Chapter 2
Analysis of User's Browsing Behavior and Their Categorization Using Markov Chain Model

Ratnesh Kumar Jain
Kendriya Vidyalaya, India

Rahul Singhai
Devi Ahilya University, India

ABSTRACT

Web server log file contains information about every access to the web pages hosted on a server like when they were requested, the Internet Protocol (IP) address of the request, the error code, the number of bytes sent to the user, and the type of browser used. Web servers can also capture referrer logs, which show the page from which a visitor makes the next request. As the visit to web site is increasing exponentially the web logs are becoming huge data repository which can be mined to extract useful information for decision making. In this chapter, we proposed a Markov chain based method to categorize the users into faithful, Partially Impatient and Completely Impatient user. And further, their browsing behavior is analyzed. We also derived some theorems to study the browsing behavior of each user type and then some numerical illustrations are added to show how their behavior differs as per categorization. At the end we extended this work by approximating the theorems.

INTRODUCTION

The discovery and analysis of useful information from the World Wide Web is called *web mining*, or Data Mining efforts associated with web is Web Mining. Web mining term can be used in two different ways, Web Content Mining & web Usage Mining. The first one describes the automatic search of information from the resources that are available on-line i.e. it is the process of information discovery from the sources across World Wide Web, and Web Usage Mining is the process of mining user browsing and access patterns while a web site is visited.

DOI: 10.4018/978-1-5225-0613-3.ch002

DATA PRE-PROCESSING WHILE MINING WEB DATA

One of the important core steps of knowledge discovery is *data pre processing*. The main goal is to create minable objects for knowledge discovery despite the presence of ambiguities and incompleteness in data. Pre-processing consists of converting the usage, content, and structure information contained in the various available data sources into the data abstractions necessary for pattern discovery.

Data Pre-Processing in Content Mining

Web content mining is strongly related to the domain of Text Mining, since in order to process and organize Web pages and their content should be first appropriately processed in order to extract patterns of interest. These selected properties are subsequently used to represent the documents and assist the clustering or classification processes.

Another essential issue during this stage is semantic analysis. Semantic analysis deals mainly with the problems of synonymy (different names for the same concept) and polysemy (different concepts having the same name). Research on the area of "Word Sense Disambiguation" (WSD) has dealt with this problem. Word sense disambiguation is achieved by assigning words to appropriate concepts. The mapping from words to concepts should be done in a reliable way, depending on the relations between words under examination.

Data Pre-Processing in Web Usage Mining

There are some important technical issues that must be taken into consideration during this phase in the context of the Web personalization process. It is necessary for Web log data to be prepared and pre-processed in order to use in the consequent phases of the process. The first issue is data preparation, depending on the application, Web log data may need to be cleaned from entries involving pages that returned an error or graphics file accesses. Furthermore, crawler activity can be filtered out, because such entries do not provide useful information about the site's usability. Another problem is with caching. Accesses to cached pages are not recorded in the Web log, therefore such information is missed. Caching is heavily dependent on the client-side technologies used and therefore cannot be dealt with ease. In such cases, cached pages can usually be inferred using the referring information from the logs. Moreover, a useful aspect is to perform page view identification, determining which page file accesses contribute to a single page view.

Most important of all is the user identification issue. There are several ways to identify individual visitors. The most obvious solution is to assume that each IP address (or each IP address/client agent pair) identifies a single visitor. Nonetheless, this is not very accurate because, for example, a visitor may access the Web from different computers, or many users may use the same IP address. Again, a user that uses more than one browser, even on the same machine, will appear as multiple users. A further assumption can then be made, that consecutive accesses from the same host during a certain time interval come from the same user. So, identification of user's browsing behavior is also one important problem during data pre-processing. More accurate approaches for a priori identification of unique visitors are the use of cookies or similar mechanisms or the requirement for user registration. However, a potential problem in using such methods might be the reluctance of users to share personal information. Assuming a user is identified, the next step is to perform session identification. Before any mining is done on web usage