Chapter IX

Indexing Techniques for Web Access Logs

Yannis Manolopoulos, Aristotle University of Thessaloniki, Greece
Mikolaj Morzy, Poznan University of Technology, Poland
Tadeusz Morzy, Poznan University of Technology, Poland
Alexandros Nanopoulos, Aristotle University of Thessaloniki, Greece
Marek Wojciechowski, Poznan University of Technology, Poland
Maciej Zakrzewicz, Poznan University of Technology, Poland

ABSTRACT

Access histories of users visiting a web server are automatically recorded in web access logs. Conceptually, the web-log data can be regarded as a collection of clients’ access-sequences, where each sequence is a list of pages accessed by a single user in a single session. This chapter presents novel indexing techniques that support efficient processing of so-called pattern queries, which consist of finding all access sequences that contain a given subsequence. Pattern queries are a key element of advanced analyses of web-log data, especially those concerning typical navigation schemes. In this chapter, we discuss the particularities of efficiently processing user access-sequences with pattern queries, compared to the
case of searching unordered sets. Extensive experimental results are given, which examine a variety of factors and illustrate the superiority of the proposed methods over indexing techniques for unordered data adapted to access sequences.

INTRODUCTION

Web servers have recently become the main source of information on the Internet. Web access logs record the access history of users who visit a web server. Each web-log entry represents a single user’s access to a web resource (HTML document, image, CGI program, etc.) and contains the client’s IP address, the timestamp, the URL address of the requested resource, and some additional information. Conceptually, the web-log data can be regarded as a collection of clients’ access-sequences, where each sequence is a list of pages accessed by a single user in a single session. Extraction of user access-sequences is a required pre-processing step in advanced analyses of web logs (called web-log mining), and it involves data cleaning and techniques of forming user sessions (see Cooley, Mobasher, & Srivastava, 1999; Lou, Liu, Lu, & Yang, 2002).

One of the most popular data mining problems in the context of web-log analysis is discovery of access patterns (Chen, Park, & Yu, 1998; Pei, Han, Mortazavi-Asl, & Zhu, 2000). Each access pattern is a sequence of web pages which occurs frequently in user access-sequences. Sequential access-patterns provide information about typical browsing strategies of users visiting a given website, e.g., “10% of users visited the page about a palmtop X, and later a page about a docking cradle for the palmtop X.” After some frequently occurring sequences have been discovered, the analyst should be able to search for user access-sequences that support (i.e., contain) the patterns. The latter operation finds several applications, e.g., searching for typical/atypical user access-sequences.

Moreover, web-log mining algorithms, such as WUM (Spiliopoulou & Faulstich, 1998), use templates to constrain the search space and to perform more focused mining, according to the user’s requirements. For instance, the user may specify the mining of sequences containing the subsequence <A * B * C>. Thus, a selection of the user access-sequences can be performed to collect those satisfying the given template. In the following, we refer to these types of queries over the database of user access-sequences as pattern queries.

Since web logs tend to be large, a natural solution to support efficient processing of pattern queries would be indexing web access-sequences.
Related Content

Analysis and Development of Load Balancing Algorithms in Cloud Computing
www.igi-global.com/article/analysis-and-development-of-load-balancing-algorithms-in-cloud-computing/204358?camid=4v1a

Engineering Wireless Mobile Applications
www.igi-global.com/article/engineering-wireless-mobile-applications/2603?camid=4v1a

A Review of Methodologies for Analyzing Websites
www.igi-global.com/chapter/review-methodologies-analyzing-websites/22000?camid=4v1a
Modeling of Complex Human-Process Interaction as Framework for Assistance and Supervisory Control of Technical Processes
www.igi-global.com/article/modeling-complex-human-process-interaction/68965?camid=4v1a