Chapter I

Combining Data Warehousing and Data Mining Techniques for Web Log Analysis

Torben Bach Pedersen, Aalborg University, Denmark

Jesper Thorhauge, Concentrate, Denmark

Søren E. Jespersen, Linkage, Denmark

Abstract

Enormous amounts of information about Web site user behavior are collected in Web server logs. However, this information is only useful if it can be queried and analyzed to provide high-level knowledge about user navigation patterns, a task that requires powerful techniques. This chapter presents a number of approaches that combine data warehousing and data mining techniques in order to analyze Web logs. After introducing the well-known click and session data warehouse (DW) schemas, the chapter presents the subsession schema, which allows fast queries on sequences
of page visits. Then, the chapter presents the so-called “hybrid” technique, which combines DW Web log schemas with a data mining technique called Hypertext Probabilistic Grammars, hereby providing fast and flexible constraint-based Web log analysis. Finally, the chapter presents a “post-check enhanced” improvement of the hybrid technique.

Introduction

With the large number of companies using the Internet to distribute and collect information, knowledge discovery on the Web—or Web mining—has become an important research area. Web mining can be divided into three areas, namely Web content mining, Web structure mining, and Web usage mining (also called Web log mining) (Cooley, Srivastava, & Mobasher, 1997). Web content mining focuses on discovery of information stored on the Internet—that is, the various search engines. Web structure mining can be used when improving the structural design of a Web site. Web usage mining, the main topic of this chapter, focuses on knowledge discovery from the usage of individual Web sites.

Web usage mining is mainly based on the activities recorded in the Web log, the log file written by the Web server recording individual requests made to the server. An important notion in a Web log is the existence of user sessions. A user session is a sequence of requests from a single user within a certain time window. Of particular interest is the discovery of frequently performed sequences of actions by the Web user—that is, frequent sequences of visited Web pages.

The work presented in this chapter has to a large extent been motivated by collaboration with the e-learning company Zenaria (www.zenaria.com). Zenaria is in the business of creating e-learning, namely interactive stories told through a series of video-sequences. The story is formed by a user first viewing a video-sequence and then choosing between some predefined options, based on the video-sequence. Depending on the user’s choice, a new video-sequence is shown and new options are presented. The choices of a user will form a complete story—a walkthrough—reflecting the choices made by the individual users. Traditionally, stories have been distributed on CD-ROM, and a consultant has evaluated the walkthroughs by observing the users. However, Zenaria now wants to distribute its stories using the Internet (i.e., a walkthrough will correspond to a Web session). Thus, the consultants will now have to use Web usage mining technology to evaluate the walkthroughs.

Data warehousing and database management system (DBMS) technologies excel in handling large amounts of data with good performance and ease-of-use. These technologies support a wide range of complex analysis queries, such as aggregation queries with constraints very well due to techniques such as bitmapped indices, ma-
Ontology-Based Construction of Grid Data Mining Workflows
www.igi-global.com/chapter/ontology-based-construction-grid-data/7578?camid=4v1a