information from texts, images and other contents.
2) Web Structure Mining: Web Structure Mining is a process of extracting information from linkages of web pages.
3) Web Usage Mining: Web Usage Mining is a process of extracting information from how to use web sites.

Phases of Web Usage Mining:
There are 3 main phases of Web Usage Mining as shown in figure 2:
1) Preprocessing: This stage includes usage preprocessing, content preprocessing and structure preprocessing. This also includes steps such as data cleaning, efficient user identification, session identification and path completion, and transaction identification.
2) Pattern Discovery: Different techniques such as association rules, clustering etc can be used for pattern discovery.

3) Pattern Analysis: This is the last step in the overall web usage mining process as described in Figure 2. The purpose of pattern analysis is to filter out uninteresting rules or patterns from the set found in the pattern discovery phase. The exact analysis methodology is usually governed by the application for which Web mining is done.

![Figure 2: High Level Web Usage Mining](image)

II. TECHNIQUES OF WEB USAGE MINING

There are various data mining techniques available which also can be applied to web data mining. Some of the techniques are listed below:

1) **Association Rules Mining:** The problem of deriving Association Rules from data was first formulated in (Agrawal, Imielinski and Swami, 1993) and is called the “market-basket problem”. The problem is that we are given a set of items and a large collection of transactions which are sets (baskets) of items. The task is to find relationships between the containments of various items within those baskets. Apart from the supermarket scenario there are many other examples where Association Rules have been used, for example users’ visits of WWW pages which the structure and its content can be optimized. The purpose of association rule mining is to find interesting associations and or correlation relationship among large set of data items. The two important measures of association rule mining are confidence and support.

2) **Sequential pattern mining:** Association rule mining does not take the time stamp into account. When we take time stamp into account we get more accurate & useful rules. To overcome above disadvantage of association rule mining, sequential pattern mining is used. Sequential pattern mining finds the relationship between occurrences of sequential events. Sequential pattern mining is mainly useful in time series data, it helps to get more accurate and useful prediction for business organisation. Use of Sequential pattern mining in web usage mining is to find sequential navigation patterns that appear in user sessions frequently.

For example: 70% of users who first visited A.html & then visited B.html afterwards, have also accessed C.html in same session.

3) **Classification:** In classification we build a model to classify a class of objects to predict the classification or missing attribute value of future objects (whose class may not be known). Classification is a two-step process:
   a) In the first process, a model is constructed to describe the characteristics of a set of data classes or concepts based on the collection of training data. Classification is a supervised learning since the data classes are predefined i.e. the training sample belongs to which class is provided.
   b) In the second step, the model is used to predict the classes of future objects or data.

4) **Clustering:** Clustering is another data mining technique which is regarded as unsupervised learning process since there is no predefined classes. Clustering is the process of grouping a set of physical or abstract objects into classes of similar objects, so that objects within the same cluster must be similar to some extent, also they should be dissimilar to those objects in other clusters. Use of clustering in web usage mining is to group together similar sessions of users.

For example: a departmental store, on the basis of different needs of customers can provide different items. In this case the store can provide a better service to its customers.
III. STEPS IN WEB USAGE MINING

1) Data Preparation
Since a web log stores all the user activities in accessing a website but of information contained in a raw web log does not represent a user session file[2] in a reliable and consistent way. This phase consist of following four main tasks:
a) Removing undesirable entries: Some of the user entries which a weblog contains may not be closely relevant to usage mining and can be removed without affecting the mining, for example: i) All log image entries. ii) Robot entries.
b) Distinguishing among Users: A user in web log mining is defined as a single individual which accesses files from web servers using browser. In order to understand the actual behavior of user, users in the log must be distinguished. Consider a sample log of a user’s different accesses to a website flipkart.com as shown in figure 3

\[ \text{FIGURE 3: SAMPLE WEBSITE STRUCTURE} \]

The access data from an IP address (103.4.253.46) recorded on the log are given in Table 1. The paths are found by Heuristics are home.php -- categories.php -- clothing.php -- clothing-men.php -- clothing-kid.php -- watches.php -- watches-kids.php

<table>
<thead>
<tr>
<th>clientip</th>
<th>reqDateTime</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>103.4.253.46</td>
<td>14/Oct/2013:02 :52:23 -0400</td>
<td>/home.php</td>
</tr>
<tr>
<td>103.4.253.46</td>
<td>14/Oct/2013:02 :52:23 -0400</td>
<td>/categories.php</td>
</tr>
<tr>
<td>103.4.253.46</td>
<td>14/Oct/2013:03 :52:23 -0400</td>
<td>/clothing.php</td>
</tr>
<tr>
<td>103.4.253.46</td>
<td>14/Oct/2013:03 :07:33 -0400</td>
<td>/software.php</td>
</tr>
<tr>
<td>103.4.253.46</td>
<td>14/Oct/2013:02 :08:21 -0400</td>
<td>/software-mobile.php</td>
</tr>
<tr>
<td>103.4.253.46</td>
<td>14/Oct/2013:02 :08:44 -0400</td>
<td>/categories.php</td>
</tr>
<tr>
<td>103.4.253.46</td>
<td>14/Oct/2013:02 :08:51 -0400</td>
<td>/clothing.php</td>
</tr>
<tr>
<td>103.4.253.46</td>
<td>14/Oct/2013:02 :08:55 -0400</td>
<td>/clothing-men.php</td>
</tr>
<tr>
<td>103.4.253.46</td>
<td>14/Oct/2013:02 :09:18 -0400</td>
<td>/categories.php</td>
</tr>
<tr>
<td>103.4.253.46</td>
<td>14/Oct/2013:02 :09:21 -0400</td>
<td>/Watches.php</td>
</tr>
<tr>
<td>103.4.253.46</td>
<td>14/Oct/2013:02 :09:24 -0400</td>
<td>/clothing-kid.php</td>
</tr>
</tbody>
</table>

c) Building Sessions (sessionization / episode identification): When a user visits a website for a long span of time, it is very likely that the user will visit the website for more than once, thus the purpose of this step is to divide user accesses of pages into individual sessions. For this purpose a time threshold is used.
Restoring the Contents of a Session: This task determines if there are important accesses that are not recorded in the access logs. For example, Web caching or using the back button will cause information discontinuance in logs.

2) Pattern Discovery/ Pattern Extraction
It deals with extracting interesting patterns from the preprocessed web logs. This is the key component of web usage mining. This can be subdivided into

**Statistical Analysis:**

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
</tr>
</thead>
</table>
| General statistics | 1) Total number of hits  
                   | 2) Total number of visitors  
                   | 3) Different errors  
                   | 4) Successful visits  
                   | 5) Incomplete visits  
                   | 6) Error reports |
| Access statistics | Request Hit and Miss count based on  
                   | 1) IP address  
                   | 2) URL |
| Periodical statistics | Access of web pages according to period of time e.g. daily, monthly, yearly. |

*Clustering:* clustering can be done according to i) Web pages, ii) Web page sequences, iii) client IP etc.
*Classification:* classify users according to their navigational behavior.
*Association Rules:* for example, thirty percent of categories page viewers will enter the software-dept pages.

IV. Applications of web usage mining

The results produced by the mining of web logs can be used for various purposes [4]:

1) To personalize the delivery of web content
2) To improve user navigation through prefetching and caching
3) To improve web design; or in e-commerce sites
4) To improve the customer satisfaction
5) Personalization of Web Content: Web Usage Mining techniques can be used to provide personalized web user experience. For instance, it is possible to anticipate, in real time, the user behavior by comparing the current navigation pattern with typical patterns which were extracted from past web log. 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 [6].

6) Prefetching 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 prefetching and caching strategies so as to reduce the server response time as done in [7].

7) Support to the 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 design of web applications [8].

V. Conclusion

The increasing demand of the Web has greatly evolved the Web mining technology. An important research area in Web mining is Web usage mining which focuses on the discovery of interesting patterns in the browsing and navigation data of Web users. This paper presents an overview of web usage mining. There are several techniques used for web usage mining proposed by various researchers, this paper discussed about various important techniques available for web usage mining. This paper also discusses about three viable steps in WUM such as preprocessing, pattern discovery and pattern analysis. If we shall be able to propose an efficient algorithm for the pattern extraction then it will help in the business of the website owners to understand their customer’s behavior properly so that they can fulfill their requirements.

References


