An Overview of Various approaches in Web Usage Mining

Authors
Prof. S.Kavitha¹, Prof. J.Umarani², Dr.S.Manikandan³

¹. Asst.Professor in dept. of Computer Applications, THRC, Perambalur, India.
   (Research scholar in Bharathiyar University, Coimbatore)
². Asst.Professor in dept. of Computer Applications, THRC, Perambalur, India.
   (Research scholar in Bharathiyar University, Coimbatore)
³. Head, Dept. of MCA, RMD Engineering College, Chennai, India. (Research Supervisor)

ABSTRACT
Web mining refers to discover useful information or knowledge from data on the web[1]. It is one of the applications of data mining. Web mining is used in various applications. Now a day’s student’s community utilizes web in lot of ways. Web provides needful essence to all. The main aim of web mining is to identify the user behavior. Web has evolved huge number of users and websites. Many of the users are school students, college students of post graduates and research scholars, Business persons, Academician The website developers can understand the need of their users using web data. In this paper we present web usage mining approaches in data mining.

Key words: Web mining, web usage mining, web personalization.

1. INTRODUCTION
World Wide Web provides lot of information based on the user request. It provides knowledge base, information retrieval, information search, etc. The web can give details by considering user’s need, user access pattern. In order to improve the quality of service web site developers understand web site user behavior. It is identified by web data[2]. Based on data used in the mining process, it is divided into: Web structure mining, Web content mining and Web usage mining. Web structure mining discovers knowledge from hyperlink structure of web. Web content mining extracts useful information from page contents. Web usage mining takes knowledge from usage patterns that people leave behind the page.
Table 1.1: Evaluation of Web Mining Approaches

<table>
<thead>
<tr>
<th>Web Mining Type</th>
<th>View of Data</th>
<th>Source</th>
<th>Collection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Usage</td>
<td>Interactivity</td>
<td>Access</td>
<td>Server/Browser logs</td>
</tr>
<tr>
<td>Content</td>
<td>Semi Structured, Unstructured</td>
<td>Pages</td>
<td>Text/Hypertext Documents</td>
</tr>
<tr>
<td>Structure</td>
<td>Link structure</td>
<td>Map</td>
<td>Link structure</td>
</tr>
</tbody>
</table>

2. WEB USAGE MINING

Web usage mining[5] is the application of data mining techniques to discover usage patterns from web data. WUM is further classified as:

Web server data: The user logs are collected by web server. It includes IP address, page reference and access time.

Application Server data: It is used to track various kinds of business events and log them in application server logs.

Application level data: New events are identified and logging can be turned on for them.

The mined data includes logs of users’ interactions with the web, web server logs, proxy server logs, browser logs and so on. Web server logs include information about access, referrer and agent. Server interaction with the user is referred as access information, referrer is about the referring page and agent is about the browser. Various tools are used to find out the routine information about the site, user. They are listed below in the following table.

2.1 Table: Tools to make out a website information

<table>
<thead>
<tr>
<th>Information</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of hits</td>
<td>Number of times each page visited</td>
</tr>
<tr>
<td>Visitors</td>
<td>Number of users came to the site</td>
</tr>
<tr>
<td>Referring website</td>
<td>URL of the site where the user came from</td>
</tr>
<tr>
<td>Referral web site</td>
<td>URL of the site where the user went</td>
</tr>
<tr>
<td>Entry point</td>
<td>The web site page the user entered from</td>
</tr>
<tr>
<td>Visitor time</td>
<td>The time and day of visit</td>
</tr>
<tr>
<td>Path analysis</td>
<td>A list of the path of pages the user took</td>
</tr>
<tr>
<td>IP address</td>
<td>Which part of world the user came from</td>
</tr>
</tbody>
</table>
We can also find which pages are most popular, what keywords are used to find the site, how the user came to the site, how the visits vary by hour, by day and by month.

An organization may design or redesign the web pages using various web usage mining applications. The following information can assist an enterprise to achieve the best visited page.

Curtail the paths to high visit pages.

Get rid of low visit pages.

Redecorate the pages including home page for easy navigation.

Spruce up some pages so that search engines can find them.

Add efficacy for advertising

2.1: WUM In Real-Time Applications

Web usage mining is used to collect the following information:

Path stretch across: The most commonly stretched across paths are identified.

Conversion rates: look-to-click, click-to-buy and basket-to-buy rates for each product are verified.

Impact of advertising: The banners that are pulling the most traffic are identified.

Impact of promotions: The promotions that produce most sales.

Web site design: The most frequently clicked links are identified.

Customer segmentation: customers are categorized based on their links.

Enterprise search: Are customers likely to purchase?

Advanced data mining methods are utilized in the context of web usage mining. Such methods include:

Association rule mining, sequential pattern discovery, clustering and classification.

2.2: Association rule mining is used to identify frequent patterns, associations and correlations among sets of items. It is also used to know correlations between pages accessed together during a server session even if they are not directly connected.

2.3: Sequential Patterns are mainly used to consider the order of transactions. In many applications orderings are significant. In market basket analysis, it is interesting to know whether people buy some items in sequence.

2.4: Clustering groups together items that have similar characteristics. The two types are user clusters and page clusters. User clusters group users that behave similarly when browsing through the web. Page clustering identifies group of pages that are conceptually related in the view of users.

2.5: Classification maps a data item into one of several predetermined classes. classes represents different user profiles and classifications is performed using selected features that describe each user’s category.
3. OVERALL ARCHITECTURE OF WEB MINING PROCESS

3.1: Identifying The User
Web usage mining is a three-phase process consisting of data preparation, pattern discovery and pattern analysis. To identify user, web data are processed. The output data is the hits registered in web usage logs. Web usage logs contain much more information. We need to retrieve required information. The unwanted information must be eliminated. These will be meaningless if they are not processed efficiently. In pattern discovery phase data mining methods are applied in order to detect interesting patterns. The patterns are analyzed in the third phase.

3.2: Web Usage Logs
Access log records each access to a web page by the user.
The common log format:
Remotehost rm123 authuser date “req” status bytes
Remotehost – name of the remotehost or IP number if hostname is not available
Rm123 – remote log name of the user
Authuser – the username with which the user has authenticated himself
Date – date and time of request
Request – request line came from the client
Status – HTTP status code returned to the client
Bytes – content-length of the documents transferred
The extended web server log file format provided by W3C permits customized log files to be recorded in a format readable by generic analysis tools. The most important fields in the log file are:
‘Referrer’ specifies URL of the client
‘user agent’ is the software the client claims to be using.
‘cookie’ the text file that has user information.
4. PROCESSING OF DATA

Web log data[7] is preprocessed to use them in the following phases of the process.

The first step in the preprocessing is data preparation.

4.1: Data preparation

Web log data must be free of problems from entries involving pages that returned an error or graphics file accesses. User activity can be filtered out, because such entries don’t provide useful information about the site. When the user accesses the cached page that information is not recorded in the log file, caching is depending on the client-side technologies.

User identification is very important for all. The individual visitor can be identified by getting IP address of a single visitor. It is not very accurate because the visitor may access the web from different computers. Many users may use same IP address. If the user is identified then perform Session Identification.

4.2: Pattern Discovery

4.2.1: Traffic analysis Tools

It takes input from web data. It extracts useful information that comprises of hits. Site activity(bustle) such as number of visits, average number of hits, successful/failed/redirected/cached hits, average view time, average length of a site, diagnostic statistics (such as server errors, and page not found errors), server statistics (such as top pages visited, entry/exit pages, and single access pages), referrers statistics (such as top referring sites, search engines, and keywords), user demographics (such as top geographical location, and most active countries/cities/organizations), client statistics (visitor’s Web browser, operating system, and cookies), and so on.

5. WEB PERSONALIZATION

Web personalization[3] is termed as any action that adapts the web site to the needs of a user or set of users, taking information from user navigational behaviors and individual interests. The steps of web personalization includes:

The collection of web data
categorization of these data
the analysis of collected data
determination of action
The ways that are employed in order to analyze the collected data include content-based filtering, collaborative filtering, rule-based filtering and Web usage mining. The site is personalized through the highlighting of existing hyperlinks, the dynamic insertion of new hyperlinks.

Content-based filtering systems are solely based on individual users’ preferences. The system tracks each user’s behaviour and recommends them items that are similar to items the user liked in the past. Collaborative filtering systems invite users to rate objects or reveal their preferences and interests and then return information that is predicted to be of interest for them. This is based on the assumption that users with similar behavior (for example users that rate similar objects) have analogous interests.

In rule-based filtering the users are asked to answer to a set of questions. These questions are derived from a decision tree, so as the user proceeds on answering them, what she/he finally receives as a result (for example a list of products) is tailored to their needs. Content-based, rule-based and collaborative filtering may also be used in combination, for deducing more accurate conclusions.

6. CONCLUSION

In this paper we are giving various approaches in web mining. WUM is one of the tool to identify user behavior and user interests. To reconstruct a web site or page the web site developer must follow some rules based on the users navigational patterns. Web usage logs tell information about the user interests. Web personalization is another important factor to attract much more users to the website.
REFERENCES


A Survey on Preprocessing Methods for Web Usage Data V.Chitraa, Dr. Antony Selvdoss Davamani


Web Page Classification Using Data Mining Keyur J. Patel1, Ketan J Sarvakar2