Association Rule Mining

Association rule mining finds interesting associations and relationships among large sets of data items. This rule shows how frequently a itemset occurs in a transaction. A typical example is Market Based Analysis.

Market Based Analysis is one of the key techniques used by large relations to show associations between items.It allows retailers to identify relationships between the items that people buy together frequently.

Given a set of transactions, we can find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction.

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Before we start defining the rule, let us first see the basic definitions.

Support Count– Frequency of occurrence of a itemset.
Here Support_Count({Milk, Bread, Diaper})=2

Frequent Itemset – An itemset whose support is greater than or equal to minsup threshold.
Association Rule – An implication expression of the form X -> Y, where X and Y are any 2 itemsets.
Example: {Milk, Diaper}->{Beer}

Rule Evaluation Metrics –

- **Support(s) –**
  The number of transactions that include items in the {X} and {Y} parts of the rule as a percentage of the total number of transaction.It is a measure of how frequently the collection of items occur together as a percentage of all transactions.
- **Support = (X+Y)/total –**
  It is interpreted as fraction of transactions that contain both X and Y.
- **Confidence(c) –**
  It is the ratio of the no of transactions that includes all items in {B} as well as the no of transactions that includes all items in {A} to the no of transactions that includes all items in {A}.
- **Conf(X=>Y) = Supp(X U Y) / Supp(X) –**
  It measures how often each item in Y appears in transactions that contains items in X also.
- **Lift(l) –**
  The lift of the rule X=>Y is the confidence of the rule divided by the expected confidence, assuming that the itemsets X and Y are independent of each other.The expected confidence is the confidence divided by the frequency of {Y}.

- **Lift(X=>Y) = Conf(X=>Y)/ Supp(Y) –**
  Lift value near 1 indicates X and Y almost often appear together as expected, greater than 1 means they appear together more than expected and less than 1 means they appear less than expected.Greater lift values indicate stronger association.

Example –
From the above table, {Milk, Diaper}=>{Beer}
s= ({Milk, Diaper, Beer})

= 2/5

= 0.4

c= (Milk, Diaper, Beer)/(Milk, Diaper)

= 2/3

= 0.67

l= Supp({Milk, Diaper, Beer})/Supp({Milk, Diaper})*Supp({Beer})

= 0.4/(0.6*0.6)

= 1.11

The Association rule is very useful in analyzing datasets. The data is collected using bar-code scanners in supermarkets. Such databases consists of a large number of transaction records which list all items bought by a customer on a single purchase. So the manager could know if certain groups of items are consistently purchased together and use this data for adjusting store layouts, cross-selling, promotions based on statistics.