Frequent Pattern Discovery Using Apriori Algorithm for Data Mining

Shikha Nandi
B. Tech (CSE), Galgotias University, Greater Noida, Uttar Pradesh, India

***Abstract*- We live in a fast-changing digital world. In today's age we expect the sellers to tell us what we might want to purchase. Most of us rely on popular website's like Amazon's recommendation system to but stuff. This gives the seller an interesting opportunity to increase their sales. If a seller can tell us what we might be interested in to buy, it doesn't only improve their sales, but also the customer experience and life time value. On the other hand, seller is unable to predict the next purchase or our shopping behavior, the customer or we might not go back to their store or website. In this paper, we will be implementing one such popular algorithm called Apriori algorithm with NoSQL Database that enables us to predict the shopping behavior of customers to know the items that are bought together frequently.**

Index terms- Data Mining, Association Rule Mining, No SQL Database, Apriori Algorithm, Frequent Item Set, Customer Segmentation

I. INTRODUCTION

Now a days a popular analogy states that data is 'the new oil'. You can think of data mining as drilling down large data for refining oil. Data mining is the technique by which we extract value from the data available. Data mining means analyzing the data in order to search for patterns in it, trends and correlations, and anomalies that can be critically important for a particular organization and its business.

For instance, data mining enables you to identify the best customers for the business. Today organizations leverage the benefits of data mining using various techniques to analyze a particular customer's purchase history and helps predict or forecast what a customer might be interested in to purchase in the future. It can also highlight a particular set of purchases that are not ordinary or do not have a recognizable pattern for a customer and hence indicates fraud.

A. Data Mining-

It can be defined as the technique of exploring and analyzing large amounts of data typically from data marts or data warehouses even from real time transactional databases with a specific goal of discovering insights and is significantly important to know about the underlying patterns and rules.

Data Mining like wisely famously is known as Knowledge Discovery in Databases (KDD), alludes to non-trivial extraction of verifiable and helpful data from information in databases.

In order to make a decision, an organization needs to have enough knowledge about the data and its business. Data analysis is done using Knowledge Discovery in data mining techniques Association Rule Mining, Apriori Algorithm etc.

B. Apriori Algorithm-

It is a widely used algorithm in data mining for mining frequent item sets and association rule mining. For example, understanding a customer's buying pattern. By finding correlations and associations between different set of items that customers buy.

The goal of Apriori algorithm is to generate association rule and to find frequent item sets with the help of candidate generation. The 2 major players of Apriori algorithm are Support and Confidence.

Say, John goes to buy a pack of milk from the supermarket. He also grabs a couple of chocolates as well. The manager there finds out that, not only John, people often tend to buy milk and chocolates together. After analyzing the pattern, the store manager starts to arrange or place these items together and notices an increase in sales.

C. Association Rule-

Association rule mining leverages use of machine learning models to analyze large data for any patterns, co-occurrence, in a database. It identifies patters using if then statements, which are called association rules.

An association rule comprises of two parts: an antecedent or the 'if' statement and a consequent or the 'then' statement. An antecedent represents an item within the dataset. A consequent represents an item found in combination along with the previous item or the antecedent.

Association rules are found by searching dataset for any frequent if then patterns and using the filters support and confidence to discover the important relationships between items. Support represents how frequently an item shows in the dataset.

On the other hand, Confidence tells us the number of times the if then statements are found to be true. And the third and last metric is called lift, that is generally used to compare confidence with expected confidence.

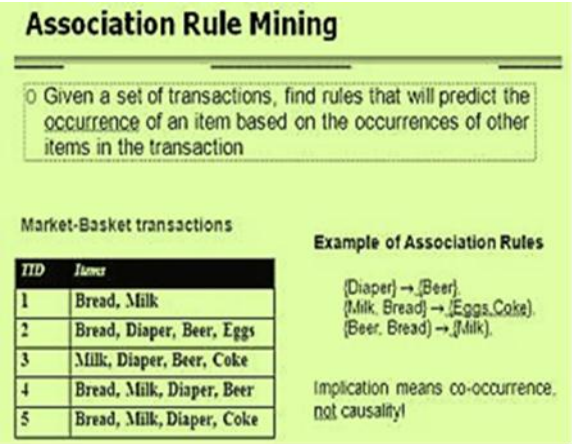

Fig.1.1 Association Rule Mining

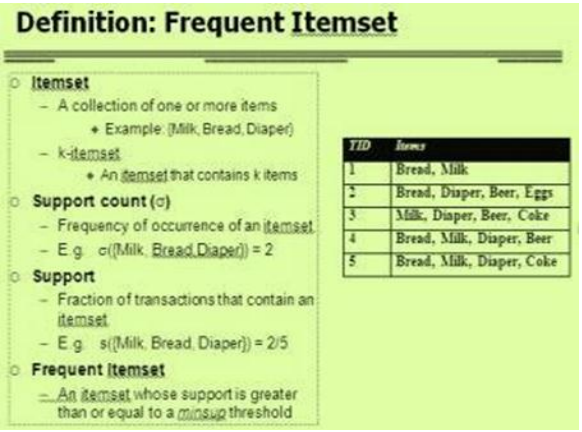

Fig.1.2 Association Rule Mining Continued

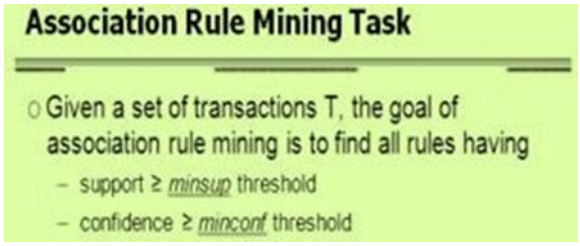

Fig.1.3 Association Rule Task

Fig. 1.4 Association Rule Frequent Item Sets

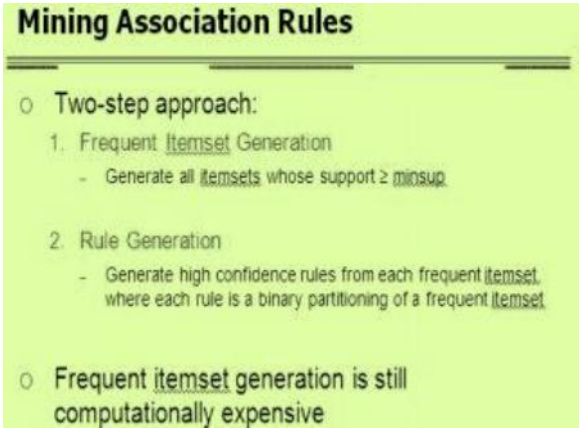

Fig.1.5 Association Rule Approaches

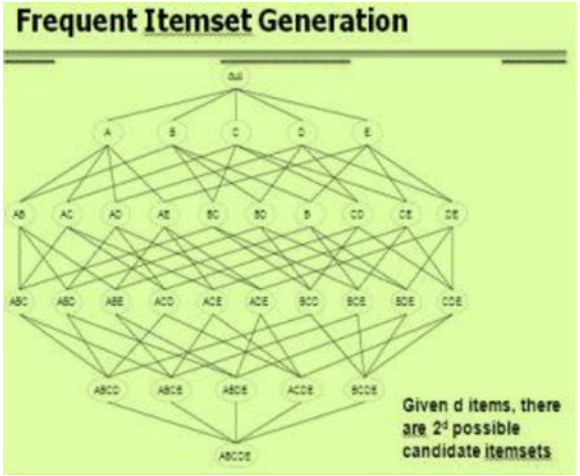

Fig .1.6 Association Rule Possible Itemset

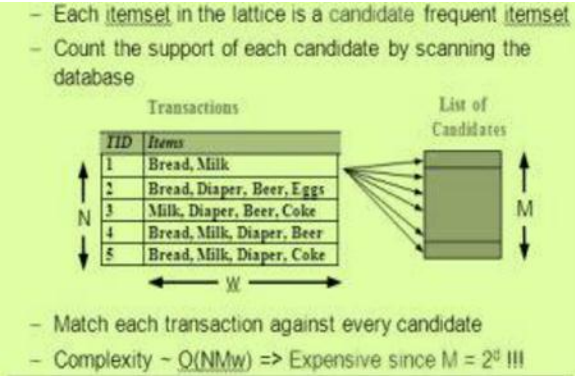

Fig.1.7 Association Rule Itemset Mapping

II. PROPOSED SYSTEM

Our project comprises of the following models-

A. Data Input.
B. Market Basket Analysis
C. Applying Association Rule.
D. Data Modelling

A. Data Input-

We're using NoSQL database as an input to our system. It has attributes like- Product Name, Quantity, Price. We're going to leverage the benefits of converting an SQL database into a NoSQL database for the implementation of an Apriori algorithm to find out the most common baskets within our data set.

We know that the use of NoSQL databases is faster than traditional RDBMS. Apriori algorithm along with NO SQL databases improves performance significantly.

In general, NoSQL has advantages over using relational databases like, it can handle very large volumes of structured/semi-structured/unstructured data, quick iteration, and frequent code pushing, object-oriented programming that is easy to use and flexible, it also has scale out architecture.

B. Market Basket Analysis-

Market Basket Analysis helps us determine which set of products sell together, we give the list of transactions data as input to the analysis system. The data has two dimensions, that represents customer and products, this analysis helps us know which products sell together for a customer. We chose this algorithm mainly because of its speed of execution and efficiency.

C. Applying Association Rule (Using Apriori Algorithm for Frequent Itemset Discovery)-

Using the Apriori algorithm, the number of itemsets that have to be analyzed can be trimmed, and the list of frequent itemsets can be discovered using the below steps-

Step I. Start with an itemset containing just a single item, such as Milk and Butter.

Step II. Find out the support of itemsets. Keep only the itemsets that match the least support threshold and remove others.

Step III. Using the remaining itemsets from Step 1, generate all the possible itemset configurations.

Step IV. Repeat the Step 1 & Step 2 until there are no newer itemsets are remaining.

We just saw how the Apriori algorithm works to identify itemsets with high support. The same technique can be used to identify item associations with high confidence or lift. Finding rules with high confidence or lift is less computationally demanding once itemsets with high support confidence have been found, because confidence and lift values are calculated with the help of support values.

Below example explains the task of calculating high confidence rules. If the rule

{milk, butter -> mango}

has low confidence, all other rules with the same items and with mango on the right side will also have low confidence. Specifically, the rules

{milk -> mango, butter}

{butter -> mango, milk}

would have low confidence as well. As mentioned earlier the lower level candidate item rules can be trimmed using the Apriori algorithm, so that very fewer candidate rules need to be analyzed/examined later.

D. Data Modelling –

Data modeling can be defined as the representation of data structures in the format of tables for an organization's database and is a very powerful presentation of the its business requirements.

The data model is a guide used by data analysts and data scientists for design and implementation of a database. Data models serves multiple purposes such as high-level conceptual models to physical data models.

III. METHODOLOGY

1) Proposed Methodology-

This system develops a NoSQL database and makes use of Apriori algorithm as techniques for the store layout. Knowledge extraction is done using association rule mining results and is illustrated as useful knowledge patterns or rules and clusters to propose suggestions and solutions for an organization for store layout. First of all, we need to measure the relationship among various products transactional data with two files. The first file comprises of transactions during a certain time period. Each transaction has the purchase date, receipt no, and bill details. The second database file contains data about the product. In the first step of data mining we're going to transform the text files into NoSQL database file. This study establishes the relationship between the database tables and transfers them on database server using OBDC environment for the data table implementation. The text formatted data of transaction records was loaded into a NoSQL database for querying. In the product data file there are product name, No of Products and price details. We are choosing all the categories to construct correlation matrix for our analysis.

2) Apriori Algorithm-

Apriori is a calculation for regular thing set mining and association rule learning over value-based databases. It continues by recognizing the regular individual things in the databases and expanding them to bigger data sets as long as those frequent item set show up adequately frequently in the database. The successive thing sets dictated by Apriori can be hence utilized to decide association rules which feature the general pattern in the database.

There are three major components of Apriori algorithm-

A. Support
B. Confidence
C. Lift

A. Support-

Support is defined as the number of transactions containing a particular item divided by total number of transactions. Suppose we want to find support for item X. This can be calculated as-

Support(X) = (Transactions containing(X)/ (Total Transactions)

For example, if, out of 10 transactions, 1 transaction contain bread then the support for item bread can be calculated with the below formula-

Support(bread) = (Transactions containing bread) / (Total Transactions)

Support(bread) = 10/1 = 10%

B. Confidence-

Confidence is defined as the chances that an item X is also purchased if item Y is purchased. It can be calculated by finding the number of transactions where X and Y are bought together, divided by total number of transactions where Y is bought. Mathematically, it can be represented as-

Confidence(A→B) = (Transactions containing both (A and B)) / (Transactions containing A)

In our problem, we have 5 transactions where milk and bread are bought together. While in 15 transactions, milk is bought. Then we can find chances or likelihood of buying bread when milk is bought can be represented as confidence of milk -> bread and presented mathematically as-

Confidence(milk→bread) = (Transactions containing both (milk and bread)) / (Transactions containing Y)

Confidence(milk→bread) = 5/15 = 33.3%

C. Lift-

Lift (X -> Y) refers to the increase in the ratio of sale of Y when X is purchased. Lift (X -> Y) can be calculated by dividing Confidence (X -> Y) divided by Support(Y). Its mathematical representation is-

Lift(X→Y) = (Confidence (X→Y))/ (Support (Y))

Coming back to our milk and bread problem, the Lift (milk -> bread) can be calculated as-

Lift(milk→bread) = (Confidence (milk→bread)) / (Support (bread))

Lift(milk→bread) = 33.3/10 = 3.33

Lift refers to likelihood of buying an item X and Y together is 3.33 times more than the likelihood of just buying the Y. A Lift of 1 represents no association between products X and Y. Lift of greater than 1

means products X and Y are more likely purchased together. Finally, Lift of less than 1 means item X and Y are unlikely to be bought together.

IV. CONCLUSION

Data mining is a technique of extracting interesting patterns or knowledge from the huge amount of data or a database. This paper discusses the implementation of Association rule mining algorithms such as Apriori that is very useful for finding simple associations between various products in our data set in NOSQL database. NoSQL database is much faster to implement, very efficient and easy to understand. The Apriori algorithm is also easy to implement and is highly efficient to discover frequent item sets. This approach allows supermarkets and stores to cluster products in a way to increase the sales and growth opportunities which in turn helps in increasing profit and this is also known as Market Basket Analysis.

ACKNOWLEDGEMENT

First and foremost, praises and thanks to the god for shower of blessings throughout my research works to complete the research successfully. I would like to express my sincere gratitude to my parents for their love, care, blessings and sacrifices for educating and preparing me for future. I would also like to thank my brother and my friends for their constant support and encouragement. Last but not the least, my special thanks go to all the people who have supported me throughout to complete this research paper.

REFERENCES

[1] Chen, Y.-L., Tang, K., Shen, R.-J., Hu, Y.-H.: Market basket analysis in a multiple store environment, Decision Support Systems, 2004.

[2] Atkin, Charles K., "Observation of Parent Child Interaction in Supermarket Decision- Making," Journal of Marketing, October, 1978. February

[3] M. Hahsler, K. Hornik, "Building on the rules infrastructure for analyzing transaction data with R", in R. Decker, H.-J. Lenz (eds.), Advances in Data Analysis, pp. 449-456, Berlin: Springer, 2007.

[4] J. Han, M. Kamber, Data Mining: Concepts and Techniques, p. cm. London: Academic Press, 2001.