MARKET BASKET ANALYSIS USING APRIORI ALGORITHM

Divyanshu Yadav[1],Aditya Swarup[2],

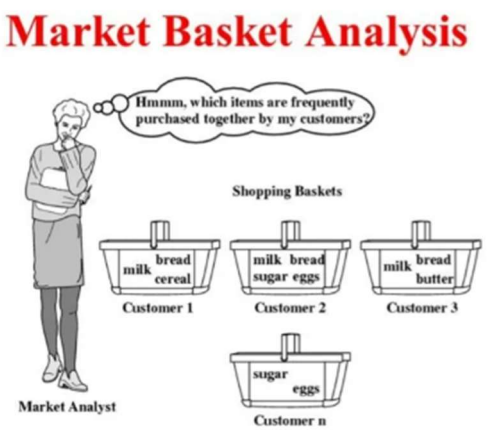
Preetish Majumdar[3], Manas Singh[4], Anurag Jain[5]

Dr. Godfrey Winster S[6]

Department of Computing Technologies SRM Institute of Science and Technology

[1]dd8353@srmist.edu.in [6]godfreys@srmist.edu.in

ABSTRACT-Market Basket analysis is the data mining process, in which the large data can be mined and several steps are involved in the mining process like data collection, preprocessing, algorithm for mining, etc. The main aim of this process is to provide only useful data to the customers to make correct decision. Market Basket Analysis identifies the relationship associatedwith different data items. We supply a large dataset collected from a store or an industry. Several industries are using this method to improve their catalog design and cross-selling of products and thus helps in making better business decisions. The Market Basket Analysis identifies the association between items thus finding the customer buying pattern. This will help the retailers to expand their business strategies. It will find the interesting hidden patterns from the large dataset and assist the owner to make business decisions. The association rules can be used in various fields like bioinformatics, education field, marketing, nuclear science, etc. There are many algorithms available to perform these tasks but they work on static data and do not capture the changes made to the data



I. INTRODUCTION

We generate a huge amount of data in our day-to-day life. This data is growing most frequently butnot every piece of information is useful. It is necessary to get useful information from the stored data. The process of extracting useful information from a large dataset is called Data mining or Knowledge Discovery in Data (KDD). This process involves data selection, data preprocessing, transformation, mining, and interpretation.

Market basket analysis is also called association rule mining. In this process large amount of data is being maintained in database in several various fields like marketing, banking, medical etc. For example we can take a grocery items such as

a customer purchase the item called bread then the probability of getting the second item is jam or milk. This will tell about the purchasing pattern of the customer.

Using this algorithm, it is easy for customers to buy the items, also it is useful for the retailer to understand the customer purchasing pattern, and retailer can use this information to make the decision according to the most selling frequent item set.

Association rule mining is one in all the important tasks of data mining. The association rule is represented in the form of A→B, where A is referredto as Antecedent and B as Consequent. It means that a customer buying product A is most likely to buy B with %C where C is called confidence. This process helps the owners to study the buying patterns of the customers. This implies that frequently purchased items are placed together within the catalog. This analysis helps to place the regularly purchased items close to each other.

Example –a) Customers purchasing computers are most likely to purchase antivirus also.

b) People purchasing cigarettes can also buy matchboxes.

Association Rule mining can be performed based on the below parameters.

1. **Support:**

It checks that how frequently an item is purchased and how frequently the item isoccurring in the transactions.

$$\text{Support}(A \rightarrow B) = \frac{\text{Transactions containing item} - \text{ and item} - B}{\text{Total no Transactions}}$$

If an item-set qualifies the minimum support value then it is considered for the calculation else item-set is ignored from the calculation. Higher the support value indicates that item is more frequently to occur.

2. **Confidence:**

The conditional probability of both the items occurring together in a transaction. Probability of purchasing an item-A over item-B and vice-versa.

$$\text{Confidence } (A \rightarrow B) = \frac{\text{Number of Transaction that Contains A and B}}{\text{Total Transactions that contain A}}$$

3. **Lift:**

Probability of purchasing first item over second item independent of each other.

$$\text{Lift } (A \rightarrow B) = \frac{\text{Confidence of A and B}}{\text{Support (B)}}$$

- Lift $(A \rightarrow B) = 1$ means that there is no correlation with the given item-set.
- Lift $(A \rightarrow B) > 1$ means that there is a positive correlation within the item-set, i.e. Products in the item-set, A, and B, are more likely to be bought together.
- Lift $A(\rightarrow B) < 1$ means that there is a negative correlation within the item-

set, i.e. products in item-set, A, and B, are unlikely to be bought together.

4. **Leverage:**
   It contrasts the items occurring together in the item-set and the expected probability of item-set.
   Leverage (A→B) = P (A and B) - [P (A) *P(B)]

5. **Conviction:**
   It is expressed as an item-A occurs without B.
   Conviction (A→B) = $\frac{1-\text{Support (B)}}{1-\text{Confidence(A}\rightarrow\text{B)}}$

Apriori algorithm:

Apriori algorithm is most widely used for finding frequent items in the transactions. It uses prior knowledge of the item sets. It states that all non-empty subsets must be frequent. The algorithm starts with simple rules and adds products to the item-set then selects the subset of the item-set by a predetermined support value. The items that qualify the given support value are considered in the subset and the rest are ignored from the subset this is called the rule generation process. This is an iterative process entire dataset is scanned to make the possible combination of items.

Item-set	Support count
Bread, Cheese	4
Bread, Milk	4
Bread, Jam	1
Bread, Sugar	2
Cheese, Milk	4
Cheese, Jam	2
Cheese, Sugar	2
Milk, Jam	0
Milk, Sugar	1
Jam, Sugar	0

Let us understand with an example:

Transaction ID	Items
T1	Bread, Cheese, Sugar
T2	Cheese, Jam
T3	Cheese, Milk
T4	Bread, Cheese, Jam
T5	Bread, Milk
T6	Cheese, Milk
T7	Bread, Milk
T8	Bread, Cheese, Milk, Sugar
T9	Bread, Cheese, Milk

Minimum support value is 2

Minimum confidence value is 60%

1Item-set

Item-set	Support count
Bread	6
Cheese	7
Milk	6
Jam	2
Sugar	2

2 Item-set

By comparing 2 Item-set with Minimum support value, few item-sets will be ignored because they do not qualify minimum support value.

Item-set	Support count
Bread, Cheese	4
Bread, Milk	4
Bread, Sugar	2
Cheese, Milk	4
Cheese, Jam	2
Cheese, Sugar	2

Now we will generate 3 Item-set

Item-set	Support count
Bread, Cheese, Milk	2
Bread, Cheese, Sugar	2

Here we cannot form 4 item sets because the 4th item is not common in the above set. As there are no frequent items we stop generating the 4th item-set and we have generated all possible frequent item-sets.

Now we will calculate the confidence for generated rules. Here we will check the possibility of purchasing 3rd item over the first 2 items.

[Bread, Cheese] => [Milk]

Confidence =

[Support (Bread and Cheese and Milk) / support (Bread and Milk)] * 100

= [2 /4] * 100

= 50

II. EXISTING SYSTEM

The existing system work on the basis of static data. It will not automatically changes with the time. Data mining is the algorithm which is existing system used, it is takes more time as it will scan the data for several time.

III. PROPOSED SYSTEM

The drawback of this we are used here is association rule mining. It is more faster than data mining. Association will provide the association between the two itemset. It will provide the relation between the two or three itemset. So it is easy to be understood by the retailer and also easy for the customer to buy the products.

IV. CONCLUSION

This paper provides a discussion on Association rules and use of Apriori principle used for Market basket Analysis. Data mining provides the way to use précised information from the large dataset. Association rules find the relationship between items by analyzing the data and provide the accurate solution to the retailer to make better business decisions.

V. REFERENCES

1. I., 2016, “Data mining” [online] http://etonlinedigitallibrary.com/bitstream/123456789/2358/1/1306.pdf, Accessed on FEB 2021
2. International Journal of Engineering Research & Technology (IJERT)ISSN:

2278-0181IJERTV9IS090345 Vol. 9
Issue 09, September-202

3. Shankar, D.D., & Shukla, V.K. (2018). Result Analysis of Cross-Validation on low embedding Feature-based Blind Steganalysis of 25 percent on JPEG images using SVM. 2018 International Conference on Circuits and Systems in DigitalEnterprise Technology (ICCSDET), 1-5.
4. Maske A, Joglekar B (2018)Survey on frequent item-set mining approaches in market basket analysis. 2018 fourth international conference on computingcommunication control and automation (ICCUBEA). pp 1–5.