Chapter X

Traversals Pattern Mining in Web Usage Data

Yongqiao Xiao, Georgia College & State University, USA
Jenq-Foung (J.F.) Yao, Georgia College & State University, USA

ABSTRACT

Web usage mining is to discover useful patterns in the web usage data, and the patterns provide useful information about the user’s browsing behavior. This chapter examines different types of web usage traversal patterns and the related techniques used to uncover them, including Association Rules, Sequential Patterns, Frequent Episodes, Maximal Frequent Forward Sequences, and Maximal Frequent Sequences. As a necessary step for pattern discovery, the preprocessing of the web logs is described. Some important issues, such as privacy, sessionization, are raised, and the possible solutions are also discussed.

INTRODUCTION

Web usage mining is to discover useful patterns in the web usage data, i.e., web logs. The web logs record the user’s browsing of a web site, and the patterns provide useful information about the user’s browsing behavior. Such patterns can be used for web design, improving web server performance, personalization, etc.
Several different types of traversal patterns have been proposed in the literature, namely, Association Rules, Sequential Patterns, Frequent Episodes, Maximal Frequent Forward Sequences, and Maximal Frequent Sequences. These patterns differ in how the patterns are defined, and they can be used for different purposes. This chapter examines these patterns and the related techniques used to uncover them.

One important issue about mining traversal patterns, or about web usage mining in general, is the preprocessing of the web logs. Since the web logs are usually not in a format for web usage mining, preprocessing is needed. Such preprocessing becomes complicated or problematic by the current use of the Web. The details about the problems and possible solutions are discussed in this chapter.

The rest of the chapter is organized as follows: The second section, Web Usage Data, gives the background to web usage mining. The third section, Preprocessing, describes the web log preprocessing. The different types of traversal patterns are described in the fourth section, Pattern Discovery. The fifth section, Pattern Analysis and Applications, describes the analyses and applications of these patterns. The sixth section, Conclusion, concludes the chapter.

Web Usage Data

To characterize the web usage data, the terms defined by the W3C Web Characterization Activity (WCA) (http://www.w3c.org/WCA) are adopted. A user is defined as an individual who is accessing the Web through a browser. A user session is a delimited set of user clicks across one or more web servers. A click corresponds to a page on the web server, which is uniquely identified by a URI (Universal Resource Identifier). A server session is a collection of user clicks to a single web server during a user session.

The web usage data can be collected from different sources, e.g., the server side, the client side (Catledge & Pitkow, 1995) and the proxy side (Cohen et al., 1998). The server usage data correspond to the logs that are collected at a web server. They provide an aggregate view of the usage of a web site by all users. Such web server log data may not be entirely reliable due to the presence of various levels of caching (e.g., client caching by the browser, proxy caching by the proxy server) within the Web environment. The client usage data can be collected by using a remote agent, e.g., Java Applets, or by asking the user to use a specialized browser. The client side data can potentially capture every click of the user, but it requires the user’s cooperation to collect.
Related Content

Search Query Recommendations in Web Information Retrieval Using Query Logs
[www.irma-international.org/chapter/search-query-recommendations-in-web-information-retrieval-using-query-logs/162895/](www.irma-international.org/chapter/search-query-recommendations-in-web-information-retrieval-using-query-logs/162895/)

A Case Study for External Users
[www.irma-international.org/chapter/case-study-external-users/8171/](www.irma-international.org/chapter/case-study-external-users/8171/)

High Performance Scheduling Mechanism for Mobile Computing Based on Self-Ranking Algorithm (SRA)
[www.irma-international.org/chapter/high-performance-scheduling-mechanism-mobile/23990/](www.irma-international.org/chapter/high-performance-scheduling-mechanism-mobile/23990/)

SWAMI: A Multiagent, Active Representation of a User’s Browsing Interests
[www.irma-international.org/chapter/swami-multiagent-active-representation-user/53060/](www.irma-international.org/chapter/swami-multiagent-active-representation-user/53060/)

Web Data Warehousing Convergence: From Schematic to Systematic
[www.irma-international.org/article/web-data-warehousing-convergence/2619/](www.irma-international.org/article/web-data-warehousing-convergence/2619/)