Web Usage Mining and Its Applications

Yongjian Fu
Cleveland State University, USA

INTRODUCTION

With the rapid development of the World Wide Web or the Web, many organizations now put their information on the Web and provide Web-based services such as online shopping, user feedback, technical support, and so on. Understanding Web usage through data mining techniques is recognized as an important area.

Web usage mining is the process to identify interesting patterns from Web server logs. It has shown great potentials in many applications such as adaptive Web sites, Web personalization, cache management, and so on.

BACKGROUND

Most commonly used Web servers maintain a server log, which consists of page requests in the form of Common Log Format. The Common Log Format specifies that a record in a log file contains, among other data, the IP address of the user, the date and time of the request, the URL of the page, the protocol, the return code of the server, and the size of the page if the request is successful (Luotonen, 1995).

A few examples of log records in Common Log Format are given in *Table 1*. The IP addresses are modified for privacy reasons. The URLs of the pages are relative to the Web server's home page address, in this example, www.csuohio.edu.

In Web usage mining, the server logs are first preprocessed to clean and transform the data. Data mining techniques are then applied on these preprocessed data to find usage patterns. The usage patterns are employed by many applications to evaluate and improve Web sites.

In preprocessing, the server log files are cleaned to filter out irrelevant information for Web usage mining, such as background images, and transformed into a set of sessions. A session is conceptually a single visit of a user (Cooley et al., 1999). For example, when a user buys an airplane ticket from a Web site, the log records related to the transaction compose a session. In practice, a session consists of pages accessed by a user in a certain period of time.

Various data mining techniques can be applied on sessions to find usage patterns, including association rules, clustering, and classification. Other techniques have also been used for Web usage analysis including data warehousing and OLAP, intelligent agent, and collaborative filtering.

Web usage mining has a broad range of applications, such as adaptive Web sites, Web personalization, and cache management, to name a few. Moreover, Web usage patterns may be combined with other information such as Web page content (texts and multimedia), hyperlinks, and user registrations to provide more comprehensive understandings and solutions.

MAIN THRUST

The preprocessing of Web server logs, techniques for web usage mining, and applications of web usage mining are discussed.

Preprocessing

In preprocessing, irrelevant records in a server log are thrown out and others are put into sessions. Log records from the same user are put into a session. The IP addresses in the log records are used to identify users. Two records with the same IP address are assumed from the same user. A session contains a unique session ID and a set of (*pid*,

Table 1. Examples from a Web server log

dan.ece.csuohio.edu -- [01/Aug/2001:13:17:45 -0700] "GET /~dan/a.html" 200 34
131.39.170.27 -- [01/Aug/2001:13:17:47 -0700] "GET /~white/Home.htm HTTP/1.0" 200 2034
dan.ece.csuohio.edu -- [01/Aug/2001:13:17:48 -0700] "GET /~dan/b.html HTTP/1.0" 200 8210
131.39.170.27 -- [01/Aug/2001:13:17:50 -0700] "GET /~white/cloud.gif HTTP/1.0" 200 4489
131.39.170.27 -- [01/Aug/2001:13:17:51 -0700] "GET /~white/hobby.htm HTTP/1.0" 200 890
117.83.344.74 -- [01/Aug/2001:13:17:51 -0700] "GET /~katz/arrow.jpg HTTP/1.0" 200 2783

t) pairs, where *pid* is a page identifier and *t* is the time the user spent on that page.

Generally, the preprocessing involves the following steps (Cooley et al., 1999):

1. Records about image files (.gif, .jpg, etc) are filtered as well as unsuccessful requests (return code not 200).
2. Requests from the same IP address are grouped into a session. A timeout threshold max_idle is used to decide the end of a session, that is, if the same IP address does not occur within a time range of max_idle minutes, the current session is closed. Subsequent requests from the same IP address will be treated as a new session.
3. The time spent on a particular page is determined by the time difference between two consecutive requests.

The introduction of max_idle is for both conceptual and practical purposes. From a conceptual point, it helps to limit a session to a single visit. For instance, a user can buy a book and comes back the next day to check movies. The activities will be separated into two sessions. From a practical point, it prevents a session from running too long. The selection of max_idle is dependent on the Web site and application. Empirically, a few studies found 30 minutes to be suitable (Cooley et al., 1999; Fu et al., 1999).

For example, the Web server log in *Table 1* will be organized into sessions as shown in *Table 2*. It should be noted that session IDs are not IP addresses since they may be shared by multiple sessions.

There are some difficulties in accurately identifying sessions and estimating times spent on pages, due to client or proxy caching of pages, sharing of IP addresses, and network traffic (Cooley et al., 1999). Besides, the time the user spent on the last page is unknown since there are no more requests after it.

Techniques

Several data mining techniques have been successfully applied in Web usage mining, including association rules, clustering, and classification. Besides, data warehousing and OLAP techniques have also been employed.

Association rules represent correlations among objects, first proposed to capture correlations among items in transactional data. For example, an association rule "hot dog $\rightarrow$ soft drink [10%, 56%]" says that 56% of people who buy hot dogs also buy soft drinks, which constitute 10% of all customers. If a session is viewed as a transaction, association rule mining algorithms can be employed to find associative relations among pages browsed (Yang et al., 2002). For example, an association rule "Page A $\rightarrow$ Page B [5%, 80%]" says 80% of users who browse page A will also browse page B, and 5% of all users browse both. Using the same algorithms, we may find frequent paths traversed by many users, for example, 40% of users browsing page A, then pages B and C, finally page D (Frias-Martinez & Karamcheti, 2002).

Clustering is the identification of classes, also called clusters or groups, for a set of objects whose classes are unknown. Using clustering techniques, we can cluster users based on their access patterns (Fu et al., 1999). In this approach, sessions are treated as objects and each page representing a dimension in the object space. Sessions containing similar pages will be grouped. For examples, if a user browses pages A, B, C, and D, and another user browses pages A, C, D, and F, they may be clustered in to a group. A more sophisticated clustering approach would use the browsing times of pages in sessions. For example, two sessions [1, (A, 15), (B, 10), (C, 1)] and [2, (A, 12), (B, 12), (D, 2)] will be clustered into one group.

In classification, a classifier is developed from a training set of objects where classes are known. Given a set of sessions in different classes, a classifier can be built using classification methods. For example, a classifier may tell whether a user will be a buyer or a non-buyer based on the browsing patterns of the user for an e-commerce site (Spiliopoulou et al., 1999).

Data warehouse techniques may be used to create data cubes from Web server logs for OLAP. The statistics along pages, IP domains, geographical locations of users, and browsing times are calculated from sessions.

Other techniques exist for Web usage mining. For example, a hybrid method, which combines hypertext probabilistic grammar and click fact table, shows promising results (Jespersen et al., 2002).

Table 2. Sessions from the server logs

Session ID	IP Address	Requested Page	Time Spent
1	dan.ece.csuohio.edu	/~dan/a.html	3 seconds
		/~dan/b.html	
2	131.39.170.27	/~white/Home.htm	4 seconds
		/~white/hobby.htm	