Web Usage Mining Data Preparation

Bamshad Mobasher
DePaul University, USA

INTRODUCTION

Web usage mining 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. 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.

An important task in any data-mining application is the creation of a suitable target dataset to which data mining and statistical algorithms are applied. This is particularly important in Web usage mining due to the characteristics of clickstream data and its relationship to other related data collected from multiple sources and across multiple channels. The data preparation process is often the most time-consuming and computationally-intensive step in the Web usage mining process and often requires the use of special algorithms and heuristics not commonly employed in other domains. This process is critical to the successful extraction of useful patterns from the data. This process may involve preprocessing the original data, integrating data from multiple sources, and transforming the integrated data into a form suitable for input into specific data-mining operations. Collectively, we refer to this process as *data preparation*.

In this article, we summarize the essential tasks and requirements for the data preparation stage of the Web usage mining process.

BACKGROUND

The primary data sources used in Web usage mining are the server log files, which include Web server access logs and application server logs. Additional data sources that are also essential for both data preparation and pattern discovery include the site files and meta-data, operational databases, application templates, and domain knowledge. In some cases and for some users, additional data may be available due to client-side or proxy-level (Internet service provider) data collection, as well as from external clickstream or demographic data sources (e.g., ComScore, NetRatings, MediaMetrix, and Acxiom).

Much of the research and practice in usage data preparation has focused on preprocessing and integrating these data sources for different types of analyses. Usage data preparation presents a number of unique challenges that have led to a variety of algorithms and heuristic techniques for preprocessing tasks, such as data fusion and cleaning, user and session identification, pageview identification (Cooley et al., 1999). The successful application of data-mining techniques to Web usage data is highly dependent on the correct application of the preprocessing tasks. Furthermore, in the context of e-commerce data analysis and Web analytics, these techniques have been extended to allow for the discovery of important and insightful user and site metrics (Kohavi et al., 2004).

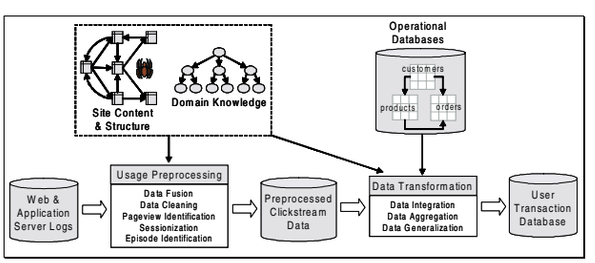
Figure 1 provides a summary of the primary tasks and elements in usage data preprocessing. We begin by providing a summary of data types commonly used in Web usage mining and then provide a brief discussion of some of the primary data preparation tasks.

The data obtained through various sources can be categorized into four primary groups (Cooley et al., 1999; Srivastava et al., 2000).

Usage Data

The log data collected automatically by the Web and application servers represents the fine-grained navigational behavior of visitors. Each hit against the server, corresponding to an HTTP request, generates a single

Figure 1. Summary of data preparation tasks for Web usage mining



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 application, status of the request, HTTP method used, the user agent (browser and operating system type and version), the referring Web resource, and, if available, client-side cookies that uniquely identify a repeat visitor. Depending on the goals of the analysis, the 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 click-through). Conceptually, each pageview can be viewed as a collection of Web objects or resources representing a specific user event (e.g., reading an article, viewing a product page, or adding a product to the shopping cart. 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 notion of a session can be abstracted further by selecting a subset of pageviews in the session that is significant or relevant for the analysis tasks at hand.

Content Data

The content data in a site are the collection of objects and relationships that are conveyed to the user. For the most part, these data are comprised of combinations of textual material and images. The data sources used to deliver or generate this data include static HTML/XML pages, multimedia files, dynamically generated page segments from scripts, and collections of records from the operational databases. The site content data also include semantic or structural meta-data embedded within the site or individual pages, such as descriptive keywords, document attributes, semantic tags, or HTTP variables. The underlying domain ontology for the site also is considered part of the content data. Domain ontologies may include conceptual hierarchies over page contents, such as product categories, explicit representations of semantic content and relationships via an ontology language such as RDF, or a database schema over the data contained in the operational databases.

Structure Data

The structure data represent the designer's view of the content organization within the site. This organization is captured via the inter-page linkage structure among pages, as reflected through hyperlinks. The structure data also include the intra-page structure of the content within a page. For example, both HTML and XML documents can be represented as tree structures over the space of tags in the page. The hyperlink structure for a site normally is captured by an automatically generated site map. A site-mapping tool must have the capability to capture and represent the inter- and intra-pageview relationships. For dynamically generated pages, the site-mapping tools either must incorporate intrinsic knowledge of the underlying applications and scripts or must have the ability to generate content segments using a sampling of parameters passed to such applications or scripts.

User Data

The operational database(s) for the site may include additional user profile information. Such data may include demographic information about registered users, user ratings on various objects such as products or movies, past purchase or visit histories of users, as well as other explicit or implicit representations of a user's interests. Some of these data can be captured anonymously, as long as there is the ability to distinguish among different users. For example, anonymous information contained in client-side cookies can be considered part of the users' profile information and can be used to identify repeat visitors to a site. Many personalization applications require the storage of prior user profile information.

MAIN THRUST

As noted in Figure 1, the required high-level tasks in usage data preprocessing include the fusion and synchronization of data from multiple log files, data cleaning, pageview identification, user identification, session identification (or sessionization), episode identification, and the integration of clickstream data with other data sources, such as content or semantic information, as well as user and product information from operational databases.

Data fusion refers to the merging of log files from several Web and application servers. This may require global synchronization across these servers. In the absence of shared embedded session ids, heuristic methods based on the referrer field in server logs, along with various sessionization and user identification methods (see following), can be used to perform the merging. This step is essential in inter-site Web usage mining, where the analysis of user behavior is performed over the log files for multiple related Web sites (Tanasa & Trousse, 2004).

Data cleaning is usually site-specific and involves tasks such as removing extraneous references to embedded objects, style files, graphics, or sound files, and removing references due to spider navigations. The latter task can be performed by maintaining a list of known