International Journal of Recent Engineering Research and Development (IJRERD)
ISSN: 2455-8761
www.ijrerd.com || Volume 02 – Issue 06 || June 2017 || PP. 08-13

An efficient Frequent Pattern algorithm for Market Basket Analysis

R.Beaulah Jeyavathana[1] K .S. Dharun Surath[2]
1: Assistant Professor (Senior Grade), Mepco Schlenk Engineering College, Sivakasi, India
2: UG Final Year Student, Mepco Schlenk Engineering College, Sivakasi, India

Abstract: Large retail business organizations maintain warehouses distributed across the world and inventory placement decisions are critical in maintaining faster delivery of products. In this work we present an algorithm to group products that are likely to be bought together so that they can be co-located in the same warehouse allowing multi-item orders to be shipped together. The effectiveness of the proposed algorithm has been validated using order details from a Belgium retail shop. Market basket analysis is an important component of analytical system in retail organizations to determine the placement of goods, designing sales promotions for different segments of customers to improve customer satisfaction and hence the profit of the supermarket. These issues are addressed here using frequent item set mining. The frequent item sets are mined from the market basket database using the efficient Frequent Pattern algorithm and then the clusters are generated.

Index Terms: Market Basket Analysis, FP-Tree algorithm, Frequent Item sets, Support, Confidence, Hypergraph, Bi-partitioning algorithm, Clustering, PPV.

I. INTRODUCTION

Market Basket Analysis (MBA) is a very prominent technique of data mining which is widely used in identifying products. Now days, there is a boom in online shopping of products due to rapid increase in e-commerce websites, and the ease in using them. This has made an enormous increase in transactional datasets. The availability of such datasets has promoted the researches to analyze the data and use it for productivity of the retailers. The increasing number of e-commerce websites has lead to development of competition between the retailers, thus this analysis of consumer purchase behavior has become a point of prime importance for them. This kind of analysis has helped them to gain the competitive advantage. To adapt to the needs of consumers retailers need to know the demands and expectations, which can be very well known by performing affinity analysis. This also helps us to know who the consumers are, understand why they make certain purchases, and gain insight about its merchandise. There are various algorithm and methods for these analysis the most commonly used is Frequent Pattern algorithm. This paper presents the process of the algorithm and the empirical results found out by them.

This algorithm identifies the frequent patterns present and gives the results based on them. This is an iterative process where the results are given in a step by step process. The frequent item sets are used for associating items together. The infrequent items are completely pruned out from the frequent sets, thus we have a very reliable technique to be sure of frequent itemset results.

A. Data Mining and Decision Support

The extraction of predictive information which is being hidden from large databases is a powerful tool with great potential to help organizations to define the information market needs of tomorrow. Data mining tools predict future trends and behaviors, allowing businesses to make knowledge-driven decisions that will affect the company, both short term and long term. The automated prospective analysis offered by data mining tools of today is much more effective than the analysis provided by tools in the past. Data mining answers business questions that traditionally were too time-consuming to resolve. Data mining tools search hidden patterns databases, finding predictive information those experts may miss because it was outside their expectations. Data mining is not new. The technology in data mining is the Decision Support Systems (DSS) which is being used for their great potential to supply executives with large amount of data needed to carry out their jobs.

After1995s, corporate intranets were developed to support information exchange and knowledge management. The primary decision support tools in use included ad hoc query and reporting tools, online analytical processing,and optimization and simulation models and data visualization. On the other hand, a data warehouse is the newest form of decision support system. Data mining is being defined as one of the hottest technologies in decision support applications till today. The use of bar codes for most commercial products, and

the computerization of many business transactions have flooded us with information, and generated an urgent need for new techniques and tools that can intelligently and automatically assist us in transforming this data into useful knowledge. Today, there is a huge amount of information locked up in the mountains of data in companies' databases, information that is potentially important but has not yet discovered. . In this context, knowledge about how customers are using the retail store is of critical importance and distinctive competencies will be built by those retailers who best succeed in extracting actionable knowledge from these data. Data Mining provides many different techniques to extract knowledge from data. It is an exciting multidisciplinary field of research which has many extremely useful applications. At present the techniques are becoming more commonly used but have not been applied adequately in the store layout. To find relationships between items purchased by customers store layout problem is motivated by applications known as market basket analysis.

II. RELATED WORK

Market basket analysis is a technique that helps us in determining which products tends to be purchased together in accordance with the association rules[1]. The primary objective is to improve the effectualness of sales and marketing strategies with the help of previously obtained customer data. Association rules aims to identify those items which frequently occur in a database. This paper presents each item is represented by Boolean value, i.e., 0 and 1, where 0 represents that item is not present whereas 1 represents that item is present. We have proposed a novel data structure, FP-Tree, frequent pattern mining, overcomes the main drawbacks of Apriori algorithm. The frequent itemsets are generated only with two scans of the database. It is an extended prefix tree structure which is used for the storage of information about patterns[3]. The nodes of the tree are arranged in such a way that the nodes occurring more frequently will have better chances of sharing nodes than nodes occurring less frequently. FP-Tree performs better than Apriori because there is no candidate set generation as well as the length of the frequent itemsets increases as support value decreases. FP-Growth algorithm is more efficient than latter one.There exists many other algorithms for mining of frequent itemsets viz., Apriori and Éclat; FP-Tree growth algorithm preprocesses the database only twice as follows: an initial scan of the database determines the frequencies of the items[2]. All the uncommon items -- the items that do not appear in a minimum number of user-specified transactions -- are discarded from it as they cannot be a part of frequent itemsets .In addition to this, all the items in the transaction are sorted in descending order in reference to their frequencies[4]. Whilst the algorithm does not depend upon the specific order of the frequencies of items, sorting in descending order may lead to much less execution time than ordered randomly. Sorting in ascending order leads to slower operations implementing even worse than in random order.

III. PROPOSED SYSTEM

The proposed design has 4 phases as shown in the figure 1. The frequent items that are bought together are identified using Frequent Item Set Mining (FISM) algorithms such as FP-Growth algorithm. This relationship is then mapped to a hyper graph. The actual product clusters are then identified by employing hyper graph clustering algorithms.

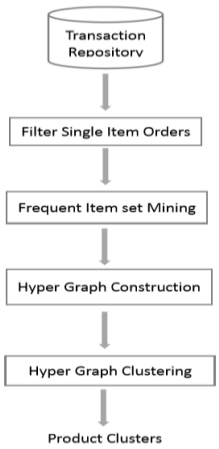

Figure 1: Steps involved in Market Basket Analysis.

A. TRANSACTION REPOSITORY

The transaction repository is the database which contains the details about the orders placed by the customers, product id and other transactions details.From these details the product id is used for the generating of the product clusters.

B. FILTER SINGLE ITEM ORDER

The Orders containing only one item does not convey any useful information about the relationship among products. It has been noted that a relatively large portion of orders is single item order. The single item orders are therefore removed to reduce the overhead of processing in the subsequent stages of the algorithm. A Frequent Pattern (FP) Tree is constructed to store the frequent pattern of the input item sets.

Algorithm 1: FP-Tree Construction

Input:A minimum support threshold and a transaction database DB.
Output: FP-tree, the frequent-pattern tree of DB.
Method: The FP-tree is constructed as follows:

1. Scan the transaction database DB once. Collect F, the set of frequent items, and the support of each frequent item. Sort F in support-descending order as FList, the list of frequent items.
2. Create the root of an FP-tree, T, and label it as "null". For each transaction Trans in DB do the following:

● Select the frequent items in Trans and sort them according to the order of FList. Let the sorted frequent-item list in Trans be [q| Q], where q is the first element and Q is the remaining list. Call insert tree ([q | Q], T).

● The function insert tree([q | Q], T) is performed as follows: If T has a child N such that N.item-name = q.item-name then, increment N's count by 1; else create a new node N, with its count initialized to 1, its parent link linked to T, and its node-link linked to the nodes with the same item-name via the node-link structure. If P is nonempty, call insert tree(Q, N) recursively.

C. FREQUENT ITEM SET MINING

The most popular algorithm for frequent itemset mining is FP-Growth algorithm. This algorithm is based on prefix tree representation of the database which saves considerable amount of memory for storing the same. The algorithm is described as a recursive elimination scheme. The efficiency of FP-Tree algorithm accounts for three reasons: Firstly, it is a compressed representation of the transaction database. Secondly, this algorithm scans the database only twice. Thirdly, it uses divide-and-conquer approach which considerably reduces the size of the conditional FP-Tree. After all the infrequent items have been deleted, the resultant is a FP-Tree. In this, each node corresponds to only one item and a set of transactions that share the same prefix are represented by one path.

Algorithm 2: FP-Growth

Input: A minimum support threshold and a database DB, represented by FP-tree constructed according to Algorithm 1, and.

```
Output: The complete set of frequent patterns.
Method: call FP-growth(FP-tree, null).
Procedure FP-growth(Tree, b) {
if Tree contains a single prefix path then // Mining single prefix-path FP-tree {
   let R be the single prefix-path part of Tree;
   let S be the multipath part with the top branching node replaced by a null root;
   for each combination (denoted as ß) of the nodes in the path R do
       generate pattern ß ∪b with support = minimum support of nodes in ß;
       Let freq pattern set(R) be the set of patterns so generated; }
else let S be Tree;
    for each item bi in Q do {// Mining multipath FP-tree
       generate pattern ß = bi∪b with support = bi .support;
  constructß's conditional pattern-base and then ß's conditional FP-tree Tree ß;
if Tree ß ≠ Ø then
    call FP-growth (Tree ß , ß);
    Let freq pattern set(S) be the set of patterns so generated; }
return (freq pattern set(R) ∪freq pattern set(S) ∪ (freq pattern set(R) × freq pattern set(S)))
}
```

D. HYPER GRAPH CONSTRUCTION

A hyper graph is a generalization of a graph in which an edge can join any number of vertices. The frequent transactions which are obtained by frequent item set mining is represented in the form of hyper graph

where each edge can join more than one frequent item sets.The relationships identified by Frequent Item Set Mining are mapped to a weighted hyper graph. The properties of the constructed hyper graph are as follows.

1. Every item represents a node in the hyper graph.
2. For every frequent item set, an edge is created that connects all nodes representing the items present in the frequent item set and the edge weight is (number of transactions containing all items in the FIS / number of transactions containing at least one item in the FIS)

E. HYPER GRAPH CLUSTERING

Once the relationships are mapped into hyper graphs, employing hyper graph clustering algorithms can identify groups of products that are likely to be purchased together. Hyper graph clustering algorithms tend to group frequent item sets together producing larger meaningful clusters. Spectral clustering techniques reduce dimensionality by mapping data points using Eigen vectors and then applying Perron-Frobenius clustering algorithms to find clusters.

The hyper graph clustering is implemented using the Perron-Frobenius clustering algorithm in three steps namely

- Preprocessing
- Spectral Representation
- Clustering

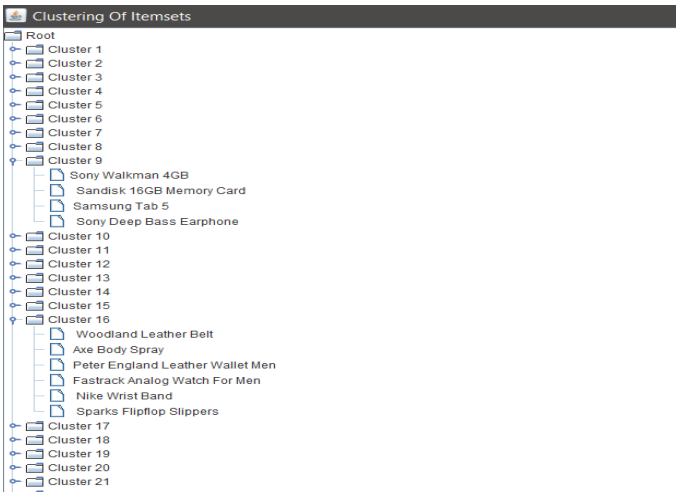

Figure 2: Representation of product clusters using jtree

1. Preprocessing:

Construct the graph and the similarity matrix representing the dataset.

2. Spectral Representation:

The laplacian matrix L is calculated by the formula

$$L = D - A$$

where D is the degree matrix and A is the adjacency or vertex matrix of the graph.

Compute eigenvalues and eigenvectors of the Laplacian matrix.

Map each point to a lower-dimensional representation based on one or more eigenvectors.

3. Clustering :

- Recursively apply bi-partitioning algorithm in a hierarchical divisive manner.
- The basic idea is to apply bi-partitioning algorithm recursively in a hierarchical way: after partitioning the graph into two, reapply the same procedure to the sub-graphs.
- The number of groups is supposed to be given or directly controlled by the threshold allowed to the objective function.

F. POSITIVE PRETECTIVE VALUES

Positive Predictive Value can be defined as follows:

PPV = number of true positives/(number of true positives + number of false positives)

The value of PPV when the number of items is increased with fixed order count was studied. The result is shown in fig 3.3. The numbers of orders were fixed at 80,000. It is clear from the chart that as number of items increases the PPV decreases.

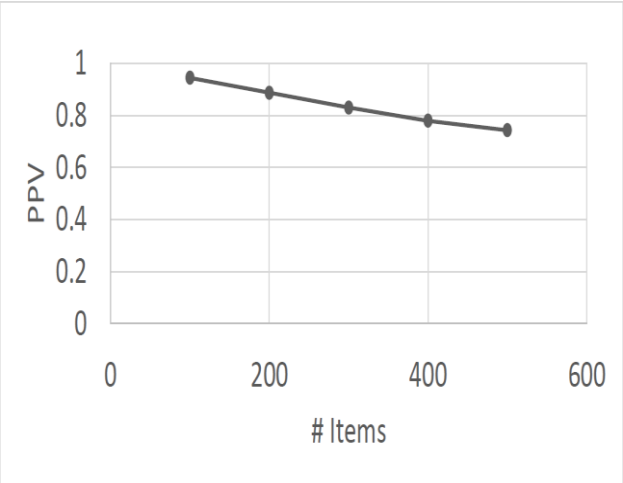

Figure 3: PPV vs. Items

Keeping the number of distinct items as 500 the effect of number of transaction is studied (Fig 3.4). When the number of transactions increase PPV increases. This is because more data is available for learning leading to better prediction.

Using this data the number of transaction required to produce a particular PPV can be found given the number of distinct items.

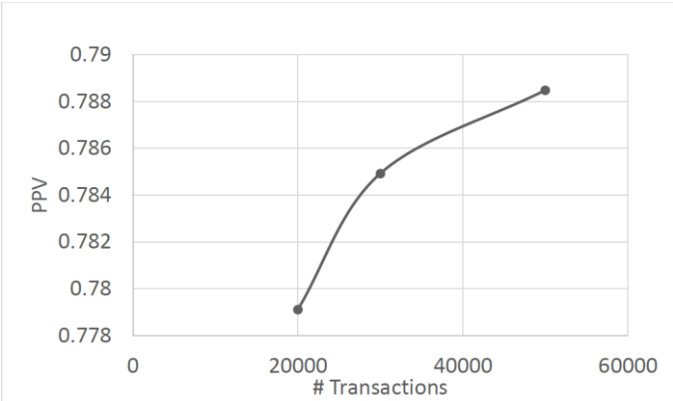

Figure 4: PPV vs Transactions

IV. CONCLUSION

The proposed "Market Basket Analysis" can be effectively used by the online retailers such as Amazon, Flipkart to maintain their products in the warehouses. These large retail business organizations can maintain warehouses distributed across the world and inventory placement decisions which are critical in maintaining faster delivery of products.

ACKNOWLEDGMENT

First and foremost, we thank the Almighty for his abundant blessings that is showered upon our past, present, future successful endeavors.

At the appellation, we would like to extend our sincere gratitude to our reverent principal, Dr. S. Arivazhagan, M.E., Ph.D., for providing sufficient working environment such as system and library facilities.

We also thank him very much for providing us with adequate lab facilities, which enable us to complete our work.

We are grateful to our charismatic Head of the Computer Science Engineering and Department Dr. K.

Muneeswaran, M.E., Ph.D., for giving us this golden opportunity to undertake a work of this nature and for his most valuable guidance given at every phase of our work.

We acknowledge with the sense of gratitude, the help given to us by all the technicians in the department.

We are very grateful to our beloved parents and friends who afforded the necessary help for us at the right time for making our work a grand success.

REFERENCES

[1] Market Basket Analysis using Association Rule Learning, "NidhiMaheshwari, Nikhilendra K. Pandey, Pankaj Agarwal " International Journal of Computer Applications Volume: 03 Issue: 02 | Feb-2016 .

[2] Affinity Analysis and Association Rule Mining using Apriori Algorithm in Market Basket Analysis ,"TejonidhiAphale, Rahul Chaudhari, JinitBansod", "R. Karthiyayini, Dr. R. Balasubramanian",International Journal of Advanced Research in Computer Science and Software Engineering , Volume 6, Issue 10, October 2016.

[3] Data Mining Based Store Layout Architecture for Supermarket , "Aishwarya Madan Mirajkar, AishwaryaPrafullaSankpal, PriyankaShashikantKoli , Rupali AnandraoPatil , AjitRatnakarPradnyavant",International Research Journal of Engineering and Technology (IRJET) Volume: 03 Issue: 02 | Feb-2016 .

[4] Information System on Market Basket Analysis, "ShrutiKawale, RajniTetwar, NirmalaSuryavanshi, PrachiWankhede, Prof. NehaTitarmare", International Journal of Computer Science and Mobile Computing, Vol.5 Issue.3, March- 2016.