Chapter VII

DWFIST: The Data Warehouse of Frequent Itemsets Tactics Approach

Rodrigo Salvador Monteiro,  
Federal University of Rio de Janeiro, Brazil &  
University of Stuttgart, Germany

Geraldo Zimbrão, Federal University of Rio de Janeiro, Brazil

Holger Schwarz, University of Stuttgart, Germany

Bernhard Mitschang, University of Stuttgart, Germany

Jano Moreira de Souza,  
Federal University of Rio de Janeiro, Brazil &  
University of Stuttgart, Germany

Abstract

This chapter presents the core of the DWFIST approach, which is concerned with supporting the analysis and exploration of frequent itemsets and derived patterns, e.g., association rules in transactional datasets. The goal of this new approach is to provide: (1) flexible pattern-retrieval capabilities without requiring the original data during the analysis phase; and (2) a
standard modeling for data warehouses of frequent itemsets, allowing an easier development and reuse of tools for analysis and exploration of itemset-based patterns. Instead of storing the original datasets, our approach organizes frequent itemsets holding on different partitions of the original transactions in a data warehouse that retains sufficient information for future analysis. A running example for mining calendar-based patterns on data streams is presented. Staging area tasks are discussed and standard conceptual and logical schemas are presented. Properties of this standard modeling allow retrieval of frequent itemsets holding on any set of partitions, along with upper and lower bounds on their frequency counts. Furthermore, precision guarantees for some interestingness measures of association rules are provided as well.

Introduction

Some data mining tasks can produce such great amounts of data that there has arisen a new knowledge management problem (Klemettinen et al., 1994). Frequent itemset mining is long known for fitting in this category. The analysis of the results of a frequent itemset mining task is far from being trivial. The same is true for many patterns built upon frequent itemsets, such as association rules. The analyst may be easily confronted with a huge number of patterns during such an analysis. Specialized analytical and exploratory tools must be devised in order to aid analysts. The lack of a standardized way for organizing, storing and accessing frequent itemsets makes the effort of developing such tools very difficult as it avoids the reuse of general solutions for different environments.

Recent applications, such as network traffic analysis, Web clickstream mining, power consumption measurement, sensor network data analysis and dynamic tracing of stock fluctuation are some examples where a new kind of data arises, the so-called data stream. A data stream is continuous and potentially infinite. It is challenging to mine frequent patterns in data streams because this task is essentially a set of join operations, whereas join is a typical blocking operator (i.e., computation for any itemset cannot complete before seeing the past and the future data sets) (Giannella et al., 2003). Providing flexibility to mine frequent itemsets in some subset of the data stream is even more challenging, especially when the subset is not known a priori.

The research field of data warehousing has been extremely successful in providing efficient and effective ways to store and organize huge amounts of data. It has succeeded also in providing a standard modeling upon which reusable analytical tools could be designed and implemented. This chapter presents a
Demonstrating and Communicating the Value of Nuclear Decommissioning to Society
www.igi-global.com/article/demonstrating-communicating-value-nuclear-decommissioning/46125?camid=4v1a