RECOMMENDATION MODEL FOR MARKET BASKET ANALYSIS USING APRIORI ALGORITHM

MR. N. MUTHURASU[1], K. AMIRTHANANTH[2], R. MUKESH KUMAR[3], B. AARTHI SIKARWAR[4]

[1] *Department of Computer Science and Engineering, SRMIST, Tamil Nadu, India*
[2] *Department of Computer Science and Engineering, SRMIST, Tamil Nadu, India*
[3] *Department of Computer Science and Engineering, SRMIST, Tamil Nadu, India*
[4] *Department of Computer Science and Engineering, SRMIST, Tamil Nadu, India*

ABSTRACT

Ecommerce website that sells products such as groceries, sauces, condiments and others perishable products. The processes done by this company are purchasing products directly from farmers and selling them to the customer. Based on the existing system, this could be implemented with the help of Data Mining using association rules. The end result of implementing this application is that it can help the process of buying, sales, and using MBA methods with predictive analysis is that it can recommend the company with items based on customers purchase patterns and it can improve the sales.

Keywords: *Data Mining, Market basket analytics, Product management, Apriori Algorithm , Recommendation system, Predictive modelling.*

1.INTRODUCTION

Generally, in an E-Commerce website there are hundreds to thousands of transactions happening daily, in the face of so much information, if we had to find the relation between merchandises and make decisions on business for selling the right product, what to sell the right customer for maximizing the profit , etc, it is a very difficult task if we had no effective way to carry out the process. Therefore, for understanding the relationship among the transacted items, many algorithms are introduced to mine the frequent itemsets. Apriori Algorithms is a well-known approach for this process. The main goal of this analysis is to develop a suitable recommendation method to find out which products are often purchased by customers together with the traits that influence it, such as time of purchase, invoice id and the products details, using MBA.

2.LITERATURE REVIEW

2.1. Data Mining

It is the process of analysing unseen patterns in data sets involving methods which are a combined approach in machine learning, database systems and in statistics. From to a different viewpoint for classification into useful information, which is collected and organized in areas, such as data-warehouses, for efficiency in analytics, mining patterns, which involves industry decision making and other information required to ultimately cut the expense and increase the ROI. Data mining is also considered to be an efficient way to discover the hiding data and it is carried out with the help of a lot of algorithms.

The methods involved in this process are:

- Extract, convert and upload data into a data warehouse
- Accumulate and maintain data in a multi-dimensional database.
- Provide access to the analysts in the industry.
- Present the analysed facts in easily understandable forms, such as graphs.

2.2. Market Basket Analysis (MBA)

Market basket analysis is an association in data mining to find attributes that appear in one time [2]. Market Basket Analysis(MBA) is a widely used practice among the Marketers to discover the best potential combination of the products or services which are repeatedly bought by the customers. Association analysis frequently based on an algorithm named "Apriori Algorithm". The result of this analysis is called association rules. Marketers use these rules to plan their recommendations.

Market Basket Analysis(MBA) looks at the purchase chance with the items purchased among the transactions. When two or more items are purchased, Market Basket Analysis(MBA) is done to check whether the purchase of one item increases the likelihood of the purchase of other item. This facts is a tool for the marketers to bundle the items or plan a product cross sell to a customer.

2.3. Apriori Algorithm

Apriori algorithm is an algorithm that is used for mining the frequent item sets and constructing the association rules from a transactional database[3]. It is devised in such a way that it operate on a database that contains a number of transactions, for example, the products bought by a customer in a shop. The association rules constructed from these kinds of details can be used to analyse or predict and find out how often one would buy PRODUCT A with PRODUCT B.

It is very important for successful MBA and it helps the customers to purchase items which increases the sales of the product. With the help of this knowledge, the owner can change placement of the goods in a shelf or design a marketing campaign by using a combination of the results that resulted in certain products. Based on the parameters of support and confidence , the association methodology can be separated into two phases, namely: Analysis of the highest frequency pattern and establishment of association rules.

3.ARCHITECTURE

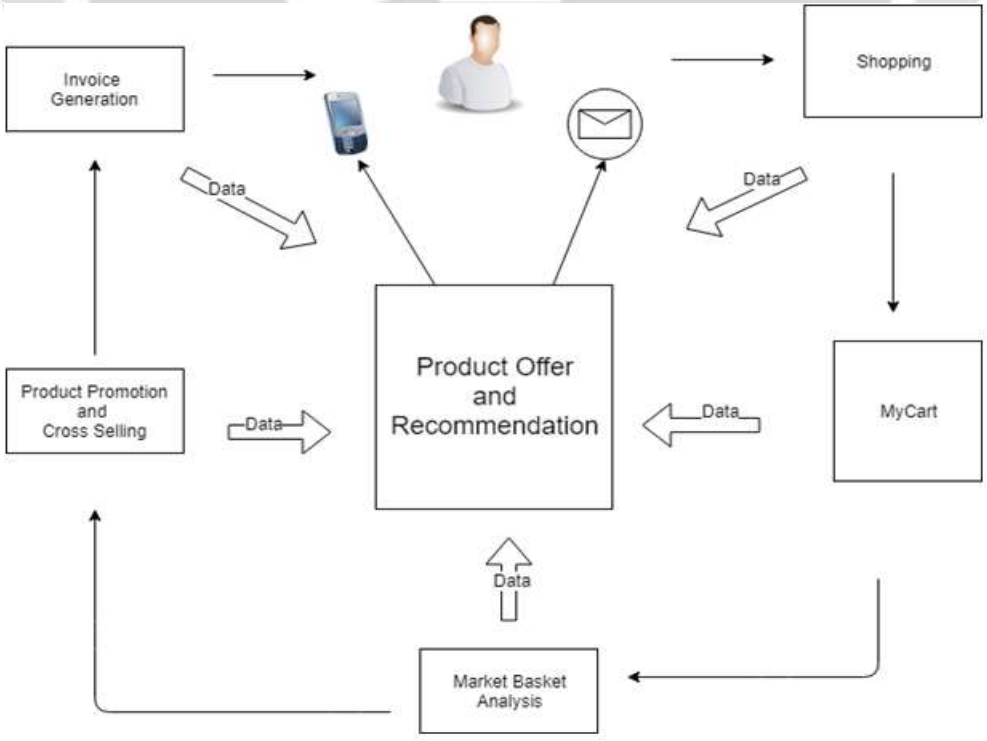

Fig-1 : Architecture of the Designed System

4.METHODOLOGY

In the designed system, we have implemented four different steps namely, Data Cleaning, Exploratory Data Analysis (EDA), Market Basket Analysis (MBA) and Predictive Modelling. We used the dataset open-sourced from the E-commerce company. This paper targets the analysis of the users based on their purchase in the website. Our ultimate aim is to produce a few graphical representations (charts, graphs, histograms, etc) to show the product sale characteristics and the show the seller which product to recommend the user. So, the parameters we wish to analyse are product-id, order-id and customer-id. We initially perform the data cleaning part followed by exploring the dataset for further inferences on the data.

4.1.Data collection

The process of gathering and quantifying the info on the variables of interest, in a fashionable method that enables a person to answer the stated questions, test hypotheses, and evaluate the outcomes. It is very important to ensure the accuracy and the integrity of the data collected.

A proper data collection process is needed as it ensures the data obtained are both definite and accurate and that further decisions based on the points mentioned in the inferences are valid[4]. Accuracy of the data collected is very essential to maintain the integrity of the work. All the steps, the selection of relevant data collection instruments (already existed, modified and newly developed systems) and clearly explained instructions for their correct use reduce the occurrence of error.

4.2.Data Cleaning

Data Cleaning or Data Cleansing is the process of identifying and cleaning (or removing) dirty or corrupt data from a database , record, or table and it refers to finding inaccurate, irrelevant, incomplete or incorrect parts of the data and then modifying, deleting or replacing the coarse or dirty data.[5]

4.3.Exploratory Data Analysis

Exploratory data analysis (EDA) is the process of analysing the dataset, database, or table to find potentially useful information, which is a difficult task. At the start of any EDA, the analyst has no idea about what he is looking at or what he is going to find.[6].

Exploratory Data Analysis (EDA) employs a variety of techniques (mostly graphical) for the analysts to look into for

1. Detecting Outliers and Anomalies in the data;
2. Extracting the important variables;
3. Uncovering the structures that pre-exist;
4. Maximizing the insight into data set;
5. Determine the optimal factor settings;
6. Test the underlying assumptions; and
7. Develop a model accordingly[7].

Exploratory Data Analysis (EDA) is the most important step in any data analysis. Here, you learn the data and the value it has and figure out methods to frame question to ask the data, as well as how best to alter the data resources to get answers for the questions you ask**.**

4.4.Predictive Analysis

Predictive analysis comprises of a variety of techniques from machine learning, data mining, predictive modelling that analyse historical and real-time stats to make predictions for the future [8]. Predictive analytics solutions involve extracting info from existing data from different sources, and determining patterns, and predicting the future trends and upcoming events. It uses different methods to make such future predictions, such as AI, Predictive Modelling, etc. These solutions are a useful and dependable methods for forecasting, since they also focus on risk management and takes anamolies into account. Moreover, it helps organizations to help adjust to the upcoming innovations and get ahead with the needs of the industry.

5.RESULT AND ANALYSIS

5.1.Exploratory Data Analysis

The data used by us includes 400,000+ records and 10 different fields: InvNo, StockCode, Desc, Qty, InvDate, UnitPr, CustID, Country, Date, and Time. Fig 1 Shows the Glimpse of the data that is gathered. Key roles are played by InvDate, CustID, Qty and InvNo

```
Observations: 406,829
Variables: 10
$ InvoiceNo   <dbl> 536365, 536365, 536365, 536365, 536365, 536365, 536365, 53636...
$ StockCode   <fct> 85123A, 71053, 84406B, 84029G, 84029E, 22752, 21730, 22633, 2...
$ Description <fct> WHITE HANGING HEART T-LIGHT HOLDER, WHITE METAL LANTERN, CREA...
$ Quantity    <int> 6, 6, 8, 6, 6, 2, 6, 6, 6, 32, 6, 6, 8, 6, 6, 3, 2, 3, 3, 4, ...
$ InvoiceDate <fct> 01-12-2010 08:26, 01-12-2010 08:26, 01-12-2010 08:26, 01-12-2...
$ UnitPrice   <dbl> 2.55, 3.39, 2.75, 3.39, 3.39, 7.65, 4.25, 1.85, 1.85, 1.69, 2...
$ CustomerID  <int> 17850, 17850, 17850, 17850, 17850, 17850, 17850, 17850, 17850...
$ Country     <fct> United Kingdom, United Kingdom, United Kingdom, United Kingdo...
$ Date        <date> 0001-12-20, 0001-12-20, 0001-12-20, 0001-12-20, 0001-12-20, ...
$ Time        <dbl> 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8...
```

Fig-2 : Glimpse of Data

In Fig 2 and 3, We are able to identify the products that are being sold in higher quantities. Fig 2 is a graphical representation in which the Frequency and the name of the product is listed on either side of the axes. These products bring a bigger market for the company and increase the credibility for the purchase of other products sold by the company.

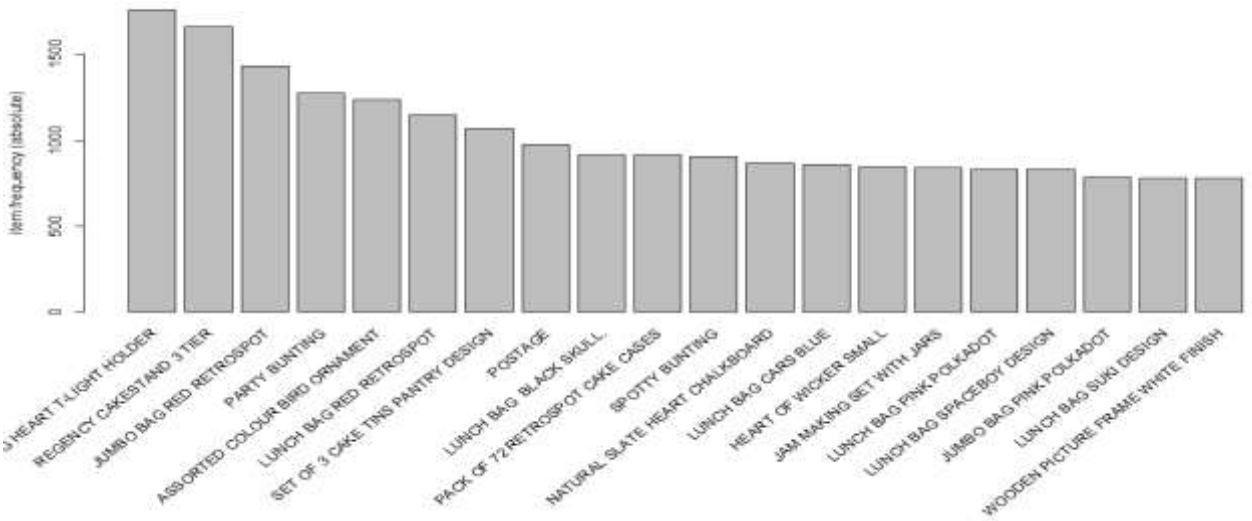

Fig-3 : Frequently Sold Item Plot

```
# A tibble: 10 x 2
   Description                          count
   <fct>                                <int>
 1 WHITE HANGING HEART T-LIGHT HOLDER    2070
 2 REGENCY CAKESTAND 3 TIER              1905
 3 JUMBO BAG RED RETROSPOT               1662
 4 ASSORTED COLOUR BIRD ORNAMENT         1418
 5 PARTY BUNTING                         1416
 6 LUNCH BAG RED RETROSPOT               1358
 7 "SET OF 3 CAKE TINS PANTRY DESIGN "   1232
 8 POSTAGE                               1196
 9 LUNCH BAG  BLACK SKULL.               1126
10 PACK OF 72 RETROSPOT CAKE CASES       1080
```

Fig-4 : Top 10 Best Selling Products Count

Most of the people who have made the purchase have bought less than 10 products per order. We are able to get the representation based on the Invoice No and the quantity of the items purchased.

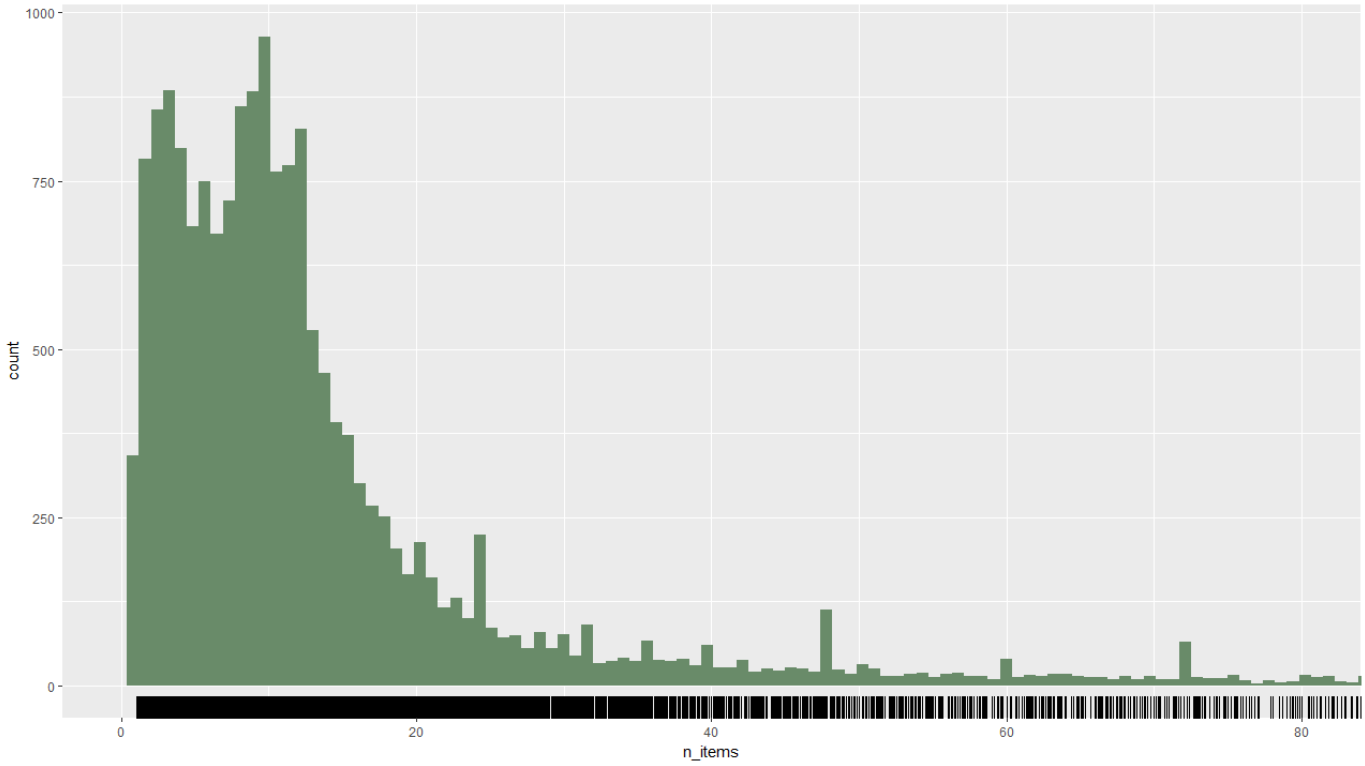

Fig-5 : Number of items per invoice distribution.

We use the Arules library from CRAN Package network to mine the rules and frequent itemsets using the Apriori Algorithm. We set the support at 0.1% and confidence at 80%. The number of rules generated were 89,697. Higher number of rules were generated for items which had a length of 6 which we were able to identify with the summary of the rules.

```
> inspect(rules[1:20])
     lhs                          rhs                  support     confidence lift      count
[1]  {WOBBLY CHICKEN}          => {METAL}              0.001454621 1          384.98000 28
[2]  {WOBBLY CHICKEN}          => {DECORATION}         0.001454621 1          384.98000 28
[3]  {DECOUPAGE}               => {GREETING CARD}      0.001194867 1          343.73214 23
[4]  {BILLBOARD FONTS DESIGN}  => {WRAP}               0.001506572 1          620.93548 29
[5]  {WOBBLY RABBIT}           => {METAL}              0.001766326 1          384.98000 34
[6]  {WOBBLY RABBIT}           => {DECORATION}         0.001766326 1          384.98000 34
[7]  {BLACK TEA}               => {SUGAR JARS}         0.002337784 1          211.52747 45
[8]  {BLACK TEA}               => {COFFEE}             0.002337784 1           60.91456 45
[9]  {FUNK MONKEY}             => {ART LIGHTS}         0.001974129 1          506.55263 38
[10] {ART LIGHTS}              => {FUNK MONKEY}        0.001974129 1          506.55263 38
[11] {CHOCOLATE  SPOTS}        => {SWISS ROLL TOWEL}   0.002233882 1          401.02083 43
[12] {WHITE TEA}               => {SUGAR JARS}         0.003324848 1          211.52747 64
[13] {WHITE TEA}               => {COFFEE}             0.003324848 1           60.91456 64
[14] {METAL}                   => {DECORATION}         0.002597538 1          384.98000 50
[15] {DECORATION}              => {METAL}              0.002597538 1          384.98000 50
[16] {MAGIC GARDEN}            => {HOOK}               0.002649488 1          377.43137 51
[17] {HOOK}                    => {MAGIC GARDEN}       0.002649488 1          377.43137 51
[18] {MAGIC GARDEN}            => {1 HANGER}           0.002649488 1          377.43137 51
[19] {1 HANGER}                => {MAGIC GARDEN}       0.002649488 1          377.43137 51
[20] {HOOK}                    => {1 HANGER}           0.002649488 1          377.43137 51
```

Fig. 6 Rules Generated.

The above 20 rules which were generated in the console were plotted into a Scatterplot graph which gives a clear understanding of the relation between the products which have a good support and confidence level.

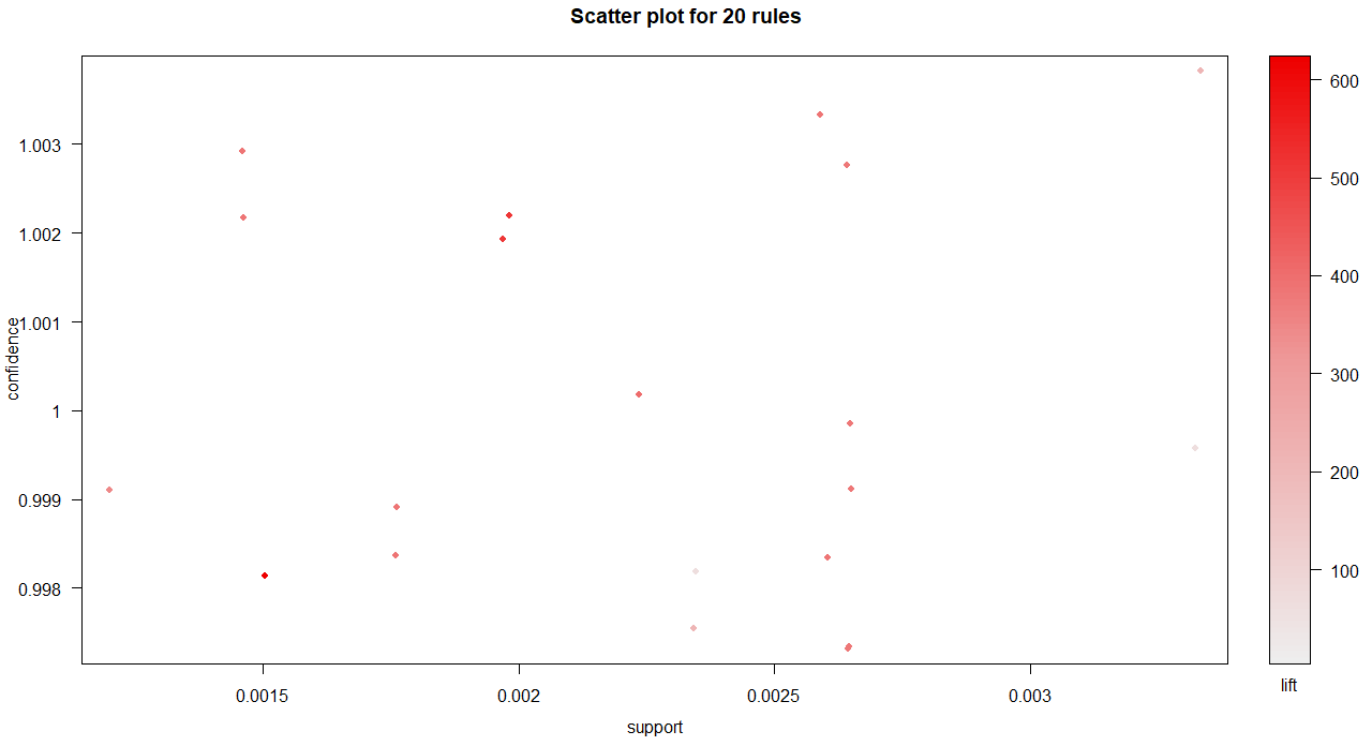

Fig-7 : Top 10 Scatter plot 20 rules.

In Fig 8, we are able to identify the relation between products that are being sold together. The items and rules are represented as vertices connecting them with the directed edges to the other product.

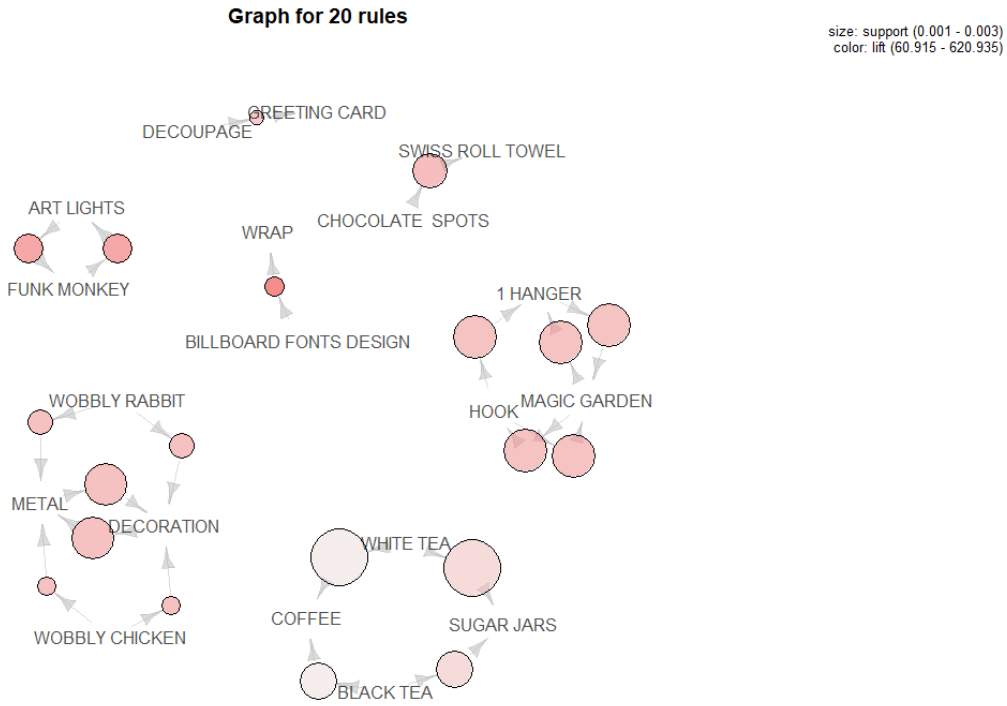

Fig-8 : Graph for 20 rules.

6.CONCLUSION

In this paper, we have shown the methodology used by us in analysing the customers purchasing patterns in a quicker and efficient way where it is done using the Apriori and Association rule mining algorithm. This can be implemented in real-time for any E-commerce company with a large number of customers and products.

7.REFERENCES

[1]. Trevor .H, Robert .T, Jerome Friedman (September 2, 2003) The Elements of Statistical Learning: Data Mining, Inference, and Prediction (3rd ed.). ISBN13: 9780387952840.

[2] Anand Rajaraman, Jeffrey David Ullman , "Mining of Massive Datasets", First Edition., Cambridge University Press December 2011, ISBN-13: 9781107015357

[3] Ramakrishnan Srikant and Rakesh Agrawal, "Fast algorithms for mining association rules in Large Databases". pages 488-500, Morgan Kaufmann Publishers Inc. San Francisco, September 12 - 15, 1994.

[4] Roger Sapsford and Victor Jupp , "Data Collection and Analysis", SAGE publications ltd, March 2006, ISBN-13: 9780761943631

[5]Shaomin Wu, Reliability Engineering & System Safety, "A review on coarse warranty data and analysis", *Volune 11 – Pages 1–11, June2013.*

[6][Daniel A. Keim, Florian Mansmann, Jorn Schneidewind, Hartmut Ziegler, Challenges in Visual Data Analysis, Proceedings of Information Visualization, Published in Proceedings of the conference on Information Visualization, pages 9-16, July 05-07, 2006.

[7] James J. Filliben, NIST Interagency/Internal Report (NISTIR), " e-Handbook of Statistical Methods", published in November 01, 2002.