Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 5- Issue 2, Paper 47 August 2022

Market Basket Analysis using Apriori Algorithm

M.Praveena , *V.Jahnavi Sai Sree ,*
T.Kalyani, P.Sunayana
, A.Abhi Lakshmi, J Jayanthi
Department Of Computer Science Andengineering,
Kalasalingam Academy Of Researchand Education,
Virudhunagar, Tamil Nadu,

Abstract— Market Basket Analysis is a useful tool in the retail industry that can assist a market owner in growing their firm and improving their sales marketing strategy. This is entirely accomplished by association rule mining, which compares client behavior to market purchases. It analyzes the buying habits of customers and generates a list of frequently purchased items.
Instead of reading large amounts of transactional data manually, it is simple to find the most popular itemset and worst item combination after creating a frequent itemset. The creation of frequent itemsets will increase market strategy, product placement, and many other aspects. As a result, goods sales improve, and anyone can run a profitable and successful business.
Keywords— Market basket analysis, Association Rulemining, Frequent Item sets, Apriori Algorithm.

I. INTRODUCTION

The problem for organisations that have invested heavily in consumer data collection these days is figuring out how to extract valuable information from their massive customer transactional databases in order to profit in the supermarket. In many retailers, market basket research has mostly been utilised to identify product associations and build a retailer's promotion programme on them.
A marketer must understand and meet the needs of his orher customers. One method of determining whether products can be placed next to one another is to do a market basket study. Market basket analyses provide retailers with useful information on linked sales on a stock-by-stock basis. Customers who purchase milk frequently also purchase other milk-related products such as bread and coffee powder. It's understandable.
Market basket analysis is used to determine which products are purchased in pairs and to plan the supermarket layout. As a result, market customer behaviour must be examined, which can be done using various data mining approaches. It's simple to make good decisions regarding positioning, pricing, and profitability, and it's also easy to see if there are any successful items that don't include any connected aspects[1]. However, the execution of the Apriori method necessitates multiple examinations of the database, resulting in overload. As a result, we try to the improvedapriori algorithm, which uses a hash table, improves the performance of the Apriori algorithm by removing limitations.

II. PROBLEM STATEMENT

People now purchase their daily necessities from a nearby market. Many supermarkets and big bazaars provide merchandise to their customers. The challenge that many merchants encounter is product allocation. Because shop managers are unaware of people's purchasing habits, they have no idea which products should be placed near each other in their Supermarket/Big Bazaar. Shop managers can use this programme to analyse the strong correlations between products, which then allows them to putproducts that go well together close together. Decisions are also made on which items to stock more of, cross-selling, up-selling, and retail shelf placement.

III. RELATED WORKS

The structure and computing speed on the binary space are used to determine the Apriori's growth. Despite the fact that a variety of ways to frequent mining have been offered, none of them have a clear bound for executing multiple attribute value of items within the non-binary space. As a transaction, multiple attribute value brings information in the form of a group of attributes in a row. There are multiple sets of valuesfor each attribute. Item A, for example, has the values for x1 and y1, while item B contains the values for x2 and y2.
Because it simply evaluates the presence or absence of data and does not discriminate the data values, computational of

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 5- Issue 2, Paper 47 August 2022

frequent items in binary space cannot differentiate between the values x1 and x2. The major goal of this research is to implement Apriori-based Association rule mining in the non-binary search space. The frequency item is determined by the attribute's value. This differs from common apriori in that the calculations are based on the frequency of the transaction while taking into account all of the item's potential worth.

Our suggested Apriori approach aims to overcome the limitation of data representation in Association rule mining,where data is generally described in binary space form as a market basket or transaction.

The project's goal is to use indicators like support confidence, lift, leverage, and conviction to uncover the items that customers buy regularly and the items that they buy together frequently. The goal of the research was to combine a hashing function with the Apriori algorithm, and hash tables were created to deal with the products found in hash tables with frequent itemsets.

This was done to improve the Apriori algorithm's efficiency while reducing the size of the itemset. The primary goal of Market Basket Analysis is to improve market and sales plan efficiency by analysing consumer transactional data collected during the sales transaction. On the basis of support and confidence, to identify the frequent things during or after the transaction. As the minimum support is reduced, the memory usage of both algorithms for generating association rules from frequent item sets increases exponentially. Applying When compared to apriori, the hashing data structure in apriori reduces the size. As a result, the execution time is shortened. As a result, Apriori's performance with hashing is improved in terms of execution time and memory space.

IV. PROPOSED SYSTEM

Our Market Basket Analysis technique collects Frequent Itemset from a set of Transactions, which are then grouped by semantic direction and intensity. The suggested system consists of software that collects frequent items from several transactional databases and then analyses the data for association mining [2]. The Apriori Algorithm is one of the many algorithms available for association rule mining, and it is employed in our Basket Analysis system. The software in our Basket Analysis system will set support and confidence for association rule mining, which will default to using the Apriori algorithm for frequent itemset mining.

V. DATA DESCRIPTION

The information was gathered from a website. This is thegrocery data, which includes a list of the things that customers purchased.

The number of items in a basket is shown on the left, followed by Item 1, 2, 3, etc., which is a list of the items.

VI. APRIORI ALGORITHM

Apriori is a popular, essential, and scalable approach for mining frequently occurring Itemsets and association rules. Agrawal and Srikant introduced Apriori in 1993. To locate all frequent itemsets in each database, the Apriori algorithm is utilised. The apriori algorithm involves scanning the database multiple times. It searches the database using a breath-first strategy. Many industries employ the apriori algorithm for transactional operations, and it may be used in real-time applications such as (shopping malls, general stores, grocery stores, and so on) by collecting the items purchased by customers over time so that frequent items can be generated. Minimal support and minimum confidence are required by the Apriori algorithm. First, we can see if the items are greater than or equal to the minimum support, and then we can look for the frequently used item set. The second point to mention is that minimum confidence is used to create association rules.

VII. ASSOCIATION RULE MINIG

In general, association rule mining is used to extract interesting correlations, patterns, and associations among sets of products in a dataset. It is made up of two primary metrics:support and confidence. Support can be described in terms of a minimum value or a value that is greater than the minimumvalue. The number of times a set of claims has become true can be used to determine confidence. The rules that satisfy the minimum support and confidence are discovered by

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 5- Issue 2, Paper 47 August 2022

association rule mining. The mining of association rules is a two-step procedure: The first step is to locate a frequently used item with less than minimum support. Step 2 is to put all of the products that are used frequently together. We look at several of these useful indicators in this algorithm, such as support, confidence, lift, leverage, and conviction [3]. Association rules are "if-then" statements that illustrate the likelihood of associations between data items in huge data sets in a variety of databases.

- **Support**

The percentage of customers that purchased item X and itemY is known as support.
Support = Probability (X & Y) Support = (Item X + ItemY) /(total number of transactions).

- **Confidence**

The percentage of purchasers who bought item X and item Ytogether, i.e. confidence, is the strength of implication of a rule; it is the percentage of purchasers who bought item X and item Y together, i.e.
Confidence = Probability (Y if X) = Confidence = (Item X
+ Item Y) / (total number of transactions that have X).

- **Lift**

The product of the probabilities of the items on the left- and right-hand sides placing as if there was no relationship between item X and item Y divided by the likelihood of all items in a rule placing side by side (otherwise called support).
Lift (X, Y) = support (X, Y)/ (support(X)* support(Y)) = lift(Y, X)
Lift can take the following values:

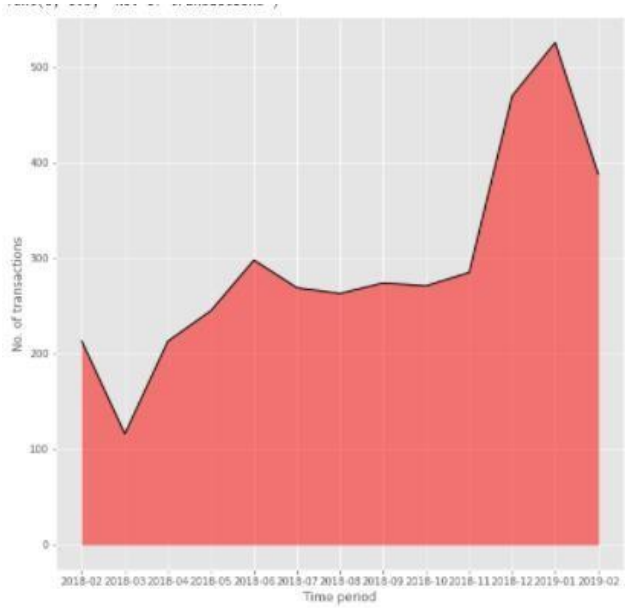
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6586 entries, 0 to 6585
Data columns (total 8 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   UserId                  6586 non-null   int64
 1   TransactionId           6586 non-null   int64
 2   TransactionTime         6586 non-null   object
 3   ItemCode                6586 non-null   int64
 4   ItemDescription         6571 non-null   object
 5   NumberOfItemsPurchased  6585 non-null   float64
 6   CostPerItem             6585 non-null   float64
 7   Country                 6585 non-null   object
dtypes: float64(2), int64(3), object(3)
memory usage: 411.8+ KB
```

1. Lift equal to 1; means that no relationship between X and Y (i.e., X and Y occur together only by chance)
2. A lift of more than one indicates that X and Y have a favourable association (i.e., X and Y occur together more often than random)
3. A lift of less than one indicates that X and Y have a negative relationship (i.e., X and Y occur together less often than random)

- **Leverage**

The difference between the frequency of A and B appearing together and the frequency expected if A and B were independent is measured by leverage.
Independence is shown by a leverage value of 0. Leverage (A, B) = support (A, B) − support(A)×support(B)

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume 5- Issue 2, Paper 47 August 2022



Conviction

A high belief value indicates that the outcome is heavily influenced by the antecedent. The denominator of a perfect confidence score, for example, becomes 0, and the conviction score is defined as 'inf'. If the objects are self-contained, the conviction is 1.

Conviction (A, B) = 1−support(B)/1 − confidence (A, B)

VIII. APRIORI ALGORITHM PROCESS

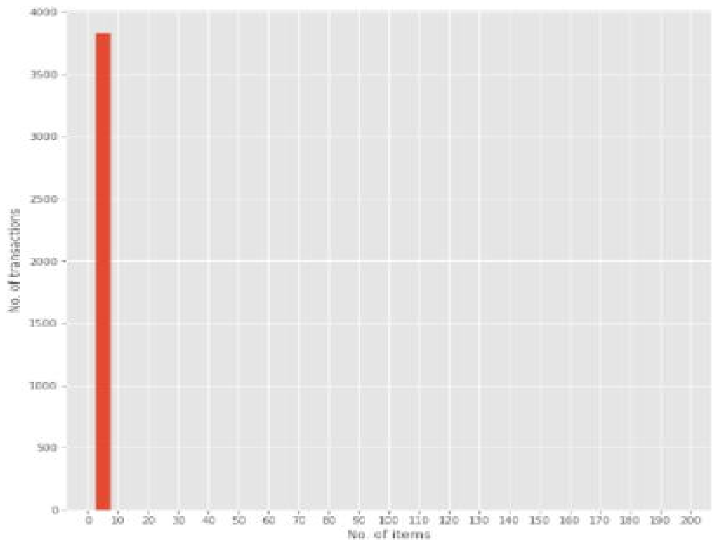
Step 1: Determine the amount of support for each itemin the dataset.
Step 2: Establishing a support threshold
Step 3: Identifying the most common things in the dataset Step 4: Identifying the frequent itemset's supportStep 5: Continue with larger sets.
Step 6: Create Association Rules and determineconfidence.
Step 7: Calculate the lift

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 5- Issue 2, Paper 47 August 2022

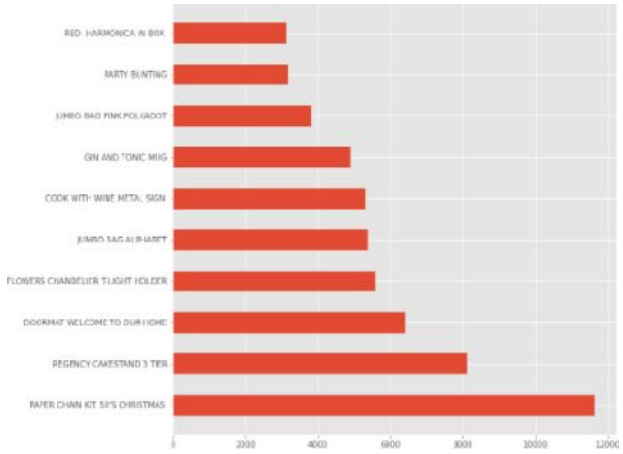
ItemDescription	4 PURPLE FLOCK DINNER CANDLES	50'S CHRISTMAS GIFT BAG LARGE	DOLLY GIRL BEAKER	NINE DRAWER OFFICE TIDY	OVAL WALL MIRROR DIAMANTE	RED SPOT GIFT BAG LARGE	SET 2 TEA TOWELS I LOVE LONDON	SPACEBOY BABY GIFT SET	TRELLIS COAT RACK	10 COLOUR SPACEBOY PEN	...	ZINC FOLKART SLEIGH BELLS	ZINC HEART FLOWER T-LIGHT HOLDER	ZINC HEART LATTICE T-LIGHT HOLDER	ZINC HERB GARDEN CONTAINER	ZINC METAL HEART DECORATION	ZINC SWEETHEART SOAP DISH	ZINC SWEETHEART WIRE LETTER RACK
TransactionId																		
580003	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	
580036	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	
580041	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	
580046	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	
580049	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
606834	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	
606522	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	
607938	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	
607991	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	
607935	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	

3831 rows × 1782 columns



Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 5- Issue 2, Paper 47 August 2022



count	3831.000000
mean	1.257374
std	0.631677
min	1.000000
25%	1.000000
50%	1.000000
75%	1.000000
max	11.000000

Name: ItemDescription, dtype: float64

IX. CONCLUSION

Market basket analysis produces a common itemset, i.e. association rules, which can be used to estimate client buying behaviour, and the marketer can use these ideas to showcase his supermarket's vast bazaar and expand the business fartherand profitably. The Apriori algorithm is the one utilised in market basket analysis. It has the potential to be a very strong algorithm for analysing customer purchase trends. Support, confidence, and lift are the three main metrics in market basket analysis. The frequency of an item in a particular transactional dataset will be evaluated by support, while the algorithm's analysing power or accuracy will be evaluated byconfidence. We looked into the transactional patterns of purchases, for example, and displayed the association patterns in a few of them. As a result, the temporal complexity of the Apriori algorithm is reduced when it is implemented using hash tables. The inadequacies of the Apriori algorithm for scanning transactional data for the purpose of creating association rules was the topic of this paper. By limiting the number of transactions when calculating the frequency of item-pairs or an item, a significant amount of time spent scanning consumer transactional data is saved. As a result, the new approach reduces the time complexity.

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 5- Issue 2, Paper 47 August 2022

REFERENCES

[1] I.H.W.E. Frank, Data Mining Practical Machine Learning Tools and Techniques, Morgan Kaufmann Publishers, 2005.
[2] A. Ceglar, and J.F. Roddick, "Association mining," ACM Computing Surveys (CSUR)2006
[3] J. Han, H. Cheng, D. Xin, and X. Yan, "Frequent pattern mining: current status and future directions," Data Mining and Knowledge Discovery2007, pp. 55-86.
[4] Data Mining and Business Analytics with RbyJohannes Ledolter. Published by John Wiley & Sons, year 2013.
[5] Jiawei Han, Micheline Kamber, Jain Pei, "Data Mining Concept and SSSTechnique 3rd Edition". Jugendra Dongre, GendLa