**EX NO: 3 DATA MINING LABORATORY**

**DATE: 18/09/2020 PRELAB – ASSOCIATION MINING**

**Aim:**

To study about the concepts of association mining.

**Questions and Answers:**

**1. What are frequent patterns?**

Frequent patterns are patterns (e.g., item sets, subsequences, or substructures) that appear frequently in a data set. For example, a set of items, such as milk and bread, that appear frequently together in a transaction data set is a frequent itemset. A subsequence, such as buying first a PC, then a digital camera, and then a memory card, if it occurs frequently in a shopping history database, is a (frequent) sequential pattern. A substructure can refer to different structural forms, such as subgraphs, subtrees, or sublattices, which may be combined with item sets or subsequences. If a substructure occurs frequently, it is called a (frequent) structured pattern. Finding frequent patterns plays an essential role in mining associations, correlations, and many other interesting relationships among data.

**2. What is a frequent item-set?**

A frequent itemset is an itemset whose support is greater than some user-specified minimum support (denoted Lk, where k is the size of the itemset) A candidate itemset is a potentially frequent itemset (denoted Ck, where k is the size of the itemset)

**3. What are the interestingness measures for association rules?**

The two interestingness measures for association rule are **support** and **confidence.**

**4. Define the above measures formally.**

Let I = {I1, I2,...,Im} be an itemset. Let D, the task-relevant data, be a set of database transactions where each transaction T is a nonempty itemset such that T ⊆ I. Each transaction is associated with an identifier, called a TID. Let A be a set of items. A transaction T is said to contain A if A ⊆ T. An association rule is an implication of the form A ⇒ B, where A ⊂ I, B ⊂ I, A 6= ∅, B 6= ∅, and A ∩B = φ. The rule A ⇒ B holds in the transaction set D with support s, where s is the percentage of transactions in D that contain A ∪B (i.e., the union of sets A and B say, or, both A and B). This is taken to be the probability, P(A ∪B). 1 The rule A ⇒ B has confidence c in the transaction set D, where c is the percentage of transactions in D containing A that also contain B. This is taken to be the conditional probability, P(B|A). That is,

**support(A⇒B) = P (A ∪ B)**

**confidence(A⇒B) = P (B|A)**

**5. What are the two steps of association rule mining?**

The two steps of association rule mining are:

1. **Frequent Itemset Generation**: Find all item sets whose support is greater than or equal

to the min support.

2. **Rule generation**: generate strong association rules from the frequent itemset whose

confidence greater than or equal to min confidence.

**6. What is the total number of frequent item-sets that a frequent item-set of length 200 has?**

* A frequent itemset of length 200, such as {a1, a2, . . . , a200}, contains 200 C1 = 200 frequent 1- item-sets: {a1}, {a2}, . . . , {a200};
* 200C2 frequent 2-itemsets: {a1, a2}, {a1, a3} , …. ,{a199, a200}; and so on.
* The total number of frequent item-sets that it contains is,

= 200C1 + 200C2 + 200C3 + …. 200C200

= 2200

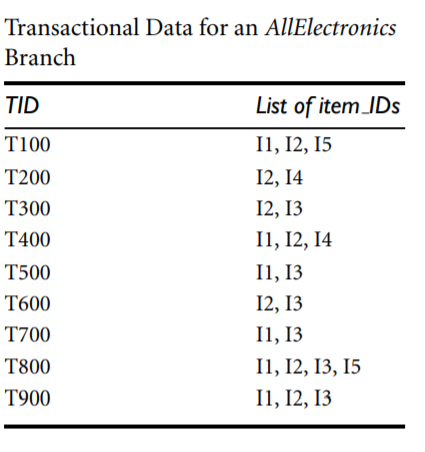
= 1606938044258990275541962092341162602522202993782792835301376

= 1.6 x 1060

**7. What is apriori property for finding frequent item-set by candidate generation method? Give an example.**

* The Apriori property is based on the following observation. By definition, if an itemset I does not satisfy the minimum support threshold, min sup, then I is not frequent, that is, P(I) < min sup. If an item A is added to the itemset I, then the resulting itemset (i.e., I ∪A) cannot occur more frequently than I. Therefore, I ∪A is not frequent either, that is, P(I ∪A) < min sup
* A two-step process is followed, consisting of **join** and **prune** actions.

**Example:**

****

**Join:**

C3 = L2 ⋈ L2

= {{I1, I2}, {I1, I3}, {I1, I5}, {I2, I3}, {I2, I4}, {I2, I5}} ⋈ {{I1, I2}, {I1, I3}, {I1, I5}, {I2, I3}, {I2, I4}, {I2, I5}} = {{I1, I2, I3}, {I1, I2, I5}, {I1, I3, I5}, {I2, I3, I4}, {I2, I3, I5}, {I2, I4, I5}}.

**Prune using the Apriori property**:

All nonempty subsets of a frequent itemset must also be frequent.

* The 2-item subsets of {I1, I2, I3} are {I1, I2}, {I1, I3}, and {I2, I3}. All 2-item subsets of {I1, I2, I3} are members of L2. Therefore, keep {I1, I2, I3} in C3.
* The 2-item subsets of {I1, I2, I5} are {I1, I2}, {I1, I5}, and {I2, I5}. All 2-item subsets of {I1, I2, I5} are members of L2. Therefore, keep {I1, I2, I5} in C3.
* The 2-item subsets of {I1, I3, I5} are {I1, I3}, {I1, I5}, and {I3, I5}. {I3, I5} is not a member of L2, and so it is not frequent. Therefore, remove {I1, I3, I5} from C3.
* The 2-item subsets of {I2, I3, I4} are {I2, I3}, {I2, I4}, and {I3, I4}. {I3, I4} is not a member of L2, and so it is not frequent. Therefore, remove {I2, I3, I4} from C3.
* The 2-item subsets of {I2, I3, I5} are {I2, I3}, {I2, I5}, and {I3, I5}. {I3, I5} is not a member of L2, and so it is not frequent. Therefore, remove {I2, I3, I5} from C3.
* The 2-item subsets of {I2, I4, I5} are {I2, I4}, {I2, I5}, and {I4, I5}. {I4, I5} is not a member of L2, and so it is not frequent. Therefore, remove {I2, I4, I5} from C3

Therefore, C3 = {{I1, I2, I3}, {I1, I2, I5}} after pruning

**Result:**

The concepts of association mining have been learnt successfully.