INTRODUCTION

In this paper, we will discuss research efforts devoted to the remaining area of Web mining, namely Web usage mining. Taken together, a complete picture of the trends in Web mining can be discerned.

BACKGROUND

Web mining is a fast developing area using data mining techniques to discover useful knowledge from Web documents and services (Etzioni, 1996; Kosala & Blockeel, 2000). Based on the types of data available on the Web, Web mining is generally divided into three categories: Web content mining, Web structure mining, and Web usage mining (Srivastava, Cooley, Deshpande, & Tan, 2000).

Both content mining and structure mining work with idealized static representations of the Web, i.e., the pages and links as they exist at a particular moment. The information discovered from content mining and structure mining is instrumental to the development of more powerful and intelligent search engines or agents (Glover, Tsioutsioulakis, Lawrence, Pennock, & Flake, 2002; Leake & Scherle, 2001). Web usage mining, on the other hand, is the discovery of useful information from users’ usage patterns. The data required to build a complete usage pattern is scattered across Web logs, application server logs, ad server logs, commerce server logs, product databases, and customer databases owned by a host of different organizations. Many of them have neither the ability nor the willingness to share the information they own. Furthermore, pages viewed by users through caching at client or proxy servers will not be recorded in the server logs, thus affecting the accuracy of server logs’ data. In addition, users are reluctant to let their Web activities be monitored due to privacy, security, and profiling concerns. The dynamic, diverse and incomplete nature of the usage data present a challenge to Web usage mining. However, as explained in the next section, a significant amount of Web usage mining research has been conducted despite the difficulty of working with an incomplete source of usage data.

MAIN THRUST

Web usage mining is the application of data mining techniques to discover usage patterns from Web data in order to understand and better serve the needs of Web-based applications (Srivastava, Cooley, Deshpande, & Tan, 2000). Usage data can be collected from three sources: Web servers, proxy servers, and Web clients.

Server access logs contain information about the name and IP address of the remote host, date and time of a user’s request, the URL of the Web page requested, size of the page requested, as well as status of the request that help characterize a user’s access to a specific Web server. Proxy server logs reveal actual requests from multiple clients to multiple Web servers served by that proxy server. The information is useful for learning the browsing behavior of a group of anonymous users sharing a common proxy server so that future page requests can be predicted to improve proxy caching services. Client side data provides detailed information about an actual user’s browsing activities. A Web client’s usage data is tracked by a remote agent deployed via JavaScript, Java applet, or modified browser. The data collected from these data sources is then used to construct data abstractions of users, user sessions, episodes, click-stream behaviors, and page views. Data abstractions are necessary for discovering usage patterns that range from single-user navigation patterns, single-site browsing patterns to multi-user, multi-site access patterns.

Usage patterns discovered have been critical for applications such as personalization, Web server performance improvement, Web site modification, and customer relationship management (Facca & Lanzi, 2003).

Web usage mining is performed in three phases, namely preprocessing, pattern discovery, and pattern analysis (Srivastava, Cooley, Deshpande, & Tan, 2000). The preprocessing phase converts the usage, content, and structure information contained in various data...