Apriori Algorithm

The apriori algorithm is used for finding frequent itemsets and generating association rules. Its primary objective is to discover relationships or associations between different objects in a database.

The apriori algorithm's main goal is to find connections between objects, known as association rules. It is also referred to as frequent pattern mining.

The algorithm operates on a database containing a large number of transactions. Each transaction consists of a set of items. The apriori algorithm works in an iterative manner, starting with frequent individual items and gradually extending to larger itemsets.

Examples

i. We take an example to understand the concept better. You must have noticed that the Pizza shop seller makes a pizza, soft drink, and breadstick combo together. He also offers a discount to their customers who buy these combos. Do you ever think why does he do so? He thinks that customers who buy pizza also buy soft drinks and breadsticks. However, by making combos, he makes it easy for the customers. At the same time, he also increases his sales performance.
ii. Similarly, you go to Big Bazar, and you will find biscuits, chips, and Chocolate bundled together. It shows that the shopkeeper makes it comfortable for the customers to buy these products in the same place.

Algorithm

1. Set the minimum support threshold - min frequency required for an itemset to be "frequent".
2. Identify frequent individual items - count the occurence of each individual item.
3. Generate candidate itemsets of size 2 - create pairs of frequent items discovered.
4. Prune infrequent itemsets - eliminate itemsets that do no meet the threshold levels.
5. Generate itemsets of larger sizes - combine the frequent itemsets of size 3,4, and so on.
6. Repeat the pruning process - keep eliminating the itemsets that do not meet the threshold levels.
7. Iterate till no more frequent itemsets can be generated.
8. Generate association rules that express the relationship between them - calculate measures to evaluate the strength & significance of these rules.

Advantages

i. Simplicity & ease of implementation.
ii. The rules are easy to human-readable * interpretable.
iii. Works well on unlabelled data.
iv. Flexibility & customisability.
v. Extensions for multiple use cases can be created easily.
vi. The algorithm is widely used & studied.

Disadvantages

i. Computational complexity.
ii. Time & space overhead.
iii. Difficulty handling sparse data.
iv. Limited discovery of complex patterns.
v. Higher memory usage.
vi. Bias of minimum support threshold.
vii. Inability to handle numeric data.
viii. Lack of incorporation of context.

How can we improve the Apriori Algorithm's efficiency?

Many methods are available for improving the efficiency of the algorithm.

a) **Hash-Based Technique:** This method uses a hash-based structure called a hash table for generating the k-itemsets and their corresponding count. It uses a hash function for generating the table.

b) **Transaction Reduction:** This method reduces the number of transactions scanned in iterations. The transactions which do not contain frequent items are marked or removed.

c) **Partitioning:** This method requires only two database scans to mine the frequent itemsets. It says that for any itemset to be potentially frequent in the database, it should be frequent in at least one of the partitions of the database.

d) **Sampling:** This method picks a random sample S from Database D and then searches for frequent itemset in S. It may be possible to lose a global frequent itemset. This can be reduced by lowering the min_sup.

e) **Dynamic Itemset Counting:** This technique can add new candidate itemsets at any marked start point of the database during the scanning of the database.

Components of the Apriori algorithm

There are three major components of the Apriori algorithm which are as follows.

1. Support
2. Confidence
3. Lift

Support

Support denotes the average popularity of any product or data item in the data set. We need to divide the total number of transactions containing that product by the total number of transactions.

Support (Men's wear)= (transactions relating MW) / (total transaction)
= 300/5000
= 16.67 %

Confidence

Confidence is the sum average of transactions/data items present in pairs/combinations in the universal dataset. To find out confidence, we divide the number of transactions that comprise both men's & women's wear by the total number of transactions.
Hence,

Confidence = (Transactions with men's & women's wear) / (total transaction)
= 250/5000
= 5%

Lift

It helps find out the ratio of the sales of women's wear when you sell men's wear. The mathematical equation of lift is mentioned below.

Lift = (Confidence (Men's wear- women's wear)/ (Support (men's wear)
= 20/18
= 1.11

Applications:

Some of the applications of Apriori Algorithm are:

1. Medical

Hospitals are generally trashed with data every day and need to retrieve a lot of past data for existing patience. Apriori algorithm help hospitals to manage the database of patients without jinxing it with other patients.

2. Education

The educational institute can use the Apriori algorithm to store and monitor students' data like age, gender, traits, characteristics, parent's details, etc.

3. Forestry

On the same line as the education and medical industry, forestry can also use the Apriori algorithm to store, analyze and manage details of every flora and fauna of the given territory.

4. New Tech Firms

Tech firms use the Apriori algorithm to maintain the record of various items of products that are purchased by various customers for recommender systems.

5. Mobile Commerce

Big data can help mobile e-commerce companies to deliver an easy, convenient and personalized shopping experience. With the Apriori algorithm, the real-time product recommendation accuracy increases, which creates an excellent customer experience and increases sales for the company.