Chapter LVII
Mining Frequent Closed Itemsets for Association Rules

Anamika Gupta
University of Delhi, India

Shikha Gupta
University of Delhi, India

Naveen Kumar
University of Delhi, India

INTRODUCTION

Association refers to correlations that exist among data. Association Rule Mining (ARM) is an important data-mining task. It refers to discovery of rules between different sets of attributes/items in very large databases (Agrawal R. & Srikant R. 1994). The discovered rules help in strategic decision making in both commercial and scientific domains.

A classical application of ARM is market basket analysis, an application of data mining in retail sales where associations between the different items are discovered to analyze the customer's buying habits in order to develop better marketing strategies. ARM has been extensively used in other applications like spatial-temporal, health care, bioinformatics, web data etc (Han J., Cheng H., Xin D., & Yan X. 2007).

Association Rule mining is decomposed in two steps 1) Generate all frequent itemsets (FI) 2) Generate confident rules using the frequent itemsets discovered in first step. The first step is computationally more expensive task and has attracted attention of most researchers. Many researchers have given different algorithms for mining FI (Han J., Cheng H., Xin D., & Yan X. 2007). However set of FI is often very large. Frequent Closed Itemsets (FCI) is a reduced, complete and loss less representation of FI and is often much less in number than FI. *Closed itemset* is an itemset whose support is not equal to support of any of its proper superset (Zaki M. J. & Hsiao C. J. 1999). Closed itemsets with support greater than the user specified support threshold (ms) are *frequent closed itemsets (FCI)*. Thus mining FCI instead of FI in association rule discovery procedure saves computation and memory efforts.

In this article we discuss the importance of mining FCI instead of FI in association rule discovery procedure. We explain different approaches and techniques for mining FCI in datasets.

BACKGROUND

Let D denotes the database of N transactions and I denotes the set of n items in D. A set of one or more items belonging to set I is termed as an itemset. A k-itemset is an itemset of cardinality k. A transaction $T \in D$ contains an itemset and is associated with a unique identifier *TID*. The probability of an itemset X being contained in a transaction is termed as *support* of X.

$$Support(X) = P(X) = \frac{No\_of\_transactions\_containing\_X}{N}$$

An itemset having support greater than the user specified support threshold (*ms*) is termed as *frequent itemset (FI)*.

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

$$Support(X \rightarrow Y) = P(X \cup Y) = \frac{No.\_of\_transactions\_containing\_X \cup Y}{N}$$

Confidence of a rule $X \rightarrow Y$ in D is '*conf*' if *conf*% of transactions in D that contain X, also contain Y. It is computed, as the conditional probability that Y occurs 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 itemsets is *strong* if its *confidence* is greater than the user specified confidence threshold (*mc*).

Three independent groups of researchers (Pasquier N., Bastide Y., Taouil R. & Lakhal L. 1999a), (Zaki M.J. & Hsiao C. J. 1999), (Stumme G., 1999) introduced the notion of mining FCI instead of FI and have given the following definitions of FCI:

Zaki et al. defines *closed itemset* as an itemset whose support is not equal to support of any of its proper superset (Closure Property) (Zaki M. J. & Hsiao C. J. 1999). In other words X is a closed itemset if there exists no proper superset X' of X such that support(X) = support(X'). Closed itemsets with support greater than the user specified support threshold (ms) are *frequent closed itemsets (FCI)*.

Pasquier et. al. defines c*losed itemset* in terms of G*alois closure operator* (Pasquier N., Bastide Y., Taouil R. & Lakhal L. 1999a). *Galois closure operator h(X)* for some $X \subseteq I$ is defined as the intersection of transactions in D containing itemset X. An itemset X is a *closed itemset* if and only if $h(X) = X$.

Stumme G. (1999) defines c*losed itemset* as the intent of *formal concept* where *formal concept* is a core structure in Formal Concept Analysis (FCA), a branch of mathematics based on c*oncepts* and c*oncept hierarchies* (Ganter B. & Wille R. 1999). A *formal concept* (A,B) is defined as a pair of set of objects A (known as *extent*) and set of attributes B (known as *intent*) such that set of all attributes belonging to *extent* A is same as B and set of all objects containing attributes of *intent* B is same as A. That is, no object other than objects of set A contains all attributes of B and no attribute other than attributes in set B is contained in all objects of set A. Stumme G. (1999) discovered that *intent* B of the *Formal Concept* (A,B) represents the closed itemset. Fig. 1 shows an example dataset (D) of five transactions with its itemset lattice and the list of frequent itemsets, frequent closed itemsets in the same example.