AFFINITY ANALYSIS AND ASSOCIATION RULE MINING USING FREQUENT PATTERN ALGORITHM IN R LANGUAGE

Namrata P. Bose, Ruchita S. Nehete, Neha R. Thakare, Rita R. Thakare
SSBT's College of Engineering and Technology,
Jalgaon, Maharashtra

Abstract : Affinity analysis is a data-mining technique, encompassing a broad set of analytic techniques that discovers co-ocurrence relationships among the activities performed by specific individuals or groups. Based on the recorded information, after the analysis, future behavior can be statistically predicted. The main purpose of the performing this analysis is to generate a set of rules to relate two or more products together. It is used in market basket analysis, in which retailer seek to understand the product buying behavior of customer which is used by them to influence their sales promotions, store design, discount plans and cross-selling. FP algorithm is implemented using 'R'. 'R' along with being programming language, is a strong statistical and analytical tool that has an excellent suite for market basket analysis in the ARules package.

IndexTerms - **Affinity Analysis, Association Rule, FP Growth, FP Tree**

I. INTRODUCTION

Affinity analysis and association rule learning encompasses a broad set of analytics techniques aimed at uncovering the associations and connections between specific object. These might be visitors to a website such as customers or audience, products in a store, or content items on a media site of this market basket analysis is perhaps the most famous example. Market basket analysis (MBA) uncovers associations between products by looking for combinations of products that frequently co-occur in transactions.

Data Mining is the essential process of discovering hidden and interesting patterns from massive amount of data where data is stored in data warehouse, OLAP (on line analytical process), databases and other repositories of information. This data may reach to more than terabytes. Data mining is also called (KDD) knowledge discovery in databases, and it includes an integration of techniques from many disciplines such as statistics, neural networks, database technology, machine learning and information retrieval, etc. Interesting patterns are extracted at reasonable time by KDDs techniques. KDD process has several steps, which are performed to extract patterns to user, such as data cleaning, data selection, data transformation, data preprocessing, data mining and pattern evaluation.

II. RELATED WORK

R. Karthiyayini and Dr. R. Balasubramanian proposed paper "Affinity analysis and Association Rule Mining using Algorithm in Market Basket Analysis" Apriori algorithm is one of the classical algorithms proposed by R. Srikant and R. Agrawal in 1994 for finding frequent patterns for Boolean association rules. Apriori is an algorithm for frequent item set mining and association rule learning over transactional databases.

Dr. Kanwal Garg and Deepak Kumar proposed "Comparing the Performance of Frequent Pattern Mining Algorithms" Eclat is a vertical database layout algorithm used for mining frequent item sets. It is based on depth first search algorithm. In the first step the data is represented in a bit matrix form. If the item is bought in a particular transaction the bit is set to 1 else to 0. After that a prefix tree needs to be constructed. To find the first item for the prefix tree the algorithm uses the intersection of the first row with all other rows and to create the second child the intersection of second row is taken with the rows following it. In the similar way all other items are found and the prefix tree get constructed. Infrequent rows are discarded from further calculations. To mine frequent item sets the depth first search algorithm is applied to prefix tree with backtracking. Frequent patterns are stored in a bit matrix structure.

Rahul Mishra proposed paper "Comparative Analysis of Apriori Algorithm and Frequent Pattern Algorithm for Frequent Pattern Mining in Web Log Data." Frequent pattern growth also labeled as FP Growth is a tree based algorithm to mine frequent pattern database unlike Apriori algorithm. It is applicable to projected 6 type database. It uses divide and conquer method unlike the Apriori and the Eclat algorithm. FP-Growth is an improvement of apriori designed to eliminate some of the heavy bottlenecks in apriori. The algorithm was planned with the benefits of map Reduce taken into account, so it works well with any distributed system focused on Map Reduce. FP-Growth simplifies all the problems present in apriori by using a structure called an FP-Tree. In it no candidate frequent item set is needed rather frequent patterns are mined from FP tree. FP tree makes use of the leafs and no des. In an FP-Tree each no de represents an item and its current count, and each branch represents a different association. Hence FP growth takes least memory because of projected layout and is storage efficient.

III. PROPOSED SYSTEM

The existing Affinity analysis system is based on Apriori algorithm, but the Apriori proves to be inefficient due to multiple scans over a database and if database is large it takes too much time to scan the whole database. So the proposed system is designed using the FP growth algorithm which reduces the multiple scans of the database by using the concept and hence reduces time of execution. As well 'R' language is used to design the system. The ARule packages in 'R' are used to implement the FP Growth algorithm efficiently.

3.1. SYSTEM ARCHITECTURE

The system architecture provide details of how the components or modules are integrated. A system architecture is the conceptual model that defines the structure, behavior, and more views of a system. An architecture description is a formal description and representation of a system, organized in a way that supports reasoning about the structures and behaviors of the system. A system architecture can comprise system components, the expand systems developed, that will work together to implement the overall system.

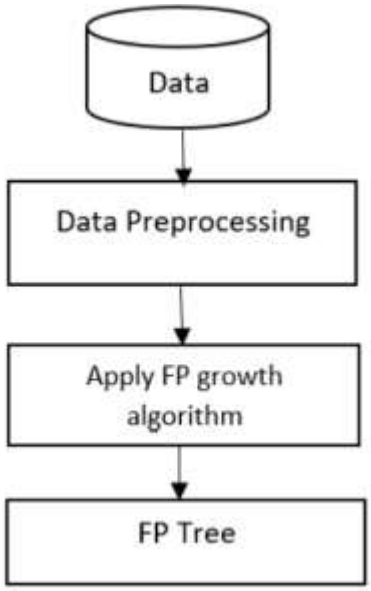

Figure 4.1: System Architecture

The system architecture here depicts the flow of the system in a short way. The system works in the main 3 steps:

1. The transactional data is the first input to the system.
2. The various data pre-processing steps like cleaning, integrating are applied on the data.
3. The FP algorithm is applied on the final database which is followed by the generation of the FP tree, showing the association rules.

3.2 FP-GROWTH ALGORITHM

1. First count all the items in all the transaction.
2. Calculate minimum support by using Formula:
min-support=(min-sup in%)*no. of transaction
3. Discard the items whose frequency is lesser than minimum support.
4. Sort the list according to the count of each item in descending order.
5. Sort the list according to the count of each item in descending order.
6. Build the tree going through each of transactions and add all the items in the order they appear in sorted list.
- **i.** For FP tree generation root node is taken as null
- **ii.** Add node with item name along with the count where count represent the number of repetition of item.
- **iii.** If another transaction contain item which are already present, then the count will be increased and the further item will be added in the tree.
- **iv.** Otherwise start a new branch or node.

3.3. IMPLEMENTATION DETAILS

The proposed system is designed to understand customer's buying habits using their transactional database and then produce association rules. The system is designed in R language using "sparklyr". R is a great statistical and graphical analysis tool,

well suited to more advanced analysis. The transactional database of grocery is taken as an input. The Frequent Pattern growth algorithm scans the dataset, compares the minimum support and minimum confidence and generates the FP tree. The FP tree gives the association rules for a random item that is selected. The process takes place only in 2 steps that is scanning the database and generating the FP tree using the FP growth algorithm.

V. RESULT AND DISCUSSION

The proposed system is designed on the basis of the Market Basket Analysis. The Market Basket Analysis is a very common method used by most of the retailers to increase their sales and profits. The algorithm like the Apriori are already in use for studying the customer product-buying habits. The proposed system uses the Frequent pattern growth algorithm to study the same which unlike Apriori takes less time and memory space.

After executing the code on the grocery database, it is observed that the number of transactions in the database are calculated properly the first step. At the second step the frequency of each item present in the transactions is calculated. Then the FP growth algorithm is applied on the them along with the approximate values of support and confidence. The FP growth algorithm then plots the expected tree which gives us the association rules required for the respective item selected. The tree changes for each item that is randomly selected.

Affinity analysis lets the retailers know about customers pro duct buying habits. Here in, Frequent Pattern growth is used to find out the association rules based up on the transactional data. Frequent Pattern algorithm gives the association rules in the form of tree which is easy to understand. The code is designed in R language because it is a statistical and analytical tool which can easily read large dataset and produce output in lesser time as compared to other programming languages. Thus the proposed system works on the grocery dataset and gives the output FP tree which gives the association rules for the item selected randomly. The output in the form of FP tree is generated within lesser time and gives a clear understanding the rules.

Experimental result of proposed system is shown in the below given diagrams. The first diagram shows the output FP tree of the transactional data. The second, third and the fourth diagram shows the associations for the items yogurt, whole milk and tropical fruit respectively.

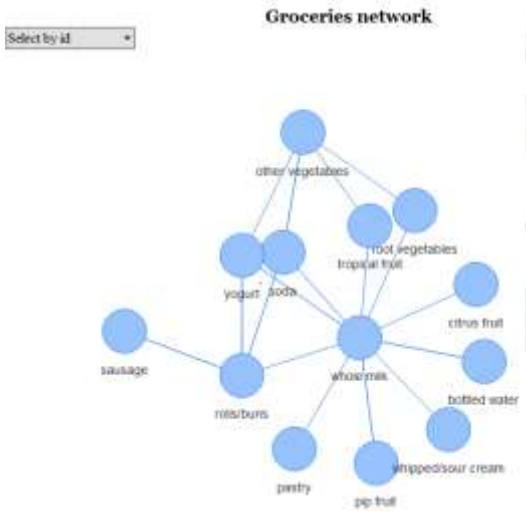

Figure 5.1: Output in the form FP-Tree

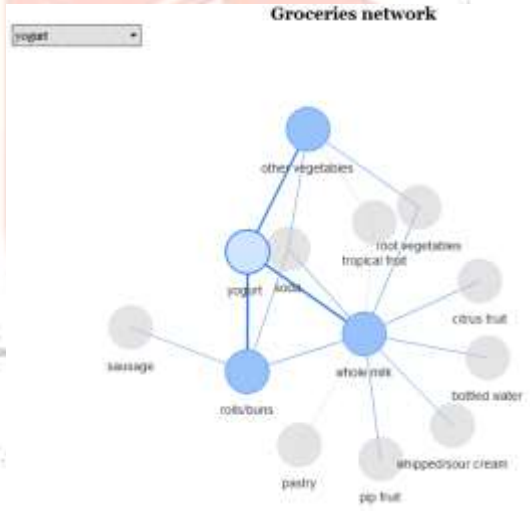

Figure 5.2: Association for yogurt

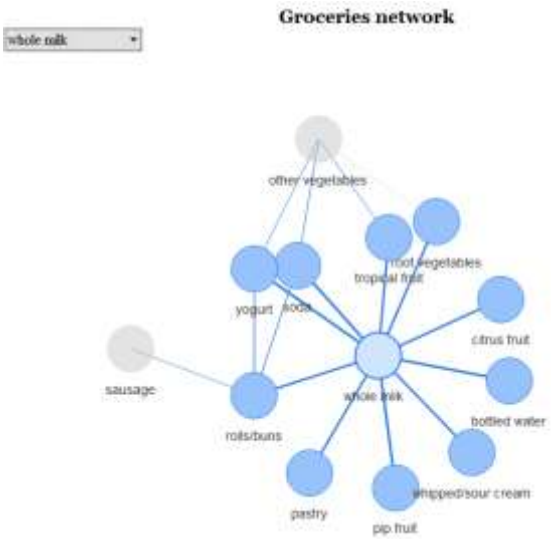

Figure 5.3: Association for whole milk

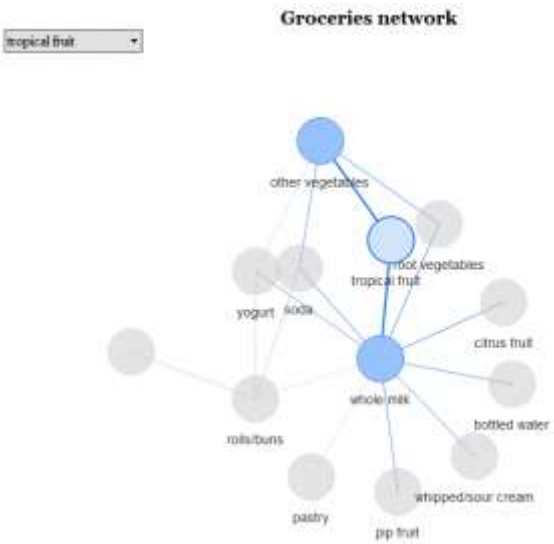

Figure 5.4: Association for tropical fruit

VI. CONCLUSION AND FUTURE WORK

The knowledge about customer's product-buying habit is of immense importance for a retailer to increase his sales. The system proposed analyses the customer buying behaviour. It solves the problem of time consumption due to involvement of FP tree as compared to the previous algorithm. It even requires less memory usage. As well, the growth will be more scalable because of its linear running time. So, it is possible that the system is useful to the users who want to analyse huge datasets with less memory and within less time.

As a thought of future work, this system may help the retailers and the publishers to analyse customer buying behaviour and then bring different plans into execution to increase their sales and profits.

REFERENCES

[1]A. Trnka, "Market basket analysis with data mining methods," ICNIT, 2010.
[2] D. R. B. R. Karthiyayini, "Affinity analysis and association rule mining using apriori algorithm in market basket analysis," vol. 6, p. 6, October 2016.
[3] V. P. Dr. M. Dhanabhakyam, Dr. M. Punithavalli, "A survey on data mining algorithm for market basket analysis," GJCST, vol. 11, July 2011.
[4] D. K. Dr. Kanwal Garg, "Comparing the performance of frequent pattern mining algorithms," International Journal of Computer Applications (0975 8887), vol. 69, no. 25, may 2013.
[5] N. S. Siddhrajsinh Solanki, "A survey on frequent pattern mining methods apriori, eclat, fp growth," International Journal of Computer Techniques.
[6] S. K. S. Jain, "Mining and optimization of association rules using effective algorithm," IJETAE, vol. 2, 2012.
[7] A. C. M. B. Sangita Chaudhari, Mayur Borkhatariya, "Implementation and analysis of improved apriori algorithm," IJESIT, vol. 5, p. 9, March 2016.