A Survey Paper on Market Basket Scaling Analysis Using Map-Reduction Technique
Ankita Shah\(^1\) Mohit Patel\(^2\) Sachin Jain\(^3\)
\(^1\)PG Research Scholar \(^2\)\(^3\)Asstant Professor
\(^1\)\(^2\)\(^3\)Department of Computer Science & Engineering
\(^1\)\(^2\)\(^3\)Parul University, Gujarat, India

Abstract— Market Basket Analysis is a technique to identify the items that are purchased together. A predictive market basket analysis is used to identify sets of products/services purchased or events that occur generally in sequence. The fundamental approach is to find the associated pairs of items in a store when there are transaction data sets. Consequently, our proposed framework performing “Market Basket Analysis” will help the retailers to settle on better choices all through the whole organization which will help in expanding the benefits and effectiveness of organization. Map/Reduce to upgrade and change over the current successive calculations to parallel programming so that the propose Market Basket Analysis calculation with Map Reduce will give better Future Prediction. Two calculations are proposed: One is to adjust current Apriori-calculation and second is to build a straightforward calculation that sort’s information sets and changes over it to (key, value) sets to fit with Map Reduce. The framework will be implemented on "MATLAB" structure utilizing Scale with clickstream datasets.

Key words: Data Mining, Market Basket Analysis, Association Rule, Apriori Algorithm, Map/Reduce, SVM

I. INTRODUCTION

Market basket analysis is a data mining technique which is widely used to determine what product customers purchase together. It is a very popular technique specially used to find association between two products. Market basket analysis is also known as association-rule mining. It is a useful method of discovering customer purchasing patterns by extracting associations or co-occurrences from stores transactional databases. The information which are obtained from the analysis that is used in forming marketing, sales, service.

The existing methods fail to discover the purchasing patterns in a multi-store environment. The proposed method is efficient and that has advantage over the traditional method when stores are diverse in size, product mix changes rapidly over time, and larger numbers of stores and periods are considered.

Fig. 1: Market Basket Analysis
If you buy a certain group of items than you are more or less likely to buy another group of items. For example, if you are in an English pub and you buy a beer and don't buy a bar meal than you are more likely to purchase crisps US. Chips at the same time than somebody who didn't buy beer. The set of items a customer buys is known as an item set, and market basket analysis find relationships between purchases.

The relationship will be in the form of a rule: IF \{beer, no bar meal\} THEN \{crisps\}. The probability that a customer will purchase beer without a bar meal is that the antecedent is true and known as the support. The conditional probability that a customer will purchase crisps is known as the confidence. The algorithms for performing market basket analysis are straightforward. The main aim is to discover knowledge patterns hidden in large data set that can produce more understanding to the data holders and identify new opportunities for imperative tasks including strategic planning and decision making. The main purpose is to find connections among the items for cross-offering their items to clients. An association rules has received a great deal of interest in the field of market basket analysis.

The paper is organized as follows: Section II have the related work of association rules and Apriori. Section III represents the methods used in this work is to achieve the desired results which are Apriori, SVM and Map Reduce. Section IV introduces the Comparative Analysis. Section V presents an Expected Solution. Section VI gives the Conclusion and the Future Scope about the methods we applied.

II. RELATED WORK

Market basket analysis [1] is one of the data meaning techniques used for inventing the system of the relation between one items to another. Association rule mining is a task and a key issue in knowledge discovery and data mining. For example association rule mining is widely used in retail industry to discover interesting association rules for better decision making.

The Apriori algorithm [2] was the first algorithm for mining association rule that uses the support based pruning to control the exponential growth of candidate item sets. Association rule mining can be combined with SVM to take advantage of knowledge represented by associated rules and use the power of SVM algorithm to create an efficient and accurate classifier model.

Association rule mining [3] is the study of identify associated items from purchasing data of the customer. The main aim is to discover the concurrence associations among data in large databases and to discover interesting associations between attributes contained in a database. An
association is if-then-rule which is supported by data. The motivation given for the development of association rules is market basket analysis which deals with the point-of-sale transaction of large retailers.

There is not any other algorithm with Market Basket Analysis [5] with Map/Reduce. The main aim is to propose the algorithm and to convert data to (key, value) pair and execute the code on Map/Reduce platform. It is composed of two functions to specify, “Map” and “Reduce”. They are both defined to process data structured in (key, value) pairs.

Frequent itemset mining [7] is a data mining topic and its task is to find all subsets of items which frequently occur. Apriori algorithm is widely used algorithm for frequent itemsets mining. The algorithm uses an iterative method to generate the candidate (k + 1) itemsets from the frequent k-itemsets. The sequential Apriori algorithm performance is inefficient, especially when the data sets include terra- or peta-bytes of data.

III. METHODS

There are various methods for Market basket Analysis:

A. Association rule mining

Association rule mining [1] is an important task and a key issue in knowledge discovery and data mining. For example, association rule mining is widely used in retail industry to discover interesting pattern to help with better decision making.

1) Support

Combined percentage of the two items; for identifying the combination of the item which fulfill the minimum requirement of support value. Support value of an item is achieved by using the following formula:

\[ S(A) = \frac{\text{Amount of transaction } A}{\text{Total transaction}} \]

The formula of support value of two items

\[ S(A \cap B) = \frac{\text{Amount of transaction } A \cap B}{\text{Total Transaction}} \]

2) Confidence

The frequencies of the item Y appear in the transaction which contains X. After all of system of high frequency found, then rules need to be found.

Confidence = \( P(Y | X) = \frac{\text{Amount A} \cap \text{B}}{\text{Amount A}} \)

B. Apriori Algorithm

The Apriori algorithm [1] was the first algorithm proposed for mining association rule that uses the support based pruning to control the exponential growth of candidate item sets. The main aim of association rule is to extract the training dataset according to the defined threshold values of the parameters Support and Confidence. Apriori algorithm is first applied on the data in order to generate association rules. This will be done through several executions each with different values of the parameter Support. Once the association rules have been generated, the rules with the highest confidence are selected. Each execution will generate different association rules which will subsequently clustered into categories based on the item appears in the right hand side of the association rule. The Apriori Pseudo code algorithm is presented as following:

Algorithm I Apriori

Input:

\[ DB: \text{transaction database;} \]
\[ sup: \text{the minimum support threshold} \]
Output: frequent itemsets

Description:

1: \( L1= \text{find frequent 1-itemsets}(DB) \)
2: for \((k=2; Lk-1 = \phi; k++)\) 
3: \( Ck= \text{Apriori gen}(Lk-1) \)
4: for each transaction \( t \in DB \) 
5: \( Ct= \text{subset}(Ck,t) \)
6: for each candidate \( c \in Ct \)
7: \( c.\text{count}++ \)
8: } 
9: \( Lk = \{c \in Ck|c.\text{count} \geq sup\} \)
10: } 
11: return \( L = \_k Lk \)
12: \( \text{Procedure Apriori gen}(Lk-1): \text{frequency}(k-1)-\text{itemsets} \)

C. Map Reduce Algorithm

Map Reduce is a programming technique [5] for analyzing data sets that do not fit in memory. Map reduce uses a data store to process data in small chunks that fit into memory. Each chunk goes into a Map phase and formats the data to be processed. Then the intermediate data chunks go through a Reduce phase and aggregates the intermediate results to produce a final result. The Map and Reduce phases are encoded by map and reduce functions, which are primary inputs to mapreduce. There are combinations of map and reduce functions to process data so the technique is both flexible and extremely powerful for handling large data processing tasks. The utility of the map reduce function lies in its ability to perform calculations on large collections of data. Thus, map reduce is not well suited for performing calculations on normal sized data sets that can be loaded directly into computer memory and analyzed with traditional techniques. Better to use map reduce to perform a statistical or analytical calculation on a data set that does not fit in memory. Each call to the map or reduce function by map reduce is independent of all others. For example, a call to the map function cannot depend on inputs or results from a previous call to the map function. It is better to break up the calculations into multiple calls to map reduce. The algorithm has the following steps:

1) Map reduce reads a chunk of data from the input data store using [data,info] = read(ds), and then calls the map function to work on that chunk.
2) The map function receives the chunk of data and organizes it or performs a calculation and then uses the add multi functions to add key-value pairs to an intermediate data storage object called a KeyValueStore. The number of calls to the map function by map reduce is equal to the number of chunks in the input data store.
3) After the map function works on all of the chunks of data in the data store, map reduce groups all of the values in the intermediate Key Value Store object by unique key.
4) Next, map reduce calls the reduce function once for each unique key added by the map function. Each unique key can have many associated values. Map reduce passes the values to the reduce function as a value Iterator object that is an object used to iterate
over the values. The Value Iterator object for each unique key contains all the associated values for that key.

5) The reduce function uses he has next and get next functions to iterate through the values in the Value Iterator object one at a time. Later aggregating the intermediate results from the map function, the reduce function adds final key-value pairs to the output using the add and add multi functions. The order of the keys in the output is the same as the order in which the reduce function adds them to the final KeyValueStore object. So, mapreduce does not explicitly sort the output.

D. Support Vector Machine

Support vector machine algorithm [2] is used to evaluate the classification accuracy for a new transaction not exists in the data used in Apriori algorithm. SVM is used to retrieve the most similar association rules from the knowledge database that best match a new query transaction. Support vector machine is a classifier using a decision boundary to separate two classes defined by solving a quadratic optimization problem. SVM finds an optimal solution that maximizes the distance between the hyperplane and the most critical training samples. The decision boundary is then specified by a subset of critical training samples named support vectors that lie on the edge. SVM extends to multi-class classification using several methods. SVM has been used in many applications because its design is well suited complex large datasets. SVM is one of the best performers for a number of classification tasks ranging from text to microarray data. Suppose we have two features, x1 and x2, and want to classify all the elements appeared. We can see the class red and the class black. The goal of the SVM is to design a hyperplane that classifies all training vectors in two categories. The black line is defined as the hyperplane that classifies all the training vectors in the two classes and have multiple hyperplanes that can classify all the instances correctly in this feature set. The best choice will be the hyperplane that leaves the maximum margin from both classes.

![Fig. 1: SVM applied onto two features, x1 and x2](image)

IV. COMPARATIVE ANALYSIS

<table>
<thead>
<tr>
<th>Method</th>
<th>Advantages</th>
<th>Limited</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scaling Method</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Association[1]</td>
<td>Purchasing pattern in multiple ways</td>
<td>Obtaining huge no of rules</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Obtaining non</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Advantages</th>
<th>Limited</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Comparative Analysis

V. EXPECTED SOLUTION

Map Reduce is a parallel programming model that is emerged in the mining of datasets of terabyte scale or larger on clusters of computers. A time efficient Map Reduce framework is to mine frequent item sets from large scale datasets. This algorithm gets better performance in term of execution time.

VI. CONCLUSION AND FUTURE SCOPE

Market basket Analysis using association rule works in determining the customer buying pattern. Further will work on different algorithm and make proposed algorithm to find relation among item set for Market Basket Analysis using parallel Processing by Map-Reduction will produce time cost for finding next generation customer purchase pattern.

REFERENCES