Chapter 16

SEQUENCE MINING IN DYNAMIC AND INTERACTIVE ENVIRONMENTS

Srinivasan Parthasarathy
Computer Science Department, University of Rochester, Rochester, NY 14627
srini@cs.rochester.edu

Mohammed J. Zaki
Computer Science Department, Rensselaer Polytechnic Institute, Troy, NY 12180
zaki@cs.rpi.edu

Mitsunori Ogihara
Computer Science Department, University of Rochester, Rochester, NY 14627
ogihara@cs.rochester.edu

Sandhya Dwarkadas
Computer Science Department, University of Rochester, Rochester, NY 14627
sandya@cs.rochester.edu

Keywords: Frequent Sequences, Sequential Patterns, Incremental and Interactive Algorithms, Data Mining, Knowledge Discovery

Abstract: The discovery of frequent sequences in temporal databases is an important data mining problem. Most current work assumes that the database is static, and a database update requires rediscovering all the patterns by scanning the entire old and new database. In this paper, we propose novel techniques for maintaining sequences in the presence of a) database updates, and b) user interaction (e.g. modifying mining parameters). This is a very challenging task, since such updates can invalidate existing sequences or introduce new ones. In both the above scenarios, we avoid re-executing the algorithm on the entire dataset, thereby reducing execution time. Experimental results confirm that our approach results
in execution time improvements of up to several orders of magnitude in practice.

1. INTRODUCTION

Dynamism and interactivity are essential features of all human endeavors, be they in a scientific or business enterprise. The collection of information and data in various fields serves as an exemplar, where every moment we are faced with new content (dynamism) and are required to manipulate it (interactivity). For example, consider a large retail store like Walmart, which has a data-warehouse more than a terabyte in size. In addition, Walmart collects approximately 20 million customer transactions every day. It is simply infeasible to mine the entire database (the original terabyte data-warehouse, and the new transactions) each time an update occurs. As another example consider Web Mining. Let's assume we have mined interesting browsing patterns at a popular portal site like Yahoo! that receives millions of hits every day. Once again it is not practical to re-mine the site logs each time an update occurs.

Given the inherently dynamic nature of data collection, it is somewhat surprising that incremental techniques have received little to no attention within knowledge discovery and data mining. It is worth noting that without incrementality, interactivity also remains a distant goal. True interactivity is not possible if once is forced to re-mine the entire database from scratch each time. This paper seeks to address the problem of mining frequent sequences in dynamic and interactive environments. For example, incremental updates of the most frequent sequence of items purchased by customers, or real-time mining of browsing patterns (i.e., sequences of web-pages) on the Internet.

Sequence mining is an important data mining task, where one attempts to discover frequent sequences over time, of attribute sets in large databases. This problem was originally motivated by applications in the retail industry (e.g. the Walmart example from above), including attached mailing, add-on sales and customer satisfaction. Besides the retail and Internet examples, it applies to many other scientific and business domains. For instance, in the health care industry it can be used for predicting the onset of disease from a sequence of symptoms, and in the financial industry it can be used for predicting investment risk based on a sequence of stock market events.

Discovering all frequent sequences in a very large database can be very compute and I/O intensive because the search space size is essentially exponential in the length of the longest transaction sequence in it. This high computational cost may be acceptable when the database is static since the discovery is done only once, and several approaches to this problem have been presented in the literature. However, many domains such as electronic commerce, stock analysis, collaborative surgery, etc., impose soft real-time constraints on the mining pro-