ISSN 2320 - 2602

Volume 7 No.1, January 2018

International Journal of Advances in Computer Science and Technology

Available Online at http://www.warse.org/IJACST/static/pdf/file/ijacst01712018.pdf

https://doi.org/ 10.30534/ijacst/2018/01712018

A Big Data Analysis and Mining Approach for IoT Big Data

Jihyun Song[1], Kyeongjoo Kim[2], Minsoo Lee[3]

[1]Dept. of Computer Science and Engineering, Ewha Womans University, Seoul, Korea, Email:
Email:ssongji7583@ewhain.net

[2] Dept. of Computer Science and Engineering, Ewha Womans University, Seoul, Korea, Email:
Email:kjkimkr@ewhain.net

[3] Dept. of Computer Science and Engineering, Ewha Womans University, Seoul, Korea,
Email:mlee@ewha.ac.kr

ABSTRACT

These days, large amounts of data are produced by various ways such as stock data, market basket transactions, IoT sensors, etc. Such data can be accumulated and analyzed to provide helpful information in our lives. With the rapid development of IoT sensors and automated markets, the market basket data can be automatically generated. For these reasons, we choose the market basket data to analyze and find the association rules between big data sets. To do the data mining, we use R programming[1], which helps to organize the data and to visualize the data sets.

Key words : Association Rules, Market Basket Analysis, R programming, Data mining

1. INTRODUCTION

Recently, as the relationship between data becomes more important, this study analyzes and visualizes the relationship by using the grocery data. Organizing the data set is a part of data mining method and we used program R, with different kinds of libraries. This study investigated the support, reliability and lift of the association rule, after designating the items as whole milk.

2. Related Research

2.1 Data Mining

Data mining[2][3] is the process of finding useful information that is not easily exposed in vast amounts of data. It is a methodology for finding patterns and trends of specific types that extract patterns from data and generate models. A particular data model forms a kind of cluster that describes the relationship between data sets. In other words, the data is analyzed by the method that easily refines data, statistically analyzes and presents hypotheses. As the kinds of data are diversified, analysis methods for unstructured data have been proposed. The most basic analytical tools for handling large-scale structured and unstructured data are Hadoop, NoSQL, and R is used as a tool to analyze the analyzed data with a focus on visualization.

2.2 Association Rule

Association rule learning[4] is a rule-based machine learning method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness. It can find frequent patterns, associations, correlations, or causal structures among sets of items in transaction databases. By using associations, it is possible to understand customer buying habits and correlations between the different items that customers place in their shopping basket.

3. Big Data Analysis Approach

Below the figures indicate the source code using R or the result of the data using different relation in each different circumstance. Preferentially programming was done with only showing the basic relations and the status of the market basket data set. Then, application using different correlation and showing the graphs that are visualized by using R.

3.1 Data Preprocessing

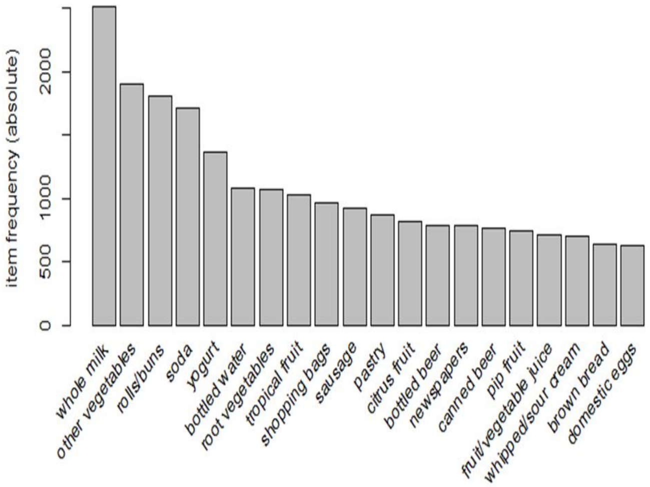
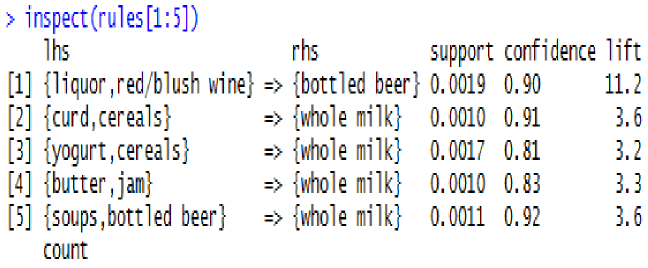


Figure 1: Item frequency plot for the top 20 items

We first loaded the necessary packages and data for data preprocessing. The package name “arules” is required. Also new name for the market basket data set is given as Groc that is composed of 169 items with 9835 transactions. If you look at the 20 most frequently viewed data prior to the data preprocessing process, it is as shown in Figure 1.

3.2 Mining Rules

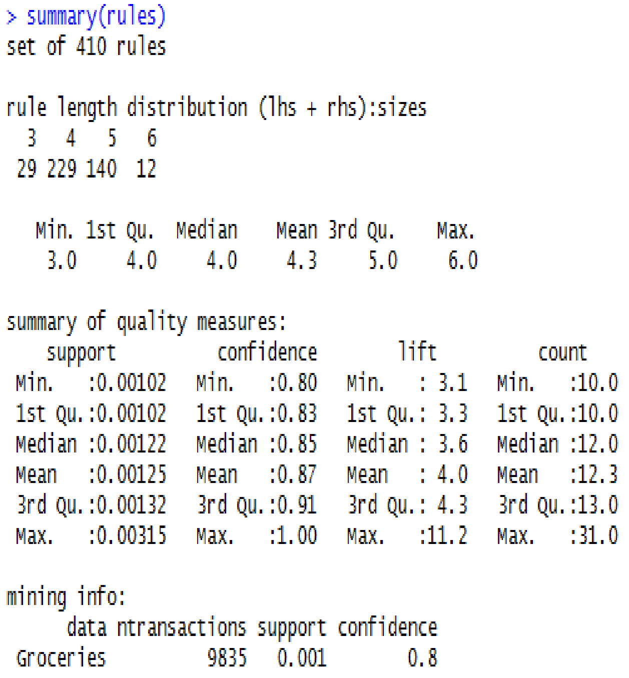
For the mining rules, we set the minimum required support and confidence. Support in association rules is an indication of how frequently the itemset appears in the dataset and confidence is an indication of how often the rule has been found to be true. So, we set the minimum support to 0.001 and minimum confidence of 0.8. After getting the rules we looked at the top5 rules, which is shown as Figure 2.



```
> inspect(rules[1:5])
    lhs                         rhs            support confidence lift
[1] {liquor,red/blush wine}  => {bottled beer} 0.0019  0.90       11.2
[2] {curd,cereals}           => {whole milk}   0.0010  0.91        3.6
[3] {yogurt,cereals}         => {whole milk}   0.0017  0.81        3.2
[4] {butter,jam}             => {whole milk}   0.0010  0.83        3.3
[5] {soups,bottled beer}     => {whole milk}   0.0011  0.92        3.6
    count
```

Figure 2: Top5 rules in support of 0.001 and confidence of 0.8

In addition to the explanation of Figure 2, when you see the third rule it indicates if someone buys yogurt and cereals, they are 81% likely to buy whole milk too.



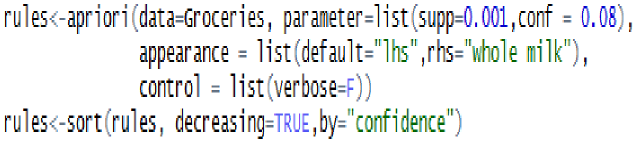
```
> summary(rules)
set of 410 rules

rule length distribution (lhs + rhs):sizes
  3   4   5   6 
 29 229 140  12 

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    3.0     4.0     4.0     4.3     5.0     6.0 

summary of quality measures:
    support           confidence          lift           count     
 Min.   :0.00102   Min.   :0.80   Min.   : 3.1   Min.   :10.0  
 1st Qu.:0.00102   1st Qu.:0.83   1st Qu.: 3.3   1st Qu.:10.0  
 Median :0.00122   Median :0.85   Median : 3.6   Median :12.0  
 Mean   :0.00125   Mean   :0.87   Mean   : 4.0   Mean   :12.3  
 3rd Qu.:0.00132   3rd Qu.:0.91   3rd Qu.: 4.3   3rd Qu.:13.0  
 Max.   :0.00315   Max.   :1.00   Max.   :11.2   Max.   :31.0  

mining info:
      data ntransactions support confidence
 Groceries          9835   0.001        0.8
```

Figure 3: Summary of Rules

When we get the summary information of rules as shown in Figure 3, we can get some information such as the number of rules generated are 410 and most of rules are 4 items long. And the summary of quality measure shows the ranges of support, confidence and lift. Lastly the information on the data mined shows the total data mined and minimum parameters. However, these rules are not sorted so we sorted them by confidence. When sorting with confidence we can find the most relevant rules. After sorting we eliminate the rules that are repeated.

4. Experimental Results

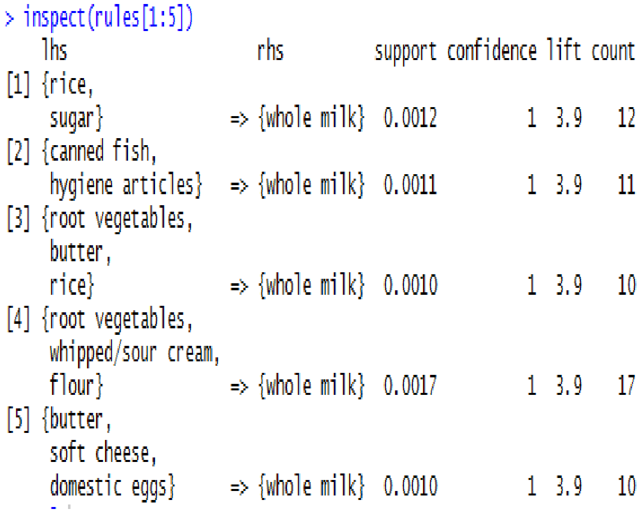
After setting the following rules that are done in implementation step we wanted to target item to generate rules. We illustrated with an example of whole milk and setting it either Left Hand Side and Right Hand Side to see what customers are likely to buy before buying whole milk and what customers likely to but if they purchase whole milk. We adjust apriori[5] function as Figure 4.



```
rules<-apriori(data=Groceries, parameter=list(supp=0.001,conf = 0.08),
               appearance = list(default="lhs",rhs="whole milk"),
               control = list(verbose=F))
rules<-sort(rules, decreasing=TRUE,by="confidence")
```

Figure 4: Targeting Item code using apriori function

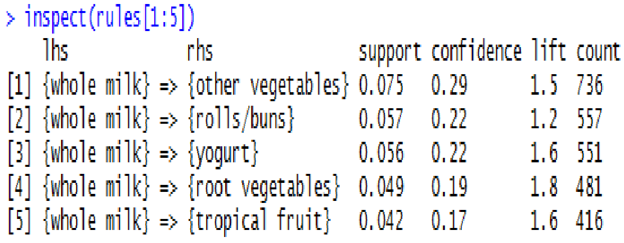
When we execute the code in Figure 4, we can get the same result as Figure 5, which is the result of showing itemset before buying milk.



```
> inspect(rules[1:5])
    lhs                     rhs          support confidence lift count
[1] {rice,                                                            
     sugar}              => {whole milk} 0.0012           1  3.9    12
[2] {canned fish,                                                     
     hygiene articles}   => {whole milk} 0.0011           1  3.9    11
[3] {root vegetables,                                                 
     butter,                                                          
     rice}               => {whole milk} 0.0010           1  3.9    10
[4] {root vegetables,                                                 
     whipped/sour cream,                                              
     flour}              => {whole milk} 0.0017           1  3.9    17
[5] {butter,                                                          
     soft cheese,                                                     
     domestic eggs}      => {whole milk} 0.0010           1  3.9    10
```

Figure 5: Targeting Item using apriori function

Likewise, we set the left hand side to be whole milk and find its antecedents. When doing this we set the confidence to 0.15 amd we set a minimum length of 2 to avoid empty left hand side items. Then the ouput was like Figure 6.



```
> inspect(rules[1:5])
    lhs             rhs                support confidence lift count
[1] {whole milk} => {other vegetables} 0.075   0.29       1.5  736
[2] {whole milk} => {rolls/buns}       0.057   0.22       1.2  557
[3] {whole milk} => {yogurt}           0.056   0.22       1.6  551
[4] {whole milk} => {root vegetables}  0.049   0.19       1.8  481
[5] {whole milk} => {tropical fruit}   0.042   0.17       1.6  416
```

Figure 6: Targeting Left Hand Side Results by using apriori function

The last step was to visualize the results. We want to map out the rules in a graph wo we hat used another library arulesViz. Figure 7 shows graph for results.

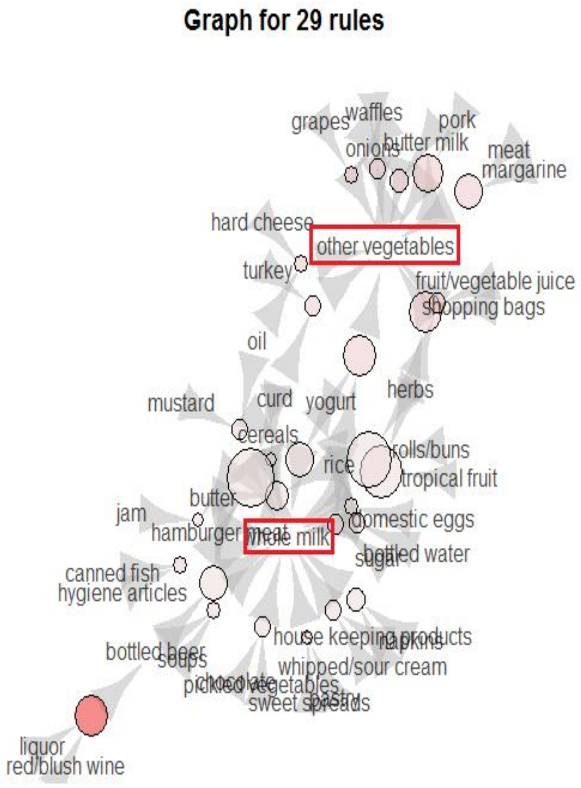


Figure 7: Graph for Rules

To illustrate the graph, the graphs directed to the middle milk will show the set of items to be purchased before purchasing the milk that was the first target. In the case of other vegetables where the arrows are facing a lot, it becomes the item that most customers buy whole milk.

5. Conclusion

In this paper, we researched about R association rules with market basket dataset. We found the relationship between things depending on what customers buy. By using R, we are able to analyze the product easily in many ways. Specifically, this analysis of what items were purchased before and after the purchase of 'whole milk'. For doing this, it will be helpful for market operation for raising sales to observe the needs of customers.

Since we are aware of analyzing association rules based on big data, we will be able to find out the other relational things beyond the market basket which will be helpful for many other fields.

ACKNOWLEDGEMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(2017R1D1A1B03034691).

REFERENCES

1. Team, R. Core. "R language definition." Vienna, Austria: R foundation for statistical computing (2000).
2. Witten, Ian H., et al. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016.A. Cichocki and R. Unbehaven. Neural Networks for Optimization and Signal Processing, 1st ed. Chichester, U.K.: Wiley, 1993, ch. 2, pp. 45-47.
3. Low, Yucheng, et al. "Distributed GraphLab: a framework for machine learning and data mining in the cloud." Proceedings of the VLDB Endowment 5.8 (2012): 716-727. https://doi.org/10.14778/2212351.2212354
4. Jochen, Hipp Ulrich Güntzer, and Gholamreza Nakhaeizadeh. "Algorithms for association rule mining—a general survey and comparison." ACM sigkdd explorations newsletter 2.1 (2000): 58-64.
5. Yanbin, Ye and Chia-Chu Chiang. "A parallel apriori algorithm for frequent itemsets mining." Software Engineering Research, Management and Applications, 2006. Fourth International Conference on. IEEE, 2006. https://doi.org/10.1109/SERA.2006.6