Closed-Itemset Incremental-Mining Problem

Luminita Dumitriu
"Dunarea de Jos" University, Romania

INTRODUCTION

Association rules, introduced by Agrawal, Imielinski and Swami (1993), provide useful means to discover associations in data. The problem of mining association rules in a database is defined as finding all the association rules that hold with more than a user-given minimum support threshold and a user-given minimum confidence threshold. According to Agrawal, Imielinski and Swami, this problem is solved in two steps:

1. Find all frequent itemsets in the database.
2. For each frequent itemset *I,* generate all the association rules $I' \Rightarrow I \setminus I'$, where $I' \subset I$.

The second problem can be solved in a straightforward manner after the first step is completed. Hence, the problem of mining association rules is reduced to the problem of finding all frequent itemsets. This is not a trivial problem, because the number of possible frequent itemsets is equal to the size of the power set of I, $2^{|I|}$.

Many algorithms are proposed in the literature, most of them based on the Apriori mining method (Agrawal & Srikant, 1994), which relies on a basic property of frequent itemsets: All subsets of a frequent itemset are frequent. This property also says that all supersets of an infrequent itemset are infrequent. This approach works well on weakly correlated data, such as market-basket data. For overcorrelated data, such as census data, there are other approaches, including Close (Pasquier, Bastide, Taouil, & Lakhal, 1999), CHARM (Zaki & Hsiao, 1999) and Closet (Pei, Han, & Mao, 2000), which are more appropriate.

An interesting study of specific approaches is performed in Zheng, Kohavi and Mason (2001), qualifying CHARM as the most adjusted algorithm to real-world data. Some later improvements of Closet are mentioned in Wang, Han, and Pei (2003) concerning speed and memory usage, while a different support-counting method is proposed in Zaki and Gouda (2001). These approaches search for closed itemsets structured in lattices that are closely related with the concept lattice in formal concept analysis (Ganter & Wille, 1999). The main advantage of a closed itemset approach is the smaller size of the resulting concept lattice versus the number of frequent itemsets, that is, search space reduction.

In this article, I describe the closed-itemset approaches, considering the fact that an association rule mining process leads to a large amount of results (most of the time, that is) difficult to understand by the user. I take into account the interactivity of the data-mining process proposed by Ankerst (2001).

BACKGROUND

In this section I first describe the use of closed-itemset lattices as the theoretical framework for the closed-itemset approach. The application of Formal Concept Analysis to the association rule problem was first mentioned in Zaki and Ogihara (1998). For more details on lattice theory, see Ganter and Wille (1999).

The closed-itemset approach is described below. I define a context (***T***, ***I***, ***D***), the Galois connection of a context ((***T***, ***I***, ***D***), *s, t*), a concept of the context *(X, Y),* and the set of concepts in the context, denoted $\boldsymbol{\beta}$(***T***, ***I***, ***D***).

The main result in the Formal Concept Analysis theory is as follows:

- **Fundamental theorem of Formal Concept Analysis (FCA):** Let (***T***, ***I***, ***D***) be a context. Then $\boldsymbol{\beta}$(***T***, ***I***, ***D***) is a complete lattice with join and meet operators given by closed set intersection and reunion operators.

How does FCA apply to the association rule problem? First, ***T*** is the set of transaction ids, ***I*** is the set of items in the database, and ***D*** is the database itself. Second, the mapping *s* associates to a transaction set *X* the maximal itemset *Y* present in all transactions in *X.* The mapping *t* associates to any itemset *Y* the maximal transaction set *X,* where each transaction comprises all the items in the itemset *Y.* The resulting frequent concepts are considered in mining application only for their itemset side, the transaction set side being ignored.

What is the advantage of FCA application? Among the results in Apriori, there are itemsets — I will call them *Y* and *Y',* where *Y'* is included in *Y,* and they have the same support. These two itemsets are two distinct results of Apriori, even if they characterize differently the same transaction set. In fact, the longest itemset is the most precise characterization of that transaction set, all the others being partial and redundant definitions. Due to the observation that $s \circ t$ and $t \circ s$ are closure operators, the concepts in the lattice of concepts eliminate the presence of any unclosed itemsets. Under these circumstances, the FCA-based approaches have only subunitary confidence association rules as results, due to the fact that all unitary confidence association rules are considered redundant behaviors. All unitary confidence association rules can be expressed through a base, the pseudo-intent set.

If I consider the data-mining process in the vision of Ankerst (2001) the resulting data model in Apriori is a long list of frequent itemsets, while FCA is a conceptual structure, namely a lattice of concepts, free of any redundancy.

MAIN THRUST

Many interesting algorithms are issued from the FCA approach to the association rule problem. Although many differences exist between the support counting, memory usage and performance of all these algorithms, the general lines are almost the same.

In the following sections, I take into account some of the main characteristics of these algorithms.

Resulting Data Model

Most of the FCA approaches do not generate a lattice of concepts but a spanning tree on that lattice. The argument is that some of the pairs of adjacent concepts in lattice can be inferred later. The main algorithms that follow this principle are CHARM and Closet.

One approach builds the entire lattice (Dumitriu, 2002), called ERA. The argument for building the entire lattice resides in the fact that missing pairs of adjacent concepts in the spanning tree-based approaches can have an identical support count, thus transforming one of the concepts in the pair in a nonconcept. In fact, CHARM has a later stage of results checking from this point of view. Another strong point of the ERA algorithm is that it offers the pseudo-intents as results as well, thus completely characterizing the data.

Itemset-Building Strategy

While Apriori is a breadth-first result-building algorithm, most of the FCA-based algorithms are depth-first. Only ERA has a different strategy: Each item in the database is used to enlarge an already existing data model built upon the previously selected items, thus generating at all times a new and extended data model. This strategy generates results layer by layer, just like an onion.

The main difference between the depth-first strategy and the layer-based strategy is that interactivity is offered to the user. Just like peeling an onion, one can take a previously found data model, reduce it or enlarge it with some items, and reach the data view that is the most revealing to the individual.

In breadth-first as well as depth-first strategies, it is impossible to provide interactivity to the user due to the fact that all items of interest for the mining process have to be available from the start.

FUTURE TRENDS

I am considering that the most challenging trends would manifest in quantitative attribute mapping as well as in online mining. The first case has a well-known problem in what concerns the quality of numerical-to-Boolean attribute mapping. Solutions to this problem are already considered in Aumann and Lindell (1999), Hong, Kuo, Chi, and Wang (2000), Imberman and Domanski (2001), and Webb (2001), but the problem is far from resolution. An interactive association rule approach may help in a generate-and-test paradigm to find the most suitable mappings in a particular database context.

Online mining is very alike cognitive processes: A data model is flooded with facts; either they are consistent with the data model, contradict it, or have a neutral character. When contradicting facts become important (in number, frequency, or any other way), the model has to change. The real problem of the data-mining process is that the model is too large in number of concepts or the concepts are too rigidly related to support change.

CONCLUSION

I have introduced the idea of data model, expressed as a frequent closed-itemset lattice, with a base for global implications, if needed. In the closed-itemset incremental-mining solution, two new and important operations are applicable to data models: extension and reduction with several items. The main advantages of this approach are:

- The construction of small models of data, which makes them more understandable for the user; also, the response time is small