Improving the Efficiency of Apriori Algorithm:

19 September 2020 09:29 AM

Many variations of the Apriori algorithm have been proposed that focus on improving the efficiency of the original algorithm

1. **Hash-based technique** (hashing itemsets into corresponding buckets): A hash-based technique can be used to reduce the size of the candidate k-itemsets, Ck, for k > 1

   For example, when scanning each transaction in the database to generate the frequent 1-itemsets, L1, from the candidate 1-itemsets in C1,

   we can generate all of the 2-itemsets for each transaction, hash (i.e., map) them into the different buckets of a hash table structure

Create hash table H_2 using hash function $h(x, y) = ((order\ of\ x) \times 10 + (order\ of\ y))\ mod\ 7$

H_2

bucket address	0	1	2	3	4	5	6
bucket count	2	2	4	2	2	4	4
bucket contents	{I1, I4} {I3, I5}	{I1, I5} {I1, I5}	{I2, I3} {I2, I3} {I2, I3} {I2, I3}	{I2, I4} {I2, I4}	{I2, I5} {I2, I5}	{I1, I2} {I1, I2} {I1, I2} {I1, I2}	{I1, I3} {I1, I3} {I1, I3} {I1, I3}

2. **Transaction reduction** (reducing the number of transactions scanned in future iterations): A transaction that does not contain any frequent k-itemsets cannot contain any frequent (k + 1)-itemsets.

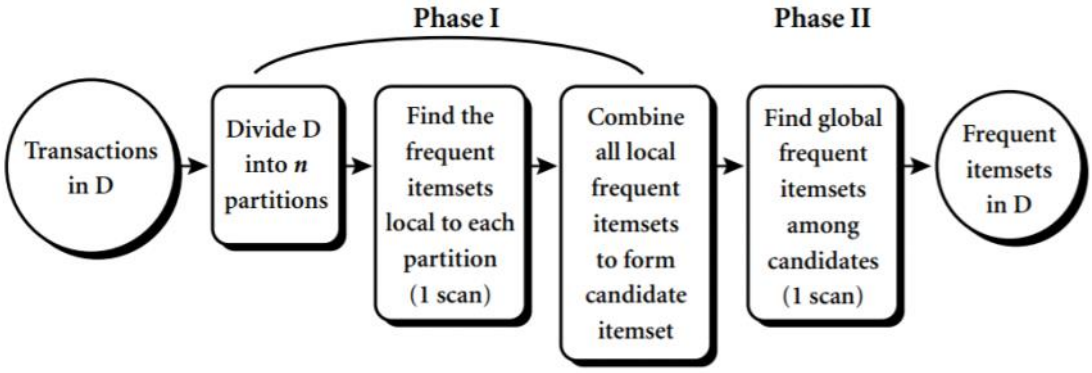
3. **Partitioning (partitioning the data to find candidate itemsets):** A partitioning technique can be used that requires just two database scans to mine the frequent itemsets .

   It consists of two phases.

   In Phase I, the algorithm subdivides the transactions of D into n nonoverlapping partitions.

   A local frequent itemset may or may not be frequent with respect to the entire database, D. Any itemset that is potentially frequent with respect to D must occur as a frequent itemset in at least one of the partitions.

   The collection of frequent itemsets from all partitions forms the global candidate itemsets with respect to D. In Phase II, a second scan of D is conducted in which the actual support of each candidate is assessed in order to determine the global frequent itemsets



4. **Dynamic itemset counting**

   In this variation, new candidate itemsets can be added at any start point, unlike in Apriori, which determines new candidate itemsets only immediately before each complete database scan.

Mining Various Kinds of Association Rules

Multilevel association rules involve concepts at different levels of abstraction.

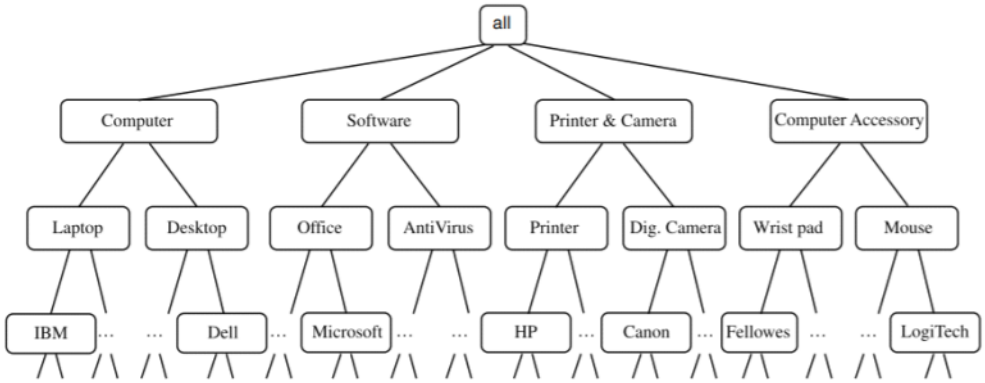
Multidimensional association rules involve more than one dimension or predicate

Quantitative association rules involve numeric attributes that have an implicit ordering among values

For many applications, it is difficult to find strong associations among data items at low or primitive levels of abstraction due to the sparsity of data at those levels. Strong associations discovered at high levels of abstraction may represent commonsense knowledge.

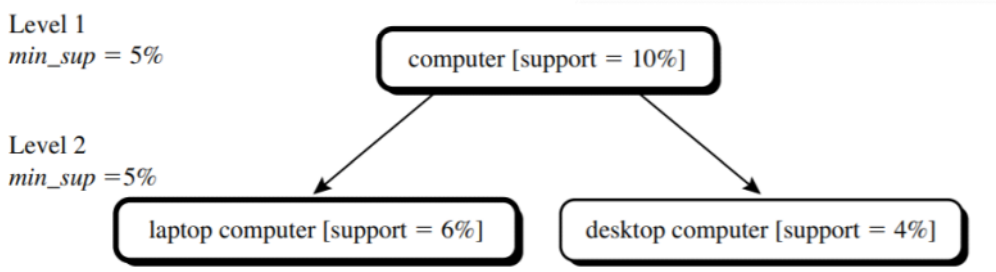
1. **Multilevel association rules**

Data mining systems should provide capabilities for mining association rules at multiple levels of abstraction. The concept hierarchy of following figure has five levels, respectively referred to as levels 0 to 4, starting with level 0 at the root node for all (the most general abstraction level). Here, level 1 includes computer, software, printer&camera, and computer accessory, level 2 includes laptop computer, desktop computer, office software, antivirus software, . . . , and level 3 includes IBM desktop computer, . . . , Microsoft office software, and so on. Level 4 is the most specific abstraction level of this hierarchy.

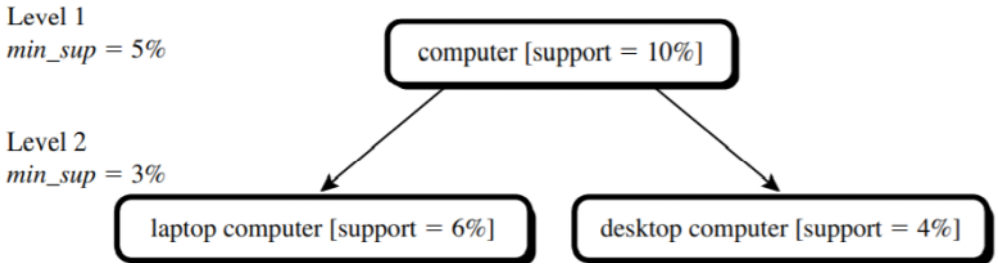


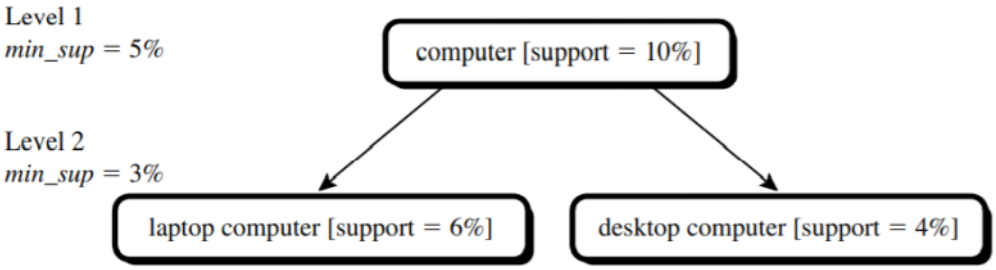
Association rules generated from mining data at multiple levels of abstraction are called multiple-level or multilevel association rules. Multilevel association rules can be mined efficiently using concept hierarchies under a support-confidence framework.

a. **Using uniform minimum support** for all levels (referred to as uniform support): The same minimum support threshold is used when mining at each level of abstraction



b. **Using reduced minimum support at lower levels (referred to as reduced support):** Each level of abstraction has its own minimum support threshold. The deeper the level of abstraction, the smaller the corresponding threshold is. For example, in following figure, the minimum support thresholds for levels 1 and 2 are 5% and 3%, respectively. In this way, "computer," "laptop computer," and "desktop computer" are all considered frequent.





c. **Using item or group-based minimum support (referred to as group-based support):** Because users or experts often have insight as to which groups are more important than others, it is sometimes more desirable to set up user-specific, item, or groupbased minimal support thresholds when mining multilevel rules

Examples:

$$buys(X, \text{“}laptop\ computer\text{”}) \Rightarrow buys(X, \text{“}HP\ printer\text{”})$$
$$[support = 8\%, confidence = 70\%]$$

$$buys(X, \text{“}IBM\ laptop\ computer\text{”}) \Rightarrow buys(X, \text{“}HP\ printer\text{”})$$
$$[support = 2\%, confidence = 72\%]$$

2. Multidimensional association rules

a. single dimensional or intradimensional association rule:

it contains a single distinct predicate (e.g., buys) with multiple occurrences (i.e., the predicate occurs more than once within the rule).

Example

buys(X, “digital camera”) ⇒ buys(X, “HP printer”).

b. multidimensional association rules

Association rules that involve two or more dimensions or predicates can be referred to as multidimensional association rules.

Example

age(X, “20...29”)∧occupation(X, “student”)⇒buys(X, “laptop”).

It contains three predicates (age, occupation, and buys), each of which occurs only once in the rule. Hence, we say that it **has no repeated predicates.** Multidimensional association rules with no repeated predicates are called **interdimensional association rules**

age(X, “20...29”)∧buys(X, “laptop”)⇒buys(X, “HP printer”)

We can also mine multidimensional association rules with repeated predicates, which contain multiple occurrences of some predicates. These rules are called hybrid-dimensional association rules. The above rule is the example of hybrid-dimensional association rule.

3. Quantitative Association Rules

Quantitative association rules are multidimensional association rules in which the numeric attributes are dynamically discretized during the mining process.

Aquan1 ∧Aquan2 ⇒ Acat where Aquan1 and Aquan2 are tests on quantitative attribute intervals (where the intervals are dynamically determined), and Acat tests a categorical attribute from the task-relevant data.

Example :

age(X, “30...39”)∧income(X, “42K...48K”)⇒buys(X, “HDTV”)