Algorithms for Association Rule Mining

Vasudha Bhatnagar
University of Delhi, India

Anamika Gupta
University of Delhi, India

Naveen Kumar
University of Delhi, India

INTRODUCTION

Association Rule Mining (ARM) is one of the important data mining tasks that has been extensively researched by data-mining community and has found wide applications in industry. An Association Rule is a pattern that implies co-occurrence of events or items in a database. Knowledge of such relationships in a database can be employed in strategic decision making in both commercial and scientific domains.

A typical application of ARM is market basket analysis where associations between the different items are discovered to analyze the customer's buying habits. The discovery of such associations can help to develop better marketing strategies. ARM has been extensively used in other applications like spatial-temporal, health care, bioinformatics, web data etc (Hipp J., Güntzer U., Nakhaeizadeh G. 2000).

An association rule is an implication of the form $X \rightarrow Y$ where X and Y are independent sets of attributes/items. An association rule indicates that if a set of items X occurs in a transaction record then the set of items Y also occurs in the same record. X is called the antecedent of the rule and Y is called the consequent of the rule. Processing massive datasets for discovering co-occurring items and generating interesting rules in reasonable time is the objective of all ARM algorithms. The task of discovering co-occurring sets of items cannot be easily accomplished using SQL, as a little reflection will reveal. Use of 'Count' aggregate query requires the condition to be specified in the where clause, which finds the frequency of only one set of items at a time. In order to find out all sets of co-occurring items in a database with n items, the number of queries that need to be written is exponential in n. This is the prime motivation for designing algorithms for efficient discovery of co-occurring sets of items, which are required to find the association rules.

In this article we focus on the algorithms for association rule mining (ARM) and the scalability issues in ARM. We assume familiarity of the reader with the motivation and applications of association rule mining

BACKGROUND

Let $I = \{i_1, i_2, \ldots, i_n\}$ denote a set of items and D denote a database of N transactions. A typical transaction $T \in D$ may contain a subset X of the entire set of items I and is associated with a unique identifier *TID*. An *item-set* is a set of one or more items i.e. X is an item-set if $X \subseteq I$. A *k-item-set* is an item-set of cardinality k. A transaction is said to contain an item-set X if $X \subseteq T$. *Support* of an item set X, also called *Coverage* is the fraction of transactions that contain X. It denotes the probability that a transaction contains X.

$$Support(X) = P(X) = \frac{No.\ of\ transactions\ containing\ X}{N}$$

An item-set having support greater than the user specified support threshold (ms) is known as *frequent item-set.*

An *association rule* is an implication of the form $X \rightarrow Y$ [*Support, Confidence*] where $X \subset I$, $Y \subset I$ and $X \cap Y = \emptyset$, where *Support* and *Confidence* are rule evaluation metrics. *Support* of a rule $X \rightarrow Y$ in D is 'S'' if S% of transactions in D contain $X \cup Y$. It is computed as:

$$Support(X \rightarrow Y) = P(X \cup Y) = \frac{No.of\ transaction\ containing\ X \cup Y}{N}$$

Support indicates the prevalence of a rule. In a typical market basket analysis application, rules with very low support values represent rare events and are likely to be uninteresting or unprofitable. *Confidence* of a rule measures its strength and provides an indication of the reliability of prediction made by the rule. A rule $X \rightarrow Y$ has a confidence 'C'' in D if C % of transactions in D that contain X, also contain Y. Confidence is computed, as the conditional probability of Y occuring in a transaction, given X is present in the same transaction, i.e.

$$Confidence(X \rightarrow Y) = P(Y/X) = \frac{P(X \cup Y)}{P(X)} = \frac{Support(X \cup Y)}{Support(X)}$$

A rule generated from frequent item-sets is *strong* if its confidence is greater than the user specified confidence threshold (*mc*). Fig. 1 shows an example database of five transactions and shows the computation of support and confidence of a rule.

The objective of Association Rule Mining algorithms is to discover the set of strong rules from a given database as per the user specified *ms* and *mc* thresholds. Algorithms for ARM essentially perform two distinct tasks: (1) Discover frequent item-sets. (2) Generate strong rules from frequent item-sets.

The first task requires counting of item-sets in the database and filtering against the user specified threshold (*ms*). The second task of generating rules from frequent item-sets is a straightforward process of generating subsets and checking for the strength. We describe below the general approaches for finding frequent item-sets in association rule mining algorithms. The second task is trivial as explained in the last section of the article.

APPROACHES FOR GENERATING FREQUENT ITEM-SETS

If we apply a brute force approach to discover frequent item-sets, the algorithm needs to maintain counters for all $2^n - 1$ item-sets. For large values of n that are common in the datasets being targeted for mining, maintaining such large number of counters is a daunting task. Even if we assume availability of such large memory, indexing of these counters also presents a challenge. Data mining researchers have developed numerous algorithms for efficient discovery of frequent item-sets.

The earlier algorithms for ARM discovered all frequent item-sets. Later it was shown by three independent groups of researchers (Pasquier N., Bastide Y., Taouil R. & Lakhal L. 1999), (Zaki M.J. 2000), (Stumme G., 1999), that it is sufficient to discover frequent closed item-sets (FCI) instead of all frequent item-sets (FI). FCI are the item-sets whose support is not equal to the support of any of its proper superset. FCI is a reduced, complete and loss less representation of frequent item-sets. Since FCI are much less in number than FI, computational expense for ARM is drastically reduced.

Figure 2 summarizes different approaches used for ARM. We briefly describe these approaches.

Discovery of Frequent Item-Sets

Level-Wise Approach

Level wise algorithms start with finding the item-sets of cardinality one and gradually work up to the frequent item-sets of higher cardinality. These algorithms use anti-monotonic property of frequent item-sets accord-

Figure 1. Computation of support and confidence of a rule in an example database

TID	Items
1	BCD
2	BCDE
3	AC
4	BDE
5	AB

Let ms=40%, mc=70%
Consider the association rule $B \rightarrow D$,
support ($B \rightarrow D$) = 3/5 = 60%
confidence($B \rightarrow D$) = support(B? D)/support(B)
= 3/4 = 75%
The rule $B \rightarrow D$ is a strong rule.