Programme	**:** Bachelor of Science
Subject	**:** Computer Science
Semester	**:** V
Course Code	**:** CSD102
Course Title	**:** Data Mining
Unit V	**:** Association Analysis
Module Name	**:** Introduction to Association Analysis, Frequent Patterns, Market Basket Analysis, Association Rule Mining - Problem Definition, Important Concepts

Notes

Introduction to Association Analysis

Association analysis is all about uncovering data relationships and using this asset to gain competitive advantage. It is a means to discover relationships in large data sets.

Hidden data relationships will be expressed as a collection of association rules and frequent item sets with association analysis. While association rules suggest a strong relationship that exists between items, frequent items are a collection of items that frequently occur together.

Frequent Patterns

Frequent patterns are patterns (such as item sets), that appear in a data set frequently. For example, a set of items, such as milk and bread that appear frequently together in a transaction data set (for example, collections of items bought by customers, or details of a website frequentation) is a frequent item set. Finding such frequent patterns play an essential role in mining associations, and many other interesting relationships among data. Moreover, it helps in data classification, clustering, and other data mining tasks as well. Thus, frequent pattern mining has become an important data mining task and a focused theme in data mining applications.

Market Basket Analysis

Frequent item set mining leads to the discovery of associations among items in large transactional or relational data sets. With massive amounts of data continuously being collected and stored, many industries are becoming interested in mining such patterns from their databases.

The discovery of interesting association relationships among huge amounts of business transaction records can help in many business decision-making processes, such as catalogue design, cross-marketing, and customer shopping behaviour analysis. A typical example of frequent item set mining or association rule mining is ***Market Basket Analysis***.

This process is based on transactional data, which are huge amounts of records of individual purchases. This process analyses customer buying habits by finding associations between different items that customers place in their shopping baskets. The discovery of such associations can help retailers develop marketing strategies by gaining insight into which items are frequently purchased together by customers. For instance, if customers are buying milk, how likely are they to also buy bread (and what kind of bread) on the same trip to the supermarket. Such information can lead to increased sales by helping retailers do selective marketing and plan their shelf space.

Example - Market Basket Analysis: Suppose, as a manager of the electronics company, AllElectronics branch, you would like to learn more about the buying habits of your customers. Specifically, you wonder, “Which groups or sets of items are customers likely to purchase on a given trip to the store?” To answer your question, market basket analysis may be performed on the retail data of customer transactions at your store. You can then use the results to plan marketing or advertising strategies, or in the design of a new catalog.

For instance, market basket analysis may help you design different store layouts. In one strategy, items that are frequently purchased together can be placed in proximity to further encourage the combined sale of such items. If customers who purchase computers also tend to buy antivirus software at the same time, then placing the hardware display close to the software display may help increase the sales of both items.

Association Rule Mining

Association rules are simple If/Then statements that help discover relationships between seemingly independent items of relational databases or other data repositories. Association rule mining is a procedure which aims to observe frequently occurring patterns or associations from datasets.

An association rule has two parts:

a. an antecedent (if) and
b. a consequent (then).

An antecedent is something that’s found in data, and a consequent is an item that is found in combination with the antecedent.

Have a look at this rule for instance:

“If a customer buys bread, he’s 70% likely of buying milk.”

In the above association rule, bread is the antecedent and milk is the consequent. Simply put, it can be understood as a retail store’s association rule to target their customers better. If the above rule is a result of thorough analysis of some data sets, it can be used to not only improve customer service but also improve the company’s revenue.

Association rule mining is a popular and well researched method for discovering interesting relations between variables in large databases.

It is intended to identify strong rules discovered in databases using different measures of interestingness. Based on the concept of strong rules, Rakesh Agrawal et al. introduced association rules.

Association Rule Mining - Problem Definition

The problem of association rule mining is defined as:

Let **$I = \{ i_1, i_2, \ldots.., i_n \}$** be a set of n-binary attributes called ***items***.

Let **$D = \{ t_1, t_2, \ldots.., t_m \}$** be a set of transactions called the ***database***.

Each transaction in ***D*** has a unique transaction ID and contains a subset of the items in ***I***.

A ***rule*** is defined as an implication of the form **X => Y** where and X, Y ⊆ I and X ⋂ Y = ∅.

The sets of items (for short item-sets) **X** and **Y** are called ***antecedent*** (left-hand-side or LHS) and ***consequent*** (right-hand-side or RHS) of the rule respectively.

Example:

Let’s consider an example from the supermarket domain.

The set of items is ***I = { Milk, Bread, Butter, Beer }*** and a small database containing the items (**1** codes presence and **0** absence of an item in a transaction) is shown in the table.

An example of association rule for the supermarket could be ***{ Butter, Bread } => { Milk }*** meaning that if butter and bread are bought, customers also buy milk. Example database with 4 items and 5 transactions (snapshot of a database with 100’s of transactions).

Transaction ID	**Milk**	**Bread**	**Butter**	**Beer**
1	1	1	0	0
2	0	0	1	0
3	0	0	0	1
4	1	1	1	0
5	0	1	0	0

Important Concepts

Rule ***Support*** and ***Confidence*** are two measures of rule interestingness. They respectively reflect the usefulness and certainty of discovered rules. The Support data of an item set X is defined as the proportion of transaction in the set which contain the item set.

In the example database, the item set { Milk, Bread, Butter } has a support of ***1 / 5 = 0.2*** since it occurs in 20% of all transactions (1 out of 5 transactions).

$$\textbf{supp(X)} = \frac{\textit{no. of transactions which contain the item set X}}{\textit{total no. of transactions}}$$

The confidence of a rule is defined as

$$\textbf{conf (X => Y) = supp (X} \cup \textbf{Y) / supp (X)}$$

For example, the rule **{ Butter, Bread } => { Milk }** has a confidence of **0.2 / 0.2 = 1.0** in the database, which means that for 100% of the transactions containing butter and bread the rule is correct (100% of the times a customer buys butter and bread, milk is bought as well).

Confidence can be interpreted as an estimate of the probability **P(*Y*|*X*),** the probability of finding the RHS of the rule in transactions under the condition that these transactions also contain the LHS.

Frequent Itemsets:

A set of k items is called a k-itemset.
For example, **I = { Milk, Bread, Butter}** is a 3-itemset.

The occurrence frequency of an item set (or simple the count) is the number of transactions that contain the itemset. For example, the database D may contain 50 transactions that contain itemset I. An itemset whose count (or probability) is greater than some pre-specified threshold (Example, Threshold >=2) is called a frequent itemset. A set of all frequent k-itemsets will be denoted as L_k.

How are we going to find interesting association rules from the database D?

It will be a two-step process:

i. Find all frequent item sets (each of these item sets will occur atleast as frequently as pre-specified by the minimum support threshold).
ii. Generate strong association rules from the frequent itemsets (these rules will satisfy both minimum support threshold and/or minimum confidence threshold).

Typically, *association rules are considered interesting* if they satisfy both a minimum support threshold and a minimum confidence threshold. These thresholds can be a set by users or domain experts.

Example: Association Rule: The information that customers who purchase computers also tend to buy antivirus software at the same time is represented in the following association rule:

***Computer ➔ antivirus software* [*support* = 2%, *confidence* = 60%]**

A support of 2% for the above Association Rule means that 2% of all the transactions under analysis show that computer and antivirus software are purchased together. A confidence of 60% means that 60% of the customers who purchased a computer also bought the software.