World Wide Web Usage Mining

Wen-Chen Hu
University of North Dakota, USA

Hung-Jen Yang
National Kaohsiung Normal University, Taiwan

Chung-wei Lee
Auburn University, USA

Jyh-haw Yeh
Boise State University, USA

INTRODUCTION

World Wide Web data mining includes content mining, hyperlink structure mining, and usage mining. All three approaches attempt to extract knowledge from the Web, produce some useful results from the knowledge extracted, and apply the results to certain real-world problems. The first two apply the data mining techniques to Web page contents and hyperlink structures, respectively. The third approach, Web usage mining (the theme of this article), is the application of data mining techniques to the usage logs of large Web data repositories in order to produce results that can be applied to many practical subjects, such as improving Web sites/pages, making additional topic or product recommendations, user/customer behavior studies, and so forth. This article provides a survey and analysis of current Web usage mining technologies and systems. A Web usage mining system must be able to perform five major functions: (i) data gathering, (ii) data preparation, (iii) navigation pattern discovery, (iv) pattern analysis and visualization, and (v) pattern applications. Many Web usage mining technologies have been proposed, and each technology employs a different approach. This article first describes a generalized Web usage mining system, which includes five individual functions. Each system function is then explained and analyzed in detail. Related surveys of Web usage mining techniques also can be found in Hu, et al. (2003) and Kosala and Blockeel (2000).

BACKGROUND

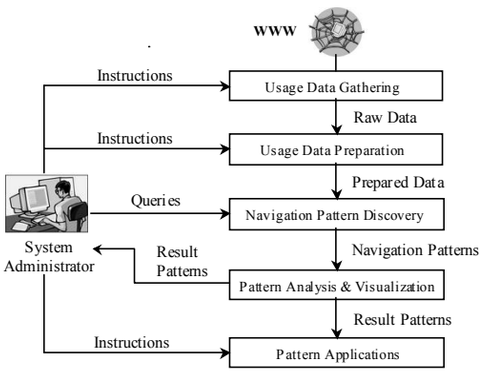
A variety of implementations and realizations is employed by Web usage mining systems. This section introduces the Web usage mining background by giving a generalized structure of the systems, each of which carries out five major tasks:

- **Usage Data Gathering:** Web logs, which record user activities on Web sites, provide the most comprehensive, detailed Web usage data.
- **Usage Data Preparation:** Log data are normally too raw to be used by mining algorithms. This task restores the user's activities that are recorded in the Web server logs in a reliable and consistent way.
- **Navigation Pattern Discovery:** This part of a usage mining system looks for interesting usage patterns contained in the log data. Most algorithms use the method of sequential pattern generation, while the remaining methods tend to be rather ad hoc.
- **Pattern Analysis and Visualization:** Navigation patterns show the facts of Web usage, but these require further interpretation and analysis before they can be applied to obtain useful results.
- **Pattern Applications:** The navigation patterns discovered can be applied to the following major areas, among others: (i) improving the page/site design, (ii) making additional product or topic recommendations, and (iii) Web personalization.

Figure 1 shows a generalized structure of a Web usage mining system; the five components will be detailed in the next section. A usage mining system also can be divided into the following two types:

- **Personal:** A user is observed as a physical person for whom identifying information and personal data/properties are known. Here, a usage mining system optimizes the interaction for this specific individual user.
- **Impersonal:** The user is observed as a unit of unknown identity, although some properties may be accessible from demographic data. In this case, a usage mining system works for a general population.

Figure 1. A Web usage mining system structure



This article concentrates on the impersonal systems. Personal systems actually are a special case of impersonal systems, so readers can easily infer the corresponding personal systems, given the information for impersonal systems.

MAIN THRUST OF THE ARTICLE

This section details the five major functions of a Web mining system: (i) data gathering, (ii) data preparation, (iii) navigation pattern discovery, (iv) pattern analysis and visualization, and (v) pattern applications.

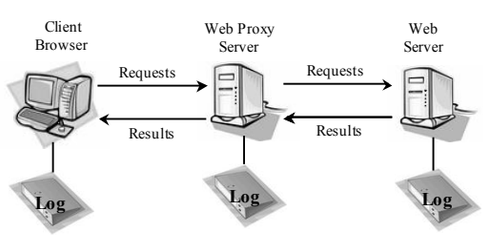
Data Gathering

Web usage data are usually supplied by two sources: trial runs by humans and Web logs. The first approach is impractical and rarely used because of the nature of its high time and expense costs and its bias. Most usage mining systems use log data as their data source. This section looks at how and what usage data can be collected.

Web Logs

A Web log file records activity information when a Web user submits a request to a Web server. A log file can be located in three different places: (i) Web servers, (ii) Web proxy servers, and (iii) client browsers, as shown in Figure 2.

Figure 2. Three Web log file locations



- **Server-Side Logs:** These logs generally supply the most complete and accurate usage data.
- **Proxy-Side Logs:** A proxy server takes the HTTP requests from users and passes them to a Web server; the proxy server then returns to users the results passed to them by the Web server.
- **Client-Side Logs:** Participants remotely test a Web site by downloading special software that records Web usage or by modifying the source code of an existing browser. HTTP cookies also could be used for this purpose. These are pieces of information generated by a Web server and stored in the user's computer, ready for future access.

Web Log Information

A Web log is a file to which the Web server writes information each time a user requests a resource from that particular site. Examples of the types of information the server preserves include the user's domain, subdomain, and host name; the resources the user requested (e.g., a page or an image map); the time of the request; and any errors returned by the server. Each log provides different and various information about the Web server and its usage data. Most logs use the format of a common log file or extended log file. For example, the following is an example of a file recorded in the extended log format:

#Version: 1.0 #Date: 12-Jan-1996 00:00:00 #Fields: time cs-method cs-uri 00:34:23 GET /foo/bar.html 12:21:16 GET /foo/bar.html 12:45:52 GET /foo/bar.html 12:57:34 GET /foo/bar.html

Data Preparation

The information contained in a raw Web server log does not reliably represent a user session file. The Web usage data preparation phase is used to restore users' activities in the Web server log in a reliable and consistent way. At a minimum, this phase should achieve the following four major tasks: (i) removing undesirable entries, (ii) distinguishing among users, (iii) building sessions, and (iv) restoring the contents of a session (Cooley, Mobasher & Srivastava, 1999).

Removing Undesirable Entries

Web logs contain user activity information, of which some is not closely relevant to usage mining and can be