International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 3 Mar 2022, pp: 917-922 www.ijaem.net ISSN: 2395-5252

Market Basket Analysis Using Apriori Algorithm with Pruning Approach

Umamaheswararao Putrevu[1] Chandrasekhar Pj[2]
[1]*Business Analyst, Silverline Techno Solutions, Vijayawada, Andhra Pradesh.*
[2]*Senior Project Manager, Andhra Pradesh Centre for Financial Systems & Services (APCFS), Vijayawada,*

Submitted: 10-03-2022 Revised: 21-03-2022 Accepted: 25-03-2022

ABSTRACT: This research paper is about Market Basket Analysis, an important component of Business Analytics in retail companies to determine the sales for different segments of customers to improve customer satisfaction and to increase profitability of the company.

This is totally done by association rule mining in which it analyses the customer behaviour against the purchasing item from market. It analyses the customer purchasing pattern and generate frequent Itemset.

After generation of frequent Itemset it is easy to find most popular Itemset and least priority Itemset from large transactional database instead of reading it manually. Generation of frequent Itemset will enhance the market strategy, placement of goods in an organized manner and many more. Market Basket Analysis helps in increase in sales of goods and for profitable business.

In the present study Market Basket Analysis for a leading shopping mall is studied and analysed using frequent Itemset mining and decision tree techniques. The frequent Itemset are extracted from the market basket database using the efficient apriori algorithm and generated association rules to discover product associations and base for retailer's promotion strategy on them.

Market basket analysis is one possible way to find out which items can be put together in super markets. Pruning of association rules resulted in best outcomes.

KEYWORDS: Market Basket Analysis, Support, Confidence, Lift Ratio, Apriori Algorithm,

I. INTRODUCTION

One of the challenges for companies that have invested heavily in customer data collection is how to extract important information from their vast customer databases and product feature databases, in order to gain competitive advantage. Several aspects of market basket analysis have been studied in academic literature, such as using customer interest profile and interests on particular products for one-to-one marketing, purchasing patterns in a multi-store environment to improve the sales.

Market basket analysis has been intensively used in many companies as a means to discover product associations and base a retailer's promotion strategy on them. Informed decision can be made easily about product placement, pricing, promotion, profitability and also finds out, if there are any successful products that have no significant related elements. Similar products can be found so those can be placed near each other or it can be cross-sold. A retailer must know the needs of customers and adapt to them. Market basket analysis is one possible way to find out which items can be put together.

Market basket analysis gives the retailer good information about related sales on group of goods basis and also it is important that the retailer could know in which channel and in which region the products can be sold more and which session (i.e) morning or evening.

Market basket analysis is one of the data mining methods focusing on discovering purchasing patterns by extracting associations or co-occurrences from a store's transactional data. Market basket analysis determines the products which are bought together and to reorganize the supermarket layout and also to design promotional campaigns such that products' purchase can be improved.

Association rules are derived from the frequent Itemset using support and confidence as threshold levels. The sets of items which have minimum support are known as Frequent Itemset. The support count of an Itemset is defined as the proportion of transactions in the dataset which contain the Itemset. Confidence is defined as the measure of certainty or trustworthiness associated with each discovered pattern. Association rules derived depends on confidence.

The aim of association analysis is to find 'interesting' relationships between items (products,

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 3 Mar 2022, pp: 917-922 www.ijaem.net ISSN: 2395-5252

documents, etc.). Example: 'purchase relationship': milk, flour and eggs are frequently bought together. or If someone purchases milk and flour then the person often also purchases eggs.

The development of hardware, software and scientific advancements made the computerization of business easier. Scientific advancements have made easy to collect the data and digitizing the information which can be stored in databases. This makes collection and storing the data at a phenomenal rate.

A retailer must know the needs of customers and adapt to their needs. Market basket analysis has been intensively used in many companies as a means to discover product associations and base for retailer's promotion strategy on them.

Market basket analysis gives retailers good information about related sales based on group of goods. Consumer behaviour needs to be analysed and it can be done through different data mining techniques. Association rule mining finds interesting association or correlation relationships among a large set of data items. Association rules are derived from the frequent Itemset using support and confidence as threshold levels.

The most influential algorithm for efficient association rule discovery from market databases is apriori algorithm which is proposed by this investigation. This algorithm shows good performance with sparse datasets and hence it is considered.

Scope of Work

The apriori algorithm is selected for processing the input data and result produced as the list of rules that are strongly associated with each other. Applications of association mining technique to various fields are also studied.

- Integrating retailer, suppliers and customers for better customer service.
- Descriptions of customer relationship patterns.
- Constantly identifying the balance between marketing, sales and service inputs against changing customer needs to maximize profit Extracting or detecting hidden customer characteristics and behaviors from large databases.
- Development of online information kiosk for customers.

Project Goal

The main Goal of Market Basket Analysis is to get better efficiency of market and sales strategy using consumer transactional data collected during the sales transactions.

Market Basket Analysis includes the following objectives:

- To spot the frequent items on or after the transaction on the basis of support and confidence.
- To generate the association rules from the frequent Itemset.
- To generate reports using R programming.

Literature / Market survey

Several aspects of Market Basket Analysis have been studied in academic literature, such as using customer interest profile and interests on particular products for one-to-one marketing, purchasing patterns in a Supermarket to improve the sales. Decisions can be made easily about product placement, pricing, promotion, profitability and it is possible to find out if there are any successful products that have no significant related elements.

Figure1.Concept of Market Basket Analysis

A number of approaches have been proposed to implement data mining techniques to perform market analysis.

Vishal et al. [1] implemented data mining in online shopping system using Tanagra tool. They made decision about the placement of product, pricing and promotion.

Sudha and Chris et al.[2] proposed the impact of customers perception and CRM on Indian retailing in the changing business scenario using data mining techniques.

Mahesh Behera, Ankush Fartale Aniket Bhagat, Prof. Nidhi Sharma [3] have proposed in their paper on Market Basket Analysis based on frequent Itemset that Market basket analysis generates the frequent itemset i.e. association rules can easily tell the customer buying behaviour and the retailer with the help of these concepts can easily setup his retail shop and can develop the business in future.

Algorithms:

Many algorithms for generating association rules have been proposed. Some well-known algorithms are apriori, Eclat and FP-Growth, but they only do

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 3 Mar 2022, pp: 917-922 www.ijaem.net ISSN: 2395-5252

half the job, since they are algorithms for mining frequent Itemset. Another step needs to be done to generate rules from frequent Itemset found in a database.

To perform a Market Basket Analysis and identify potential rules, a data mining algorithm called the 'apriori algorithm' is commonly used, which works in two steps

- Systematically identify Itemset that occur frequently in the data set with a support greater than a pre-specified threshold.
- Calculate the confidence of all possible rules given the frequent Itemset and keep only those with a confidence greater than a pre-specified threshold.

The thresholds at which to set the support and confidence are user-specified and are likely to vary between transaction data sets.

Eclat (alt. ECLAT, stands for Equivalence Class Transformation) is a depth-first search algorithm using set intersection. It is a naturally elegant algorithm suitable for both sequential as well as parallel execution with locality enhancing properties. It was first introduced by Zaki, Parthasarathy, Li and Ogihara in a series of papers written in 1997[4].

FP stands for frequent pattern: In the first pass, the algorithm counts occurrence of items (attribute-value pairs) in the dataset, and stores them to 'header table'. In the second pass, it builds the FP-tree structure by inserting instances. Items in each instance have to be sorted by descending order of their frequency in the dataset, so that the tree can be processed quickly. Items in each instance that do not meet minimum coverage threshold are discarded. If many instances share most frequent items, FP-tree provides high compression close to tree root.

Recursive processing of this compressed version of main dataset grows large itemsets directly, instead of generating candidate items and testing them against the entire database. Growth starts from the bottom of the header table (having longest branches), by finding all instances matching given condition. New tree is created, with counts projected from the original tree corresponding to the set of instances that are conditional on the attribute, with each node getting sum of its children counts. Recursive growth ends when no individual items conditional on the attribute meet minimum support threshold, and processing continues on the remaining header items of the original FP-tree.

apriori uses a breadth-first search strategy to count the support of Itemset and uses a candidate generation function which exploits the downward closure property of support. The apriori algorithm was proposed by Agrawal and Srikant in 1994 [5] . Apriori is designed to operate on databases containing transactions (for example, collections of items bought by customers, or details of a website frequentation). Other algorithms are designed for finding association rules in data having no transactions (Winepi and Minepi), or having no timestamps (DNA sequencing).

Each transaction is seen as a set of items (an itemset). Given a threshold the Apriori algorithm identifies the item sets which are subsets of at least transactions in the database. Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time (a step known as candidate generation), and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found. Apriori, while historically significant, suffers from a number of inefficiencies or trade-offs, which have spawned other algorithms. Candidate generation generates large numbers of subsets (the algorithm attempts to load up the candidate set with as many as possible before each scan). Bottom-up subset exploration (essentially a breadth-first traversal of the subset lattice) finds any maximal subset S only.

Later algorithms such as Max Miner try to identify the maximal frequent item sets without enumerating their subsets, and perform "jumps" in the search space rather than a purely bottom-up approach

Apriori is an algorithm for frequent item set mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database: this has applications in domains such as market basket analysis.

- Most widely used approach for efficiently searching large databases for association rules
- The algorithm employs a simple 'a priori' belief to reduce the association rule search space – all subsets of a frequent item-set must also be frequent

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 3 Mar 2022, pp: 917-922 www.ijaem.net ISSN: 2395-5252

Block Diagram of Market Basket Analysis:

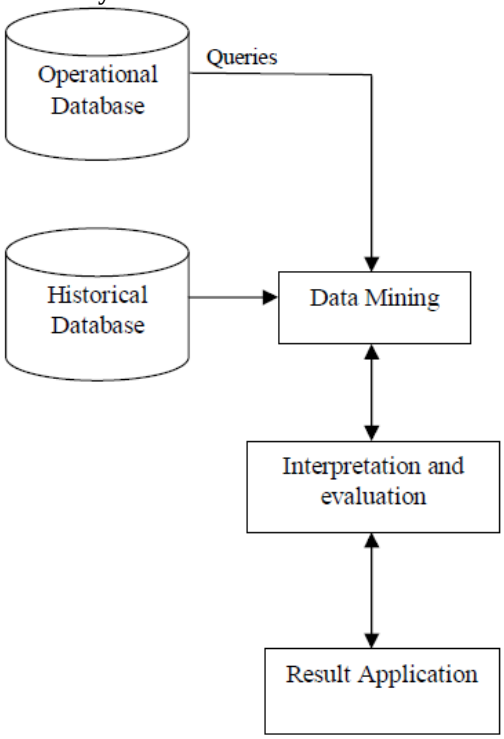


Figure 2. Block Diagram of Market Basket Analysis

II. METHODOLOGY

Business use of market basket analysis has significantly increased since the introduction of electronic point of sale. Amazon uses affinity analysis for cross-selling when it recommends products to people based on their purchase history and the purchase history of other people who bought the same item.

- Market Basket Analysis is a machine learning-based technique for identifying buying pattern from numerous retail transactions for helping the retailer in increasing the sales
- Market Basket Analysis involves –
  - Using simple performance. measures to find associations in large databases.
  - Understanding the peculiarities of transactional data
  - Knowing how to identify the useful and actionable patterns.
- The results of a market basket analysis are actionable patterns which can be used to suggest which set of items are frequently bought with which other set of items.
- Building block of market basket analysis is Association Rules.

Association rule learning:

Association rule learning is a rule-based machine learning method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness.

- Transactions are specified in terms of itemsets, such as follows:

{whole milk, other vegetables, rolls/buns , soda , yogurt, Others}

- Result of Market Basket Analysis is a collection of Association Rules that specify the patterns found in the relationships among items in the item-sets.
- Association rules are always composed from subsets of itemsets and are denoted by relating one itemset on the left hand side (LHS) of the rule to another item-set on the right hand side (RHS) of the rule.
- The LHS is the condition that needs to be met in order to trigger the rule, and the RHS is the expected result of meeting that condition.
- A rule identified from the example transaction is as follows:

{peanut butter, jelly} → {Bread}

- This association rule states that if peanut butter and jelly are purchased together, then bread is also likely to be purchased
- Association rules have been developed in the context of large retail databases.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 3 Mar 2022, pp: 917-922 www.ijaem.net ISSN: 2395-5252

- Association rules are used for unsupervised knowledge discovery in large databases.

Process: Association rules are usually required to satisfy a user-specified minimum support and a user-specified minimum confidence at the same time. Association rule generation is usually split up into two separate steps:

- A minimum support threshold is applied to find all frequent Itemset in a database.
- A minimum confidence constraint is applied to these frequent Itemset in order to form rules.

Finding all frequent Itemset in a database is difficult since it involves searching all possible item combinations.

Rule Form:

Antecedent→ Consequent [Support, Confidence]

Support and Confidence are user defined measures of interestingness.

$$Rule: X \Rightarrow Y \quad Support = \frac{frq(X,Y)}{N} \quad Confidence = \frac{frq(X,Y)}{frq(X)} \quad Lift = \frac{Support}{Supp(X) \times Supp(Y)}$$

Identifying frequently purchased groceries with Association Rules:

- Identifying all itemsets that meet a minimum support threshold
- Creating rules from these item-sets using those meeting a minimum confidence threshold

Steps for Performing Market Basket Analysis

1. Load the library and data
2. Inspect the transactions
3. Examine the frequency of the items
4. Plot the frequency of the items
5. Visualize the sparse matrix
6. Train a model – Set better support and confidence levels to learn more rules
7. Evaluate Model Performance
8. Write the rules in output file

III. RESULTS AND DISCUSSION

Output of intermediate steps

examine the frequency of items

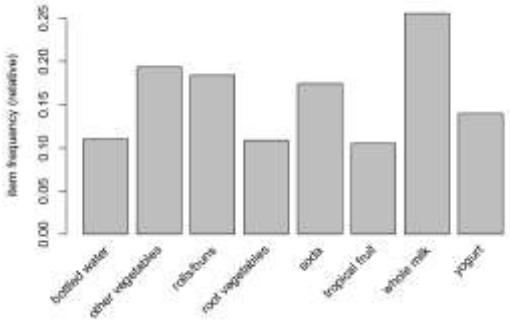
itemFrequency (groceries[1:3])

abrasive cleaner artif. sweetener baby cosmetics

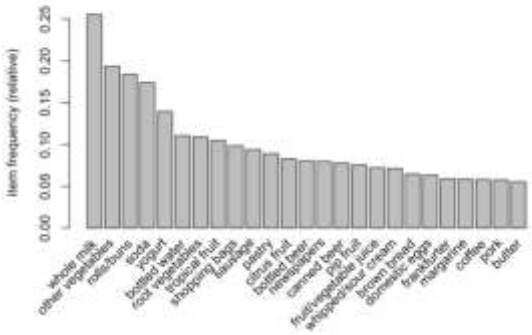
0.0035587189 0.0032536858 0.0006100661

plot the frequency of items

itemFrequency Plot (groceries, support = 0.1)



Fugure 3. ItemFrequency Plot (groceries, topN = 25)



R Programming output:

inspect(grocery.rules[1:10])

lhs rhs support confidence

1 {liquor} => {bottled beer} 0.004677173 0.4220183

2 {cereals} => {whole milk} 0.003660397 0.6428571

3 {candles} => {whole milk} 0.003050330 0.3409091

4 {soups} => {other vegetables} 0.003152008 0.4626866

5 {Instant food products} => {hamburger meat} 0.003050330 0.3797468

6 {Instant food products} => {whole milk} 0.003050330 0.3797468

7 {specialty cheese} => {other vegetables} 0.004270463 0.5000000

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 3 Mar 2022, pp: 917-922 www.ijaem.net ISSN: 2395-5252

8 {specialty cheese} => {whole milk} 0.003762074 0.4404762

9 {chocolate marshmallow} => {whole milk} 0.003152008 0.3483146
10 {flower (seeds)} => {other vegetables} 0.003762074 0.3627451

lift
1 5.240594
2 2.515917
3 1.334199
4 2.391236
5 11.421438
6 1.486196
7 2.584078
8 1.723869
9 1.363181
10 1.874723

Analysis of the results

The store manager should focus on whole milk product, vegetables, rolls / bons and yogurt to gain more profits.

IV. CONCLUSION

Summary of the project outcome:

- Market Basket Analysis is an important component of Business Analytics in retail companies to determine the sales for different segments of customers to improve customer satisfaction and to increase profitability of the company.
- This is totally done by association rule mining in which it analyses the customer behaviour against the purchasing item from market. It analyses the customer purchasing pattern and generate frequent Itemset.
- After generation of frequent Itemset it is easy to find most popular Itemset and least priority Itemset from large transactional database instead of reading it manually. Generation of frequent Itemset will enhance the market strategy, placement of goods in an organized manner and many more. Market Basket Analysis helps in increase in sales of goods and for profitable business.
- In the present study Market Basket Aanalysis for the groceries dataset is studied and analysed using frequent Itemset mining technique. The frequent Itemset are extracted from the market basket database using the efficient Apriori algorithm and generated association rules to discover product associations and base a retailer's promotion strategy on them.
- Market basket analysis is one possible way to find out which items can be put together in super markets.

Applications:

- Market basket analysis can be used in

Retail – each customer purchases different set of products, different quantities, and different times.

- Identify who customers are (not by name).
- Understand why they make certain purchases.
- Gain insight about its merchandise (products).
- Fast and slow movers.
- Products which are purchased together.
- Products which might benefit from promotion.
- Take action: Store layouts Which products to put on specials, promote, coupons etc.,
- Combining all of this with a customer loyalty card it becomes even more valuable.

Scope of future work:

For big datasets, we can implement FP Growth Algorithm with Spark for extracting good association rules.

REFERENCES

[1]. Vishal et al. implemented data mining in online shopping system using Tanagra tool. They made decision about the placement of product, pricing and promotion.
[2]. Sudha and Chris et al.proposed the impact of customers perception and CRM on Indian retailing in the changing business scenario using data mining techniques.
[3]. Mahesh Behera, Ankush Fartale Aniket Bhagat, Prof. Nidhi Sharma.
[4]. Zaki, Parthasarathy, Li and Ogihara, "Evaluation of Sampling for Data Mining of Association Rules" May 1996.
[5]. Agrawal, Rakesh and Srikant, Ramakrishnan. Fast algorithms for mining association rules in large databases. In VLDB, pp. 487–499, 1994