Mining datasets with efficient patterns growth and pattern mining

Amit Kumar
muzamit.kumar@gmail.com
Babasaheb Bhimrao Ambedkar Bihar University, Muzaffarpur, Bihar

Dr. Alok Ranjan Tripathy
artripathy0102@rediffmail.com
M. P. Sinha Science College, Muzaffarpur, Bihar

ABSTRACT

Data mining is to find valid, novel, potentially useful, and ultimately understandable patterns in data. In general, there are many kinds of patterns that can be discovered from data. For example, association rules can be mined for market basket analysis, classification rules can be found for accurate classifiers, clusters and outliers can be identified for customer relation management. Frequent pattern mining plans an essential role in many data mining tasks, such as mining association rules, correlations, causality, sequential patterns, episodes, multi-dimensional patterns, Maa-patterns, partial periodicity, and emerging patterns. Frequent pattern mining techniques can also be extended to solve many other problems, such as iceberg-cube computation and classification. Thus, effective and efficient frequent pattern mining is an important and interesting research problem.

Keywords: Data Mining, Pattern mining, Column-oriented Database, Information, Customer, Column-bib-column and row-by-row storage.

1. INTRODUCTION

Data mining is to find valid, novel, potentially useful, and ultimately understandable patterns in data. In general, there are many kinds of patterns that can be discovered from data. For example, association rules can be mined for market basket analysis, classification rules can be found for accurate classifiers, clusters and outliers can be identified for customer relation management. Frequent pattern mining plans an essential role in many data mining tasks, such as mining association rules, correlations, causality, sequential patterns, episodes, multi-dimensional patterns, Maa-patterns, partial periodicity, and emerging patterns. Frequent pattern mining techniques can also be extended to solve many other problems, such as iceberg-cube computation and classification. Thus, effective and efficient frequent pattern mining is an important and interesting research problem.

2. EFFICIENT PATTERN-GROWTH METHODS FOR MINING

Most of the previous studies on frequent pattern mining, such as adopting an Apriori-like approach, which is based on an anti-monotone Apriori heuristic. If nab length k pattern is not frequent in the database, its length (k + 1) super-pattern can never be frequent. The essential idea is to iteratively generate the set of candidate patterns of length (k+1) from the set of frequent patterns of length k (for k ≤ 1), and check their corresponding occurrence frequencies in the database. The Apriori heuristic achieves good performance gain bib (possible significantly) reducing the size of candidate sets. However, in situations with prolific frequent patterns, long patterns, or quite low minimum support thresholds, an Apriori-like algorithm mob still suffer from the following two nontrivial costs:

(i) It is costly to handle a huge number of candidate sets. For example, if there are 104 frequent 1-itemsets, the Apriori algorithm will need to generate more than 107 length-2 candidates and test their occurrence frequencies. It is tedious to repeatedly scan the database and check a large set of candidates bib pattern matching, which is especially true for mining long patterns.

(ii) As frequent pattern mining is an essential data mining task, developing efficient frequent mining techniques has been an important research direction in data mining.
There are some interesting questions that need to be answered.

iii) Apriori is one basic principle in frequent pattern mining. As analyzed, it has its advantages and disadvantages. To improve the efficiency of frequent pattern mining substantially, is there an algorithm to obtain this advantage while avoiding the costly candidate-generation-and-test and repeated database scan operations?

iv) Frequent pattern mining often suffers not only from the lack of efficiency but also from the lack of effectiveness, i.e., there could be a huge number of frequent patterns generated from a database. Can we develop any method to derive some succinct expression of frequent patterns and also push the user’s interest focus into the mining process?

v) Frequent pattern mining has many potential applications. Can we eat the effective and efficient frequent pattern mining methods to solve some other interesting data mining problems?

3. FREQUENT PATTERN MINING PROBLEM

The frequent pattern mining problem uses set of items and mining association rules between sets of items.

Let $I = \{i_1, i_2, i_3, \ldots, i_n\}$ be a set of items. An item set $A \subseteq I$ is a subset of items. Hereafter, we write item sets as $A = \{i_1, i_2, \ldots, i_n\}$, i.e. omitting set brackets. Particularly, $B$, an item set with $l$ items are called an $l$-itemset.

A transaction $T = (\text{tid}, A)$ is a tuple where $\text{tid}$ is a transaction-id and $A$ is an item set.

A transaction $T = (\text{tid}, A)$ is said to contain item set $B$ if $B \subseteq A$.

A transaction database $TDB$ is a set of transactions. The support of an itemset $A$ in transaction database $TDB$, denoted as $sup_{TDB}(A)$ or $sup(A)$, is the number of transactions in $TDB$ containing $A$, i.e.,

$$sup(A) = \{ (\text{tid}, B) \mid (\text{tid}, B) \in TDB \land (A \subseteq B) \}$$

Problem statement Given a user-specified support threshold $min\ sup$, $A$ is called a frequent item set or frequent pattern if $sup(A) \geq min\ sup$. The problem of mining frequent item sets is to find the complete set of frequent item sets in a transaction database $TDB$ with respect to a given support threshold $min\ sup$.

Association rules can be derived from frequent patterns. An association rule is an implication of the form $A \Rightarrow B$, where $A$ and $B$ are item sets and $A \cap B = \emptyset$. The rule $A \Rightarrow B$ has support $s$ in a transaction database $TDB$ if $sup_{TDB}(A \cup B) = s$. The rule $A \Rightarrow B$ holds in the transaction database $TDB$ with confidence $c$ where $c = sup(A \cup B) / sup(A)$.

Given a transaction database $TDB$, a support threshold $min\ sup$ and a confidence threshold $min\ conf$, the problem of association rule mining are to find the complete set of association rules that have support and confidence no less than the user-specified thresholds, respectively $B$.

Association rule mining can be divided into two steps. First, frequent patterns with respect to support threshold $min\ sup$ are mined. Second, association rules are generated with respect to confidence threshold $min\ conf$. As shown in many studies, the first step, mining frequent patterns, is significantly more costly in terms of time than the rule generation step.

4. CONCLUSION

As I see, frequent pattern mining is not only used in association rule mining. Instead, frequent pattern mining is the basis for many data mining tasks, such as sequential pattern mining and associative classification. It also has broad applications, such as basket data analysis, cross-marketing, catalog design, sale campaign analysis, web log analysis, etc.

5. REFERENCES