Adaptive Web Presence and Evolution through Web Log Analysis

Xueping Li
University of Tennessee, Knoxville, USA

INTRODUCTION

The Internet has become a popular medium to disseminate information and a new platform to conduct electronic business (e-business) and electronic commerce (e-commerce). With the rapid growth of the WWW and the intensified competition among the businesses, effective web presence is critical to attract potential customers and retain current customer thus the success of the business. This poses a significant challenge because the web is inherently dynamic and web data is more sophisticated, diverse, and dynamic than traditional well-structured data. Web mining is one method to gain insights into how to evolve the web presence and to ultimately produce a predictive model such that the evolution of a given web site can be categorized under its particular context for strategic planning. In particular, web logs contain potentially useful information and the analysis of web log data have opened new avenues to assist the web administrators and designers to establish adaptive web presence and evolution to fit user requirements.

BACKGROUND

People have realized that web access logs are a valuable resource for discovering various characteristics of customer behaviors. Various data mining or machine learning techniques are applied to model and understand the web user activities (Borges and Levene, 1999; Cooley et al., 1999; Kosala et al., 2000; Srivastava et al., 2000; Nasraoui and Krishnapuram, 2002). The authors in (Kohavi, 2001; Mobasher et al., 2000) discuss the pros and cons of mining the e-commerce log data. Lee and Shiu (Lee and Shiu, 2004) propose an adaptive website system to automatically change the website architecture according to user browsing activities and to improve website usability from the viewpoint of efficiency. Recommendation systems are used by an ever-increasing number of e-commerce sites to help consumers find products to purchase (Schafer et al, 2001). Specifically, recommendation systems analyze the users' and communities' opinions and transaction history in order to help individuals identify products that are most likely to be relevant to their preferences (e.g. Amazon.com, eBay.com). Besides web mining technology, some researches investigate on Markov chain to model the web user access behavior (Xing et al., 2002; Dhyani et al., 2003; Wu et al., 2005). Web log analysis is used to extract terms to build web page index, which is further combined with text-based and anchor-based indices to improve the performance of the web site search (Ding and Zhou, 2007). A genetic algorithm is introduced in a model-driven decision-support system for web site optimization (Asllani and Lari, 2007). A web forensic framework as an alternative structure for clickstream data analysis is introduced for customer segmentation development and loyal customer identification; and some trends in web data analysis are discussed (Sen et al., 2006).

MAIN FOCUS

Broadly speaking, web log analysis falls into the range of web usage mining, one of the three categories of web mining (Kosala and Blockeel, 2000; Srivastava et al., 2002). There are several steps involved in web log analysis: web log acquisition, cleansing and preprocessing, and pattern discovery and analysis.

Web Log Data Acquisition

Web logs contain potentially useful information for the study of the effectiveness of web presence. Most websites enable logs to be created to collect the server and client activities such as access log, agent log, error log, and referrer log. Access logs contain the bulk of data including the date and time, users' IP addresses, requested URL, and so on. Agent logs provide the information of the users' browser type, browser version, and operating system. Error logs provide problematic and erroneous links on the server such as "file not found", "forbidden to access", et al. Referrer logs provide information about web pages that contain the links to documents on the server.

Because of the stateless characteristic of the Hyper Text Transfer Protocol (HTTP), the underlying protocol used by the WWW, each request in the web log seems independent of each other. The identification of user sessions, in which all pages that a user requests during a single visit, becomes very difficulty (Cooley et al., 1999). Pitkow (1995, 1997, 1998) pointed out that local caching and proxy servers are two main obstacles to get reliable web usage data. Most browsers will cache the recently pages to improve the response time. When a user clicks the "back" button in a browser, the cached document is displayed instead of retrieving the page from the web server. This process can not be recorded by the web log. The existence of proxy servers makes it even harder to identify the user session. In the web server log, requests from a proxy server will have the same identifier although the requests may come from several different users. Because of the cache ability of proxy servers, one requested page in web server logs may actually be viewed by several users. Besides the above two obstacles, the dynamic content pages such as Active Server Pages (ASP) and Java Server Pages (JSP) will also create problems for web logging. For example, although the same Uniform Resource Locator (URL) appears in a web server log, the content that is requested by users might be totally different.

To overcome the above obstacles of inaccuracy web log resulting from caching, proxy server and dynamic web pages, specialized logging techniques are needed. One way is to configure the web server to customize the web logging. Another is to integrate the web logging function into the design of the web pages. For example, it is beneficial to an e-commerce web site to log the customer shopping cart information which can be implemented using ASP or JSP. This specialized log can record the details that the users add items to or remove items from their shopping carts thus to gain insights into the user behavior patterns with regard to shopping carts.

Besides web server logging, package sniffers and cookies can be used to further collection web log data. Packet sniffers can collect more detailed information than web server log by looking into the data packets transferred on the wire or air (wireless connections). However, it suffers from several drawbacks. First, packet sniffers can not read the information of encrypted data. Second, it is expensive because each server needs a separate packet sniffer. It would be difficult to manage all the sniffers if the servers are located in different geographic locations. Finally, because the packets need to be processed by the sniffers first, the usage of packet sniffers may reduce the performance of the web servers. For these reasons, packet sniffing is not widely used as web log analysis and other data collecting techniques.

A cookie is a small piece of information generated by the web server and stored at the client side. The client first sends a request to a web server. After the web server processes the request, the web server will send back a response containing the requested page. The cookie information is sent with the response at the same time. The cookie typically contains the session id, expiration date, user name and password and so on. This information will be stored at the client machine. The cookie information will be sent to the web server every time the client sends a request. By assigning each visitor a unique session id, it becomes easy to identify the sessions. However, some users prefer to disable the usage of cookies on their computers which limits the wide application of cookies.

Web Log Data Cleansing and Preprocessing

Web log data cleansing and preprocessing is critical to the success of the web log analysis. Even though most of the web logs are collected electronically, serious data quality issues may arise from a variety of sources such as system configuration, software bugs, implementation, data collection process, and so on. For example, one common mistake is that the web logs collected from different sites use different time zone. One may use Greenwich Mean Time (GMT) while the other uses Eastern Standard Time (EST). It is necessary to cleanse the data before analysis.

There are some significant challenges related to web log data cleansing. One of them is to differentiate the web traffic data generated by web bots from that generated by "real" web visitors. Web bots, including web robots and