Chapter XVIII

Web Usage Mining: Algorithms and Results

Yew-Kwong Woon
Nanyang Technological University, Singapore

Wee-Keong Ng
Nanyang Technological University, Singapore

Ee-Peng Lim
Nanyang Technological University, Singapore

ABSTRACT

The rising popularity of electronic commerce makes data mining an indispensable technology for several applications, especially online business competitiveness. The World Wide Web provides abundant raw data in the form of Web access logs. However, without data mining techniques, it is difficult to make any sense out of such massive data. In this chapter, we focus on the mining of Web access logs, commonly known as Web usage mining. We analyze algorithms for preprocessing and extracting knowledge from such logs. We will also propose our own techniques to mine the logs in a more holistic manner. Experiments conducted on real Web server logs verify the practicality as well as the efficiency of the proposed techniques as compared to an existing technique. Finally, challenges in Web usage mining are discussed.

INTRODUCTION

The rate of growth of the World Wide Web (Web) may be slowing down, but Online Computer Library Center researchers concluded that the Web would continue to grow rapidly in their annual review of the Web (Dean, 2000). In addition, Forrester Research affirmed the continued popularity of electronic commerce through its prediction that
global online trade would expand to $12.8 trillion by 2006 (Sharrard, Kafka & Tavilla, 2001). Hence, to stay competitive and profitable in a fast paced environment like the Web, companies must be able to extract knowledge from their Web access logs, Web transaction logs and Web user profiles to ensure the success of Customer Relationship Management (CRM) (Berson, Smith & Thearling, 2000). However, the immense amount of Web data makes manual inspection virtually impossible and thus, data mining techniques become indispensable in the quest for cutting-edge knowledge. In fact, electronic commerce has even been touted as the killer domain for data mining (Kohavi, 2001).

Our focus here is on Web Usage Mining (WUM), which we define as the extraction of meaningful user patterns from Web server access logs using data mining techniques. For brevity, we shall use the term logfile to refer to a Web server access log. WUM is fast gaining importance because of the wide availability of logfiles as well as its applicability in CRM (Woon, Ng & Lim, 2002b). In addition, it has diversified applications such as Web personalization (Mobasher, Cooley, & Srivastava; Mobasher, Dai, Luo, Sun & Zhu, 2000), Website structuring (Masseglia, Poncelet & Teisseire, 1999; Perkowitz & Etzioni, 1999; Spiliopoulou, 2000), marketing (Buchner & Mulvenna, 1998), user profiling (Heer & Chi., 2002; Mobasher et al., 2000), caching and prefetching (Yang, Zhang & Li, 2001).

Unfortunately, as most logfiles are originally meant for debugging purposes, they are non-ideal candidates for WUM raw data (Kohavi, 2001). However, due to their adoption by a vast number of existing Web servers, we postulate that they will not be replaced as the de facto Web data sources in the near future. Sources of logfiles include Web servers, Web clients, proxy servers and application servers (Kohavi, 2001; Srivastava, Cooley, Deshpande & Tan, 2000). A standard logfile has the following format (Consortium, 1995): 

\[
\text{remotehost logname username date request status bytes}
\]

where:
- \text{remotehost} is the remote hostname or its IP address,
- \text{logname} is the remote logname of the user,
- \text{username} is the username with which the user has authenticated himself,
- \text{date} is the date and time of the request,
- \text{request} is the exact request line as it came from the client,
- \text{status} is the HTTP status code returned to the client, and
- \text{bytes} is the content-length of the document transferred.

The following is a fragment of a common logfile:

\begin{verbatim}
ntu.edu.sg - - [30/May/2003:00:01:15 -0400] "GET /html/faq.html HTTP/1.0" 200 4855
155.69.181.254 - - [1/Jun/2003:00:03:22 -0400] "GET /pub/home.html HTTP/1.0" 200 165
\end{verbatim}

As observed in the example, the fields logname and username are usually not recorded. Therefore, it is difficult to identify the activities of individual users. An extended format is also available to capture demographic data and session identifiers but we shall only focus on the standard common logfile format because of its wider adoption and fewer privacy/security concerns (Hallam-Baker & Behlendorf, 1996).

There are several existing works on logfile mining but they deal separately on specific issues of mining and make certain assumptions without taking a holistic view,