Incremental Association Rule Mining Using Materialized Data Mining Views

Mikołaj Morzy, Tadeusz Morzy, and Zbyszko Króliskowski

Institute of Computing Science, Poznań University of Technology, Piotrowo 3A, 60-965 Poznań, Poland

Abstract. Data mining is an interactive and iterative process. Users issue series of similar queries until they receive satisfying results, yet currently available data mining systems do not support iterative processing of data mining queries and do not allow to re-use the results of previous queries. Consequently, mining algorithms suffer from long processing times, which are unacceptable from the point of view of interactive data mining. On the other hand, the results of consecutive data mining queries are usually very similar. This observation leads to the idea of reusing materialized results of previous data mining queries. We present the notion of a materialized data mining view and we propose two novel algorithms which aim at efficient discovery of association rules in the presence of materialized results of previous data mining queries.

1 Overview of Data Mining Processing

Data mining, also referred to as knowledge discovery in databases, is a non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [4]. Data mining systems are evolving from systems dedicated to and specialized in particular tasks or domains to general-purpose systems, which are tightly coupled with the existing relational database technology. Most data mining queries are expensive in terms of processing cost and differ significantly from typical database queries. Hence, novel methods of query processing and optimization need to be developed in order to achieve satisfying data mining query performance.

From a user’s point of view the execution of a data mining algorithm and the discovery of a set of patterns is an answer to a sophisticated database query. A user limits the mined dataset and determines the values of parameters that control a given algorithm. In return, the system discovers relevant patterns and presents them to the user. When the process starts, a user does not know the exact goal of the exploration. Rather, they achieve satisfying results in several consecutive steps. In each step the user verifies discovered patterns and, suitably to the needs, expectations, and experience modifies either the mined dataset, or algorithm parameters, or both. Mining practice shows that the vast majority of data mining queries are only minor modifications of former queries. Given these
circumstances it is necessary to be able to exploit the results of previous queries in order to be able to answer a given query efficiently. A data mining system should be capable of answering a query in an incremental manner where the results of previous queries are maintained and tested against the current data set and parameter set and the base algorithm should be run only on the difference set. This principle applies also to the situation when the mining algorithm is run after a data warehouse refresh to discover new patterns. Usually, the volume of new or changed data after the data warehouse refresh is significantly smaller when compared to the size of the original data warehouse.

The basic problem in data mining is the processing time of data mining queries. In addition, the size of the result can easily surpass the size of the queried database. Such properties of mining process make it unsuitable for interactive and iterative pattern discovery. One possible solution is to use materialized views. Data mining query results can be materialized automatically or at a user’s request. Materialized views have been thoroughly examined and successfully applied in traditional database systems. We propose to follow this path and introduce materialized views to data mining systems.

In this paper we present the concept of materialized data mining views. Section 2 contains definitions of basic terms used throughout the paper. The notion of a data mining query is presented in Sec. 3. Data mining views and materialized data mining views are presented in Sec. 4. We demonstrate the use of materialized views in association rule discovery in Sec. 5. Section 6 presents novel algorithms of complementary association rule mining using materialized data mining views. The paper concludes with the presentation of experimental results in Sec. 7.

2 Basic Definitions

Let \( L = \{l_1, \ldots, l_n\} \) be a set of literals called items. Let \( D \) be a set of variable length transactions and \( \forall T \in D : T \subseteq L \). A transaction \( T \) supports an item \( x \) if \( x \in T \). The transaction \( T \) supports an itemset \( X \) if it supports every element \( x \in X \). The support of an itemset is the number of transactions supporting the itemset. The problem of discovering frequent itemsets consists in finding all itemsets with the support higher than user-defined minimum support threshold denoted as \( \text{minsup} \). An itemset with the support higher than \( \text{minsup} \) is called a frequent itemset.

An association rule is an implication of the form \( X \rightarrow Y \) where \( X \subset L, Y \subset L \) and \( X \cap Y = \emptyset \). \( X \) is called the head of a rule whilst \( Y \) is called the body of a rule. Two measures represent statistical significance and strength of a rule. The support of a rule is the number of transactions that support \( X \cup Y \). The confidence of a rule is the ratio of the number of transactions that support the rule to the number of transactions that support the head of the rule.

The problem of discovering association rules consists in finding all rules with support and confidence higher than the user-specified thresholds of minimum support and confidence, called \( \text{minsup} \) and \( \text{minconf} \) respectively. The problem