Association Rule Mining using Apriori Algorithm for Extracting Product Sales Patterns in Groceries

Mrs. M.Kavitha,
Assistant Professor & Ph.D Part time Research Scholar,
PG & Research Department of Computer Science and Applications, Vivekanandha College of Arts and Sciences for Women, Tiruchengodu, Tamilnadu, India

Dr. S. Subbaiah
Assistant Professor
Department of Computer Science,
Sri Krishna Arts and Science College,
Coimbatore, Tamilnadu, India

Abstract:- **Association Rule Mining is used for finding the patterns, associations and relationships in dataset. The rule is used for identifying the frequently occurs in item set. It helps to retailers to identify relationships among the items that people buy together frequently. It involves machine learning models to analyze the dataset for predicting patterns and co-occurrence. Many algorithms are used to generate the association rules. In this paper, we implemented apriori algorithm using R tool.**

Keywords: Data mining, Association Rule Mining, Apriori algorithm

I. INTRODUCTION

The software plays main part in businesses and organizations. A large amount of data is generated with the use of software. This datasets must be analyzed. Using that information the organization take some decision for their growth. For this process data mining is used for fulfilling this type of requirements. It applies some of the algorithms on the data and provides some useful information to organizations. To extract useful information from large data sets some of the techniques we applied. There are classification, clustering and association rule mining.

In retail industries collect and maintain the large amount of transactions data. It is very important using that data correctly and finds the hidden patterns and relationships among items should be identified. Using that information the retailers can find which products are frequently bought by customer along with other products. This information can help the person very useful for taking the business towards growth. It could help in supporting sales.

For this, Association Rule mining is used. This technique helps to find the frequent patterns, association, relationships and correlations and structures among the datasets in transactional database. Apriori algorithm works better for finding the frequent pattens. In this paper, we used R tool for implementing the apriori algorithm on groceries dataset.

The remaining of our paper is structured as follows: Section 2 specifies related research. Section 3 describes Association Rule Mining. Section 4 describes our use of R data mining tool for generating rules and their experimental results and Section 5 shows the conclusion of the research work.

II. RELATED WORK

Pramod Prasad [1] et al implemented an association rule mining in extracting patterns that occurred frequently .This will help to manage retail businesses and gave reports. These reports are very useful regarding prediction of product sales styles and customer behaviour. This will help the retailers to make better decisions. For their experimentation, they tested the Apriori algorithm in Weka. They successfully implemented the Apriori algorithm in a visual C#.Net application.

M Harahap [2] et al analyzed the patient prescriptions details and then identify the relationship among the diseases and what are the medicines used by the physicians for treating the patient's disease. The medicine selection is more important. They collected the details of patient's prescriptions and applied k-means clustering method to classify the top most 10 diseases and finally they applied apriori algorithm to find relationships among prescriptions and diseases. They used MySql database for cleaning and transforming of data. The assessment of support, confidence and lift between prescriptions and diseases is useful for recommending the correct medicine based on the state of disease of the particular patient.

Jayakumar Kaliappan [3] et al used Apriori algorithm for finding association rules to promote the sales and user interaction. They proposed the modified apriori algorithm. The proposed algorithm is 89.4% efficient than the normal algorithm.

Charanjeet Kaur [4] presented a survey of research paper regarding the association rule mining using Apriori algorithm.

III. ASSOCIATION RULE MINING USING APRIORI ALGORITHM

Association Rule Mining is called as a Association Rule Learning. It is technique for used to find the association among the variables which are present in a dataset. It is applied in many areas like grocery stores, business websites having transactional database.

Association Rule Mining is applied on the transactional database. A rule is a notation that represents which items are frequently bought with which items. It has two parts. They are LHS and RHS.

Defintion by Agarwal [5], The association rule mining defined as follows:

Let I = { i_1,i_2,i_3,….i_n}

Here I denote as a set of items.

Let D={ t_1,t_2,t_3,….t_n}

Here t denotes set of trasactions in database.

Each transactions has a unique transaction id is assigned. It contains a subset of the items I.

So a rule is defined as follows

$$X => Y, \quad X, Y \in I$$

For Example,

Itemset A => Itemset B

Here Itemset A is a LHS and Itemset B is a RHS. This denotes the items which are present in right are frequently bought items which are present in a left.

For this apriori algorithm is used to find rules from a given trasanctional database. In this algorithm, mostly three measures are used. They are

1. Support
2. Confidence
3. Lift

These are explained as follows:

Let's Consider the rule A => B

1. Support

   It is an indication used to find how frequently the itemset appears in a particular data set. It calculates as

$$\frac{Number\ of\ Transactions\ With\ Both\ A\ and\ B}{Total\ Number\ of\ Transactions}$$

   =P (A∩B)

2. Confidence

   It is an indication to find how often the rule has been found to be true.

$$\frac{Number\ of\ Transactions\ With\ Both\ A\ and\ B}{Total\ Number\ of\ Transactions\ with\ \ A}$$

$$= \frac{P(A \cap B)}{P(A)}$$

3. Lift

   It is calculated using these two values. Confidence and Expected Confidence Values.

$$\frac{Confidence}{Expected\ Confidence}$$

$$= \frac{P(A \cap B)}{P(A).P(B)}$$

IV. RESULTS

We implemented the Apriori algorithm for groceries dataset using R Tool. R tool is mainly used in data mining. It is an open source tool and its very interesting to learn the concepts. It has number of predefined packages are available to do the algorithms.

For implementing Apriori Algorithm the arules package is used. The arulesViz package is used for visualizing the results.

```
>library(arules)
>library(arulesViz)
```

After loaded the packages, we load the dataset.

```
>data(Groceries)
```

The dataset of Groceries will be loaded into environment. Then summary of dataset checked.

```
>summary(Groceries)
```

It displays the following results. The dataset has 9835 transactions and 169 items.

```
transactions as itemMatrix in sparse format with
 9835 rows (elements/itemsets/transactions) and
 169 columns (items) and a density of 0.02609146

most frequent items:
     whole milk other vegetables       rolls/buns             soda
yogurt
           2513            1903            1809            1715
1372
        (Other)
          34055

element (itemset/transaction) length distribution:
sizes
   1    2    3    4    5    6    7    8    9   10   11   12   13   14   15
16   17   18
2159 1643 1299 1005  855  645  545  438  350  246  182  1
17   78   77   55   46   29   14
  19   20   21   22   23   24   26   27   28   29   32
  14    9   11    4    6    1    1    1    1    3    1

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   2.000   3.000   4.409   6.000  32.000

includes extended item information - examples:
       labels  level2          level1
1 frankfurter sausage meat and sausage
2     sausage sausage meat and sausage
3  liver loaf sausage meat and sausage
```

```
> apriori(Groceries,parameter = list(support=0.002,confide
nce=0.5))->rule1
>inspect(head(rule1,5))
```

lhs	rhs	support	confidence	lift	count
[1] {cereals}	=> {whole milk}	0.003660397	0.6428571	2.515917	36
[2] {jam}	=> {whole milk}	0.002948653	0.5471698	2.141431	29
[3] {specialty cheese}	=> {other vegetables}	0.004270463	0.5000000	2.584078	42
[4] {rice}	=> {other vegetables}	0.003965430	0.5200000	2.687441	39
[5] {rice}	=> {whole milk}	0.004677173	0.6133333	2.400371	46

In this above result a person who likes to buy cereals also like to buy whole milk. Like that a person who likes to buy jam also like to buy whole milk. The support value 0.003660397 represents 0.3% of the all the orders represents the LHS and RHS combination. The Confidence value 0.6428571 represents the orders contains cereals 64% of them like to buy whole milk. The Lift value 2.515917 tell us how significant the consequent with respect to the antecedent. So all of these values are 2 times significant.

```
>inspect(head(sort(rule1,by="lift"),5))
```

lhs rhs support confidence lift count

[1] {butter, hard cheese} => {whipped/sour cream} 0.002033554 0.5128205 7.154028 20

[2] {beef, citrus fruit, other vegetables} => {root vegetables} 0.002135231 0.6363636 5.838280 21

[3] {citrus fruit, tropical fruit, other vegetables, whole milk} => {root vegetables} 0.003152008 0.6326531 5.804238 31

[4] {citrus fruit, other vegetables, frozen vegetables} => {root vegetables} 0.002033554 0.6250000 5.734025 20

[5] {beef, tropical fruit, other vegetables} => {root vegetables} 0.002745297 0.6136364 5.629770 27

Then order by lift value we used like this. These result will be plotted.

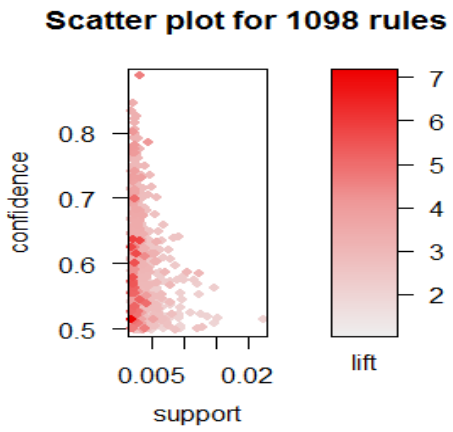

Fig1: Scatter plot

X axis represents support values and y axis represents confidence value. The dark red dots are representing occurs low supporting values. The Fig 2: represents Grouped matrix for 1098 rules.

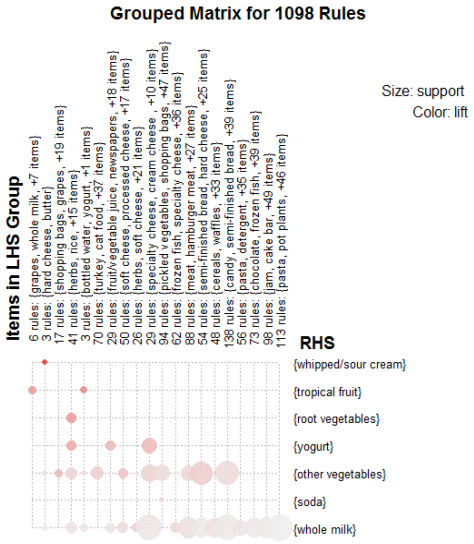

Fig 2: Grouped Matrix

It displayed set of LHS and RHS Values. And the support value is based on the size and the lift value based on the dark color.

```
>apriori(Groceries,parameter=list(support=0.002,confidence=0.5,minlen=5))->rule2
>inspect(head(rule2,4))
```

This is the second rule we build that minval=5.The result is

lhs rhs support confidence lift count

[1] {tropical fruit, other vegetables, butter, yogurt} => {whole milk} 0.002338587 0.7666667 3.000464 23

[2] {tropical fruit, whole milk, butter, yogurt} => {other vegetables} 0.002338587 0.6969697 3.602048 23

[3] {tropical fruit, other vegetables, whole milk, butter} => {yogurt} 0.002338587 0.6969697 4.996135 23

[4] {other vegetables, whole milk, butter, yogurt} => {tropical fruit} 0.002338587 0.5348837 5.097463 23

The Fig 3 represents the grouped matrix of rule 2. Then rule3 will be formed and displayed the following results.

```
>apriori(Groceries,parameter=
list(support=0.007,confidence=0.6))->rule3
> inspect(head(rule3,4))
```

```
     lhs                     rhs            support confidence     lift coun
t
[1] {root vegetables,
     butter}            => {whole milk} 0.008235892  0.6377953
2.496107    81
[2] {butter,
     yogurt}            => {whole milk} 0.009354347  0.638888
9 2.500387    92
[3] {tropical fruit,
     other vegetables,
     yogurt}            => {whole milk} 0.007625826  0.619834
7 2.425816    75
[4] {root vegetables,
     other vegetables,
     yogurt}            => {whole milk} 0.007829181  0.606299
2 2.372842    77
```

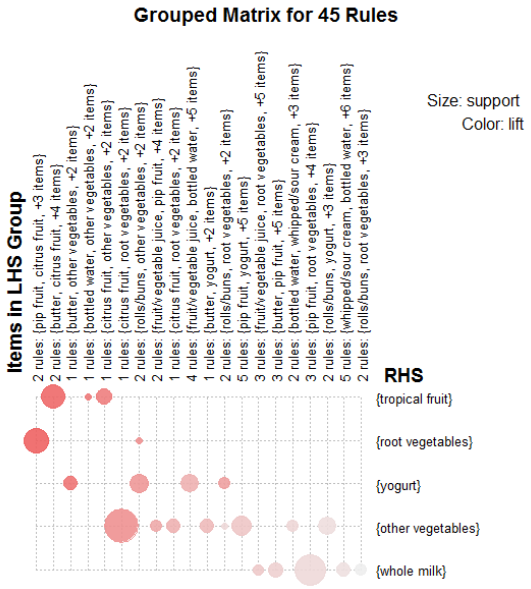

Fig 3: Grouped Matrix of rule2

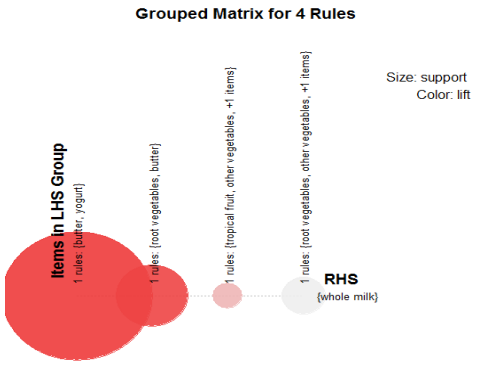

Fig 4: Grouped matrix for rule3

The figure 4 represents the grouped matrix for rule 3. The LHS and RHS value is listed. In this figure a person who likes to buy butter and yogurt definitely liked to buy whole milk.

V. CONCLUSION

The Association rule mining is very useful for analyzing datasets which are collected in supermarket. So the manager can know what are products purchased frequently with what are the items buying together by the customer. It will be used for taking decisions and promotes their sales. And also they can provide combo offers and some other offers also.

REFERENCES

[1] Prasad et al, "Using Association Rule Mining for Extracting Product sales patterns in Retail Store Transactions", International Journal on Computer Science and Engineering (IJCSE) ISSN:0975-3397, Vol.3(5) May 2011, 2177-2182

[2] M Harahap et al, "Mining association rule based on the diseases for recommendation of medicine need", Journal of physics, IOP Conf. Series: Journal of Physics: Conf. Series 1007 (2018) 012017 doi :10.1088/1742-6596/1007/1/012017

[3] Jayakumar Kaliappan et al," Weblog and retail industries Analysis using a robust modified Apriori algorithm", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-6, April 2019

[4] Charanjeet Kaur. "Association Rule Mining using Apriori algorithm: A survey", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 2, Issue 6, June 2013 ISSN: 2278 – 1323

[5] Agrawal R, Imielinski T, Swami A, "Mining association rules between sets of items in large databases". Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD'93. CiteSeerX 10.1.1.40.6984. doi:10.1145/170035.170072. ISBN 978-0897915922.