Web Usage Mining

Bamshad Mobasher
DePaul University, USA

INTRODUCTION

With the continued growth and proliferation of e-commerce, Web services, and Web-based information systems, the volumes of clickstream and user data collected by Web-based organizations in their daily operations have reached astronomical proportions. Analyzing such data can help these organizations determine the lifetime value of clients, design cross-marketing strategies across products and services, evaluate the effectiveness of promotional campaigns, optimize the functionality of Web-based applications, provide more personalized content to visitors, and find the most effective logical structure for their Web space. This type of analysis involves the automatic discovery of meaningful patterns and relationships from a large collection of primarily semi-structured data often stored in Web and applications server access logs as well as in related operational data sources.

Web usage mining (Cooley et al., 1997; Srivastava et al., 2000) refers to the automatic discovery and analysis of patterns in clickstream and associated data collected or generated as a result of user interactions with Web resources on one or more Web sites. The goal of Web usage mining is to capture, model, and analyze the behavioral patterns and profiles of users interacting with a Web site. The discovered patterns usually are represented as collections of pages, objects, or resources that are frequently accessed by groups of users with common needs or interests.

The overall Web usage mining process can be divided into three interdependent tasks: data preprocessing, pattern discovery, and pattern analysis or application. In the preprocessing stage, the clickstream data are cleaned and partitioned into a set of user transactions representing the activities of each user during different visits to the site. In the pattern discovery stage, statistical, database, and machine learning operations are performed to obtain possibly hidden patterns reflecting the typical behavior of users, as well as summary statistics on Web resources, sessions, and users. In the final stage of the process, the discovered patterns and statistics are further processed, filtered, and used as input to applications, such as recommendation engines, visualization tools, and Web analytics and report generation tools.

In this article, we provide a summary of the analysis and data-mining tasks most commonly used in Web usage mining and discuss some of their typical applications.

BACKGROUND

The log data collected automatically by the Web and application servers represent the fine-grained navigational behavior of visitors. Each hit against the server generates a single entry in the server access logs. Each log entry (depending on the log format) may contain fields identifying the time and date of the request, the IP address of the client, the resource requested, possible parameters used in invoking a Web applications, status of the request, HTTP method used, the user agent (browser and operating system types and versions), the referring Web resource, and, if available, client-side cookies that uniquely identify repeat visitors.

Depending on the goals of the analysis, these data need to be transformed and aggregated at different levels of abstraction. In Web usage mining, the most basic level of data abstraction is that of a *pageview*. A pageview is an aggregate representation of a collection of Web objects contributing to the display on a user's browser resulting from a single user action (such as a clickthrough). At the user level, the most basic level of behavioral abstraction is that of a *session*. A session is a sequence of pageviews by a single user during a single visit. The process of transforming the preprocessed clickstream data into a collection of sessions is called *sessionization*.

The goal of the preprocessing stage in Web usage mining is to transform the raw clickstream data into a set of user sessions, each corresponding to a delimited sequence of pageviews (Cooley et al., 1999). The sessionized data can be used as the input for a variety of data-mining algorithms. However, in many applications, data from a variety of other sources must be integrated with the preprocessed clickstream data. For example, in e-commerce applications, the integration of both customer and product data (e.g., demographics, ratings, purchase histories) from operational databases with usage data can allow for the discovery of important business intelligence metrics, such as customer conversion ratios and lifetime values (Kohavi et al., 2004). The integration of semantic knowledge from the site content or semantic attributes of products can be used by personalization systems to provide more useful recommendations (Dai & Mobasher, 2004; Gahni & Fano, 2003).

A detailed discussion of the data preparation and data collection in Web usage mining can be found in the article

"Data Preparation for Web Usage Mining" in this volume (Mobasher, 2005).

MAIN THRUST

The types and levels of analysis performed on the integrated usage data depend on the ultimate goals of the analyst and the desired outcomes. This section describes the most common types of pattern discovery and analysis employed in the Web usage mining domain and discusses some of their applications.

Session and Visitor Analysis

The statistical analysis of preprocessed session data constitutes the most common form of analysis. In this case, data are aggregated by predetermined units, such as days, sessions, visitors, or domains. Standard statistical techniques can be used on these data to gain knowledge about visitor behavior. This is the approach taken by most commercial tools available for Web log analysis. Reports based on this type of analysis may include information about most frequently accessed pages, average view time of a page, average length of a path through a site, common entry and exit points, and other aggregate measures. Despite a lack of depth in this type of analysis, the resulting knowledge potentially can be useful for improving system performance and providing support for marketing decisions. Furthermore, commercial Web analytics tools are increasingly incorporating a variety of data-mining algorithms resulting in more sophisticated site and customer metrics.

Another form of analysis on integrated usage data is Online Analytical Processing (OLAP). OLAP provides a more integrated framework for analysis with a higher degree of flexibility. The data source for OLAP analysis is usually a multidimensional data warehouse, which integrates usage, content, and e-commerce data at different levels of aggregation for each dimension. OLAP tools allow changes in aggregation levels along each dimension during the analysis. Analysis dimensions in such a structure can be based on various fields available in the log files and may include time duration, domain, requested resource, user agent, and referrers. This allows the analysis to be performed on portions of the log related to a specific time interval or at a higher level of abstraction with respect to the URL path structure. The integration of e-commerce data in the data warehouse further can enhance the ability of OLAP tools to derive important business intelligence metrics (Buchner & Mulvenna, 1999). The output from OLAP queries also can be used as the input for a variety of data-mining or data visualization tools.

Association and Correlation Analysis

Association rule discovery and statistical correlation analysis on usage data result in finding groups of items or pages that are commonly accessed or purchased together. This, in turn, enables Web sites to organize the site content more efficiently or to provide effective cross-sale product recommendations.

Association rule discovery algorithms find groups of items (e.g., pageviews) occurring frequently together in many transactions (i.e., satisfying a pre-specified minimum support threshold). Such groups of items are referred to as *frequent itemsets*. Association rules that satisfy a minimum confidence threshold are then generated from the frequent itemsets. An association rule r is an expression of the form $X \rightarrow Y (\sigma_r, \alpha_r)$, where X and Y are itemsets, σ is the support of the itemset $X \cup Y$ representing the probability that X and Y occur together in a transaction, and α is the confidence for the rule r, representing the conditional probability that Y occurs in a transaction, given that X has occurred in that transaction.

The discovery of association rules in Web transaction data has many advantages. For example, a high-confidence rule, such as {special-offers/, /products/software/} $\rightarrow$ {shopping-cart/}, might provide some indication that a promotional campaign on software products is positively affecting online sales. Such rules also can be used to optimize the structure of the site. For example, if a site does not provide direct linkage between two pages A and B, the discovery of a rule {A} $\rightarrow$ {B} would indicate that providing a direct hyperlink from A to B might aid users in finding the intended information. Both association analysis (among products or pageviews) and statistical correlation analysis (generally among customers or visitors) have been used successfully in Web personalization and recommender systems (Herlocker et al., 2004; Mobasher et al., 2001).

Cluster Analysis and Visitor Segmentation

Clustering is a data-mining technique to group together a set of items having similar characteristics. In the Web usage domain, there are two kinds of interesting clusters that can be discovered: user cluster and page clusters.

Clustering of user records (sessions or transactions) is one of the most commonly used analysis tasks in Web usage mining and Web analytics. Clustering of users tends to establish groups of users exhibiting similar browsing patterns. Such knowledge is especially useful for inferring user demographics in order to perform market segmentation in e-commerce applications or to