Trends in Web Usage Mining

Anita Lee-Post
University of Kentucky, USA

Haihao Jin
University of Kentucky, USA

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, Tsioutsiouliklis, 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

sources into data abstractions necessary for pattern discovery. Pattern discovery draws upon methods and algorithms such as statistical analysis, association rules, clustering, classification, sequential pattern and dependency modeling to characterize usage patterns in the form of frequency of page views, Web pages that frequently appear together in users' sessions, usage clusters, page clusters, inter-session patterns, and user profiles. The usage patterns discovered can be analyzed to personalize the Web experience for a user, improve the performance of Web servers and Web-based applications, modify the design of a website, or manage customer relationships.

Personalization or personalizing the Web experience for a user can be achieved by making dynamic recommendations to a Web user based on his/her profile or navigation pattern (Eirinaki & Vazirgiannis, 2003). The recommendations can be a site cluster that contains dynamically selected page links. They can also be an adaptive website that automatically improves its organization and presentation to suit a user's access pattern (Perkowitz & Etzioni, 2000).

Pre-fetching and caching improve the performance of Web servers and Web-based applications by reducing server response time (Jespersen, Pedersen, & Thorhauge, 2003; Lan, Bressan, Ooi, & Tan, 2000). Frequently accessed pages are "pre-fetched" from the Web server and "cached" to anticipate and satisfy future user requests expediently.

Website redesign or modifying the design of a website to improve its usability and quality can be guided by usage mining discoveries (Srikant & Yang, 2001). Detailed feedback on user behavior can be used by Website designer to improve the content and structure of its website. Usage patterns discovered from server log can also be clustered to generate index pages or home pages.

E-commerce intelligence or mining business intelligence from web usage data is important for e-commerce operations, in particular, customer relationship management. Data generated about customers from e-commerce transactions can be analyzed to improve marketing, sales, and customer services. Effective customer relationship management is made possible by tailoring these customer service activities to maximize customer satisfaction.

Other recent applications of usage mining include search relevance ranking, adaptive Web site navigation, and social network mining. A ranking of search topic relevance was established as a result of mining a user's browsing records (Wang, Chen, Tao, Ma, & Liu, 2002) or a user's past search history (Mukhopadhyay, Giri, & Singh, 2003). Zhu, Hong, and Hughes (2004) paved the way for adaptive Web site navigation by analyzing Web log files to discover conceptual link hierarchies among Web pages. The intent was to allow for less specific search criteria involving more Web pages on multiple conceptual levels. Domingos and Richardson (2001) proposed calculating an online customer's network value based on correlations between online customers. The idea was to recognize the exponential growth potential of the expected profit from a customer in a social network as he/she influenced others to shop online.

The integration of Web content, structure, and usage mining has shown promising results. Fang & Sheng (2004) proposed an approach to maximize the efficiency and effectiveness of a portal page to a Website. The hyperlinks in the portal page was built based on relationships among hyperlinks extracted from the Website as well as relationships among access patterns discovered from Web logs. Cooley (2003) was able to identify interesting Web usage patterns from Web content and structure data. In addition, combining information discovered from both content and structure mining was instrumental in enhancing Web personalization (Eirinaki & Vazirgiannis, 2003; Eirinaki, Vazirgiannis, & Varlamis, 2003).

FUTURE TRENDS

The future trends of web usage mining should be directed to address the following needs:

- Web usage mining is currently restricted to modeling the behavior of visitors to a particular site or a network of sites or to data provided by a small sample of users who have volunteered to have their movements tracked across unrelated sites. A broader and more complete source of data for Web usage mining is critically needed.
- Combining the tasks of content and structure mining has been shown to enhance the quality of search results, returning pages that are both relevant and authoritative. Information discovered from usage mining needs to work with content and structure data for effective personalization and website redesign. Therefore, there is a need to integrate the three areas of content, structure, and usage mining so that a more complete data source can be used and analyzed to yield more useful information and applications.
- The area of Web mining is increasingly multidisciplinary. More advanced extraction technology is needed for unstructured content mining. More intelligent analytical tools are needed for both structure and usage mining. There is a need to involve more academic fields, in particular, machine learning and linguistic science to join efforts in advancing the growth and development of the Web mining field.