Association Rule Mining and Application to MPIS

Raymond Chi-Wing Wong
The Chinese University of Hong Kong, Hong Kong

Ada Wai-Chee Fu
The Chinese University of Hong Kong, Hong Kong

INTRODUCTION

Association rule mining (Agrawal, Imilienski, & Swami, 1993) has been proposed for understanding the relationships among items in transactions or market baskets. For instance, if a customer buys butter, what is the chance that he/she buys bread at the same time? Such information may be useful for decision makers to determine strategies in a store.

BACKGROUND

Given a set $I = \{I_1, I_2, ..., I_n\}$ of *items* (e.g., carrot, orange and knife) in a supermarket. The database contains a number of transactions. Each transaction t is a binary vector with $t[k]=1$ if t bought item I_k and $t[k]=0$ otherwise (e.g., {1, 0, 0, 1, 0}). An association rule is of the form $X \rightarrow I_j$, where X is a set of some items in I, and I_j is a single item not in X (e.g., {Orange, Knife} ➔Plate).

A transaction t satisfies X if for all items I_k in X, $t[k] = 1$. The *support* for a rule $X \rightarrow I_j$ is the fraction of transactions that satisfy the union of X and I_j. A rule $X \rightarrow I_j$ has *confidence* c% if and only if c% of transactions that satisfy X also satisfy I_j.

The mining process of association rule can be divided into two steps:

1. **Frequent Itemset Generation:** Generate all sets of items that have support greater than or equal to a certain threshold, called *minsupport*
2. **Association Rule Generation:** From the frequent itemsets, generate all association rules that have confidence greater than or equal to a certain threshold called *minconfidence*

Step 1 is much more difficult compared with Step 2. Thus, researchers have focused on the studies of frequent itemset generation.

The *Apriori Algorithm* is a well-known approach, which was proposed by Agrawal & Srikant (1994), to find frequent itemsets. It is an iterative approach and there are two steps in each iteration. The first step generates a set of candidate itemsets. Then, the second step prunes all disqualified candidates (i.e., all infrequent itemsets). The iterations begin with size 2 itemsets and the size is incremented at each iteration. The algorithm is based on the *closure property* of frequent itemsets: if a set of items is frequent, then all its proper subsets are also frequent. The weaknesses of this algorithm are the generation of a large number of candidate itemsets and the requirement to scan the database once in each iteration.

A data structure called *FP-tree* and an efficient algorithm called FP-growth are proposed by Han, Pei, & Yin (2000) to overcome the above weaknesses. The idea of FP-tree is fetching all transactions from the database and inserting them into a compressed tree structure. Then, algorithm FP-growth reads from the FP-tree structure to mine frequent itemsets.

MAIN THRUST

Variations in Association Rules

Many variations on the above problem formulation have been suggested. The association rules can be classified based on the following (Han & Kamber, 2000):

1. **Association Rules Based on the Type of Values of Attribute:** Based on the type of values of attributes, there are two kinds – Boolean association rule, which is presented above, and quantitative association rule. *Quantitative association rule* describes the relationships among some quantitative attributes (e.g., income and age). An example is income(40K..50K) ➔ age(40..45). One proposed method is grid-based — dividing each attribute into a fixed number of partitions [Association Rule Clustering System (ARCS) in Lent, Swami & Widom (1997)]. Srikant & Agrawal (1996) proposed to partition quantitative attributes *dynamically* and to

merge the partitions based on a measure of *partial completeness*. Another non-grid based approach is found in Zhang, Padmanabhan, & Tuzhilin (2004).

2. **Association Rules based on the Dimensionality of Data:** Association rules can be divided into *single-dimensional association rules* and *multi-dimensional association rule*s. One example of single-dimensional rule is buys({Orange, Knife}) ➔ buys(Plate), which contains only the dimension *buys*. Multi-dimensional association rule is the one containing attributes for more than one dimension. For example, income(40K..50K) ➔ buys(Plate). One mining approach is to borrow the concept of *data cube* in the field of data warehousing. *Figure 1* shows a lattice for the data cube for the dimensions age, income and buys. Researchers (Kamber, Han, & Chiang, 1997) applied the data cube model and used the *aggregate* techniques for mining.
3. **Association Rules based on the Level of Abstractions of Attribute:** The rules discussed in previous sections can be viewed as single-level association rule. A rule that references different levels of abstraction of attributes is called a *multilevel association rule*. Suppose there are two rules – income(10K..20K) ➔ buys(fruit) and income(10K..20K) ➔ buys(orange). There are two different levels of abstractions in these two rules because "fruit" is a higher-level abstraction of "orange." Han & Fu (1995) apply a top-down strategy to the concept hierarchy in the mining of frequent itemsets.

Other Extensions to Association Rule Mining

There are other extensions to association rule mining. Some of them (Bayardo, 1998) find *maxpattern* (i.e., maximal frequent patterns) while others (Zaki & Hsiao, 2002) find *frequent closed itemsets*. Maxpattern is a frequent itemset that does not have a frequent item superset. A frequent itemset is a frequent closed itemsets if there exists no itemset X' such that (1) $X \subset X'$ and (2) $\forall$ transactions t, X is in t implies X' is in t. These considerations can reduce the resulting number of frequent itemsets significantly.

Figure 1. A lattice showing the data cube for the dimensions age, income, and buys

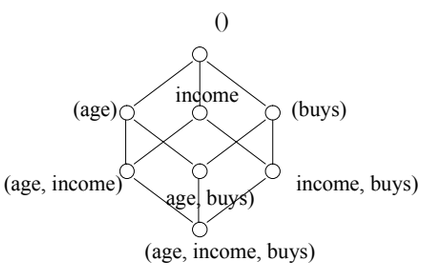
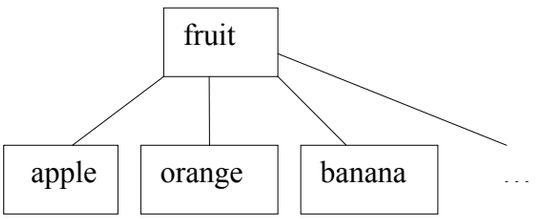


Figure 2. A concept hierarchy of the fruit



Another variation of the frequent itemset problem is mining *top-K frequent* itemsets (Cheung & Fu, 2004). The problem is to find K frequent itemsets with the greatest supports. It is often more reasonable to assume the parameter K, instead of the data-distribution dependent parameter of minsupport because the user typically would not have the knowledge of the data distribution before data mining.

The other variations of the problem are the *incremental update of mining association rules* (Hidber, 1999), *constraint-based rule mining* (Grahne & Lakshmanan, 2000), *distributed and parallel association rule mining* (Gilburd, Schuster, & Wolff, 2004), *association rule mining with multiple minimum supports/without minimum support* (Chiu, Wu, & Chen, 2004), *association rule mining with weighted item and weight support* (Tao, Murtagh, & Farid, 2003), and *fuzzy association rule mining* (Kuok, Fu, & Wong, 1998).

Association rule mining has been integrated with other data mining problems. There have been the integration of classification and association rule mining (Wang, Zhou, & He, 2000) and the integration of association rule mining with relational database systems (Sarawagi, Thomas, & Agrawal, 1998).

Application of the Concept of Association Rules to MPIS

Other than market basket analysis (Blischok, 1995), association rules can also help in applications such as intrusion detection (Lee, Stolfo, & Mok, 1999), heterogeneous genome data (Satou et al., 1997), mining remotely sensed images/data (Dong, Perrizo, Ding, & Zhou, 2000) and product assortment decisions (Wong, Fu, & Wang, 2003; Wong & Fu, 2004). Here we focus on the application on product assortment decisions, as it is one of very few examples where the association rules are not the end mining results.