8_Association_Rule_Mining

2023-08-14

```
groceries = read.csv('~/STA380-Exercise/datafile/groceries.txt', header = FALSE)
```

```
library(arules)
```

```
## Loading required package: Matrix

##
## Attaching package: 'arules'

## The following objects are masked from 'package:base':
##
##     abbreviate, write
```

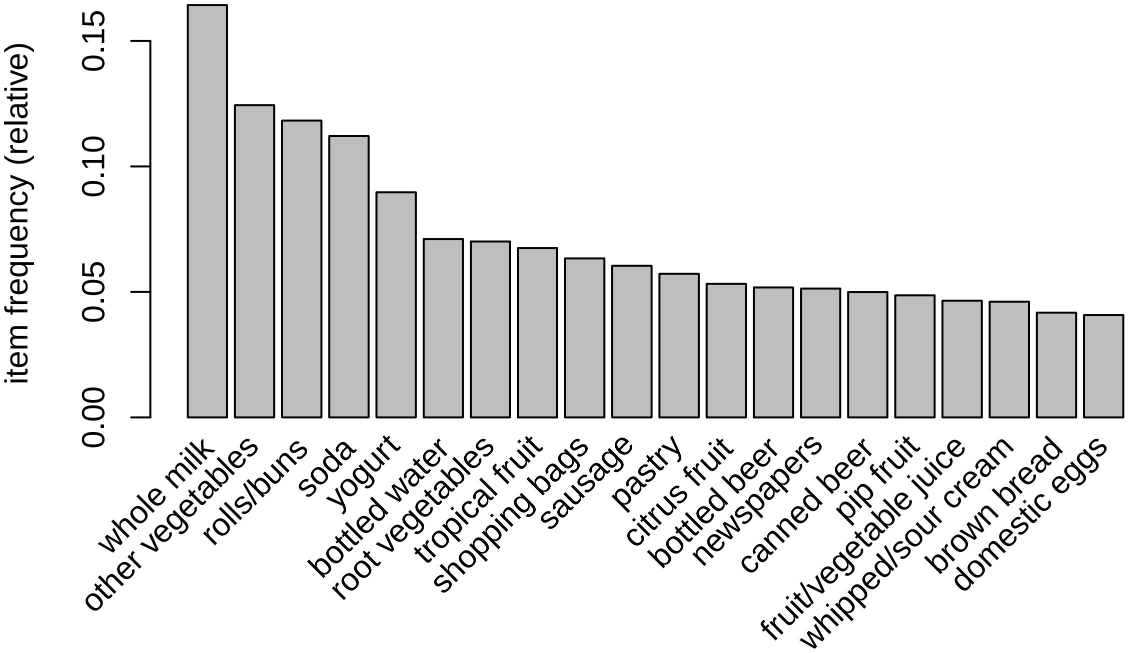
```
# Replace empty cells with NA
groceries[groceries == ""] <- NA

# Create baskets by removing NA values
baskets <- lapply(1:nrow(groceries), function(row) {
  basket <- unlist(groceries[row, ])
  basket <- basket[!is.na(basket)]
  basket
})

# Convert baskets to transactions
gtrans <- as(baskets, "transactions")

# Display summary of transactions
summary(gtrans)
```

```
## transactions as itemMatrix in sparse format with
##  15296 rows (elements/itemsets/transactions) and
##  169 columns (items) and a density of 0.01677625
##
## most frequent items:
##       whole milk other vegetables       rolls/buns             soda
##             2513             1903             1809             1715
##           yogurt          (Other)
##             1372            34055
##
## element (itemset/transaction) length distribution:
## sizes
##    1    2    3    4
## 3485 2630 2102 7079
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   2.835   4.000   4.000
##
```

```
## includes extended item information - examples:
##             labels
## 1 abrasive cleaner
## 2 artif. sweetener
## 3   baby cosmetics
```

The data consists of 15,296 shopping baskets (transactions) and 169 unique items (columns). The density of the data indicates that, on average, each transaction contains approximately 1.68% of the total available items.

```
# Item frequency plot
itemFrequencyPlot(
  gtrans, topN = 20,
  type = 'relative',
)
```



Frequent Items: Among the most frequent items, we find that “whole milk,” “other vegetables,” “rolls/buns,” “soda,” and “yogurt” dominate the baskets. These items are commonly purchased by the customers.

Transaction Length Distribution: The distribution of transaction lengths provides an overview of the number of items per basket. It appears that:

3,485 baskets contain 1 item, 2,630 baskets contain 2 items 2,102 baskets contain 3 items, 7,079 baskets contain 4 items Item

Labels: We also have additional information about some of the items. For instance, we can see examples of items like “abrasive cleaner,” “artificial sweetener,” and “baby cosmetics.”

Finding interesting association rules

```
# Perform association rule mining using Apriori algorithm
grocr <- apriori(gtrans,
                 parameter = list(support = 0.005, confidence = 0.1, maxlen = 4))
```

```
## Apriori
##
```

```
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##         0.1    0.1    1 none FALSE            TRUE       5   0.005      1
##  maxlen target  ext
##       4  rules TRUE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 76
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[169 item(s), 15296 transaction(s)] done [0.00s].
## sorting and recoding items ... [101 item(s)] done [0.00s].
## creating transaction tree ... done [0.01s].
## checking subsets of size 1 2 3 done [0.00s].
## writing ... [118 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```

```
# Display summary of the mined rules
summary(grocr)
```

```
## set of 118 rules
##
## rule length distribution (lhs + rhs):sizes
##   1   2   3
##   4 108   6
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   2.000   2.017   2.000   3.000
##
## summary of quality measures:
##     support           confidence        coverage            lift
##  Min.   :0.005034   Min.   :0.1006   Min.   :0.01589   Min.   :0.902
##  1st Qu.:0.006636   1st Qu.:0.1277   1st Qu.:0.03707   1st Qu.:1.428
##  Median :0.008270   Median :0.1757   Median :0.05250   Median :1.821
##  Mean   :0.014550   Mean   :0.1912   Mean   :0.09165   Mean   :1.918
##  3rd Qu.:0.012667   3rd Qu.:0.2408   3rd Qu.:0.07008   3rd Qu.:2.323
##  Max.   :0.164291   Max.   :0.4037   Max.   :1.00000   Max.   :3.865
##      count
##  Min.   :  77.0
##  1st Qu.: 101.5
##  Median : 126.5
##  Mean   : 222.6
##  3rd Qu.: 193.8
##  Max.   :2513.0
##
## mining info:
##    data ntransactions support confidence
##  gtrans         15296   0.005        0.1
##                                                                                      call
##  apriori(data = gtrans, parameter = list(support = 0.005, confidence = 0.1, maxlen = 4))
```

Rule Characteristics: The Apriori algorithm discovered a total of 118 association rules, characterized by varying lengths of items on both the left-hand side (lhs) and right-hand side (rhs) of the rules. These rules provide insights into how items are related in the context of grocery shopping baskets.

Support: The minimum support for an item set is approximately 0.5%, indicating that a rule is relevant if it appears in at least this proportion of transactions.

Confidence: The minimum confidence threshold is set at 10%, meaning that a rule should have a confidence level of at least this value to be considered significant.

Coverage: The coverage reflects the proportion of transactions covered by the rule.

Lift: The lift measures the strength of the association between the items in the rule. A lift greater than 1.0 indicates a positive association.

```
# Filter rules based on lift threshold
high_lift_rules <- subset(grocr, subset = lift > 2.5)

# Display and interpret the high lift rules
inspect(high_lift_rules)
```

```
##      lhs                                rhs                support
## [1]  {onions}                        => {root vegetables}  0.005295502
## [2]  {onions}                        => {other vegetables} 0.007452929
## [3]  {beef}                          => {citrus fruit}     0.005099372
## [4]  {beef}                          => {root vegetables}  0.008695084
## [5]  {root vegetables}               => {beef}             0.008695084
## [6]  {pork}                          => {root vegetables}  0.006733787
## [7]  {frankfurter}                   => {sausage}          0.006472280
## [8]  {sausage}                       => {frankfurter}      0.006472280
## [9]  {pip fruit}                     => {citrus fruit}     0.008172071
## [10] {citrus fruit}                  => {pip fruit}        0.008172071
## [11] {pip fruit}                     => {tropical fruit}   0.012683054
## [12] {tropical fruit}                => {pip fruit}        0.012683054
## [13] {citrus fruit}                  => {tropical fruit}   0.012486925
## [14] {tropical fruit}                => {citrus fruit}     0.012486925
## [15] {root vegetables}               => {other vegetables} 0.025366109
## [16] {other vegetables}              => {root vegetables}  0.025366109
## [17] {root vegetables, whole milk}  => {other vegetables} 0.008172071
## [18] {other vegetables, whole milk} => {root vegetables}  0.008172071
##      confidence coverage   lift     count
## [1]  0.2655738  0.01993985 3.789381  81
## [2]  0.3737705  0.01993985 3.004306 114
## [3]  0.1511628  0.03373431 2.840523  78
## [4]  0.2577519  0.03373431 3.677774 133
## [5]  0.1240672  0.07008368 3.677774 133
## [6]  0.1816578  0.03706851 2.592013 103
## [7]  0.1706897  0.03791841 2.825616  99
## [8]  0.1071429  0.06040795 2.825616  99
## [9]  0.1680108  0.04864017 3.157116 125
## [10] 0.1535627  0.05321653 3.157116 125
## [11] 0.2607527  0.04864017 3.864800 194
## [12] 0.1879845  0.06746862 3.864800 194
## [13] 0.2346437  0.05321653 3.477820 191
## [14] 0.1850775  0.06746862 3.477820 191
## [15] 0.3619403  0.07008368 2.909216 388
## [16] 0.2038886  0.12441161 2.909216 388
## [17] 0.3612717  0.02262029 2.903842 125
## [18] 0.2000000  0.04086036 2.853731 125
```

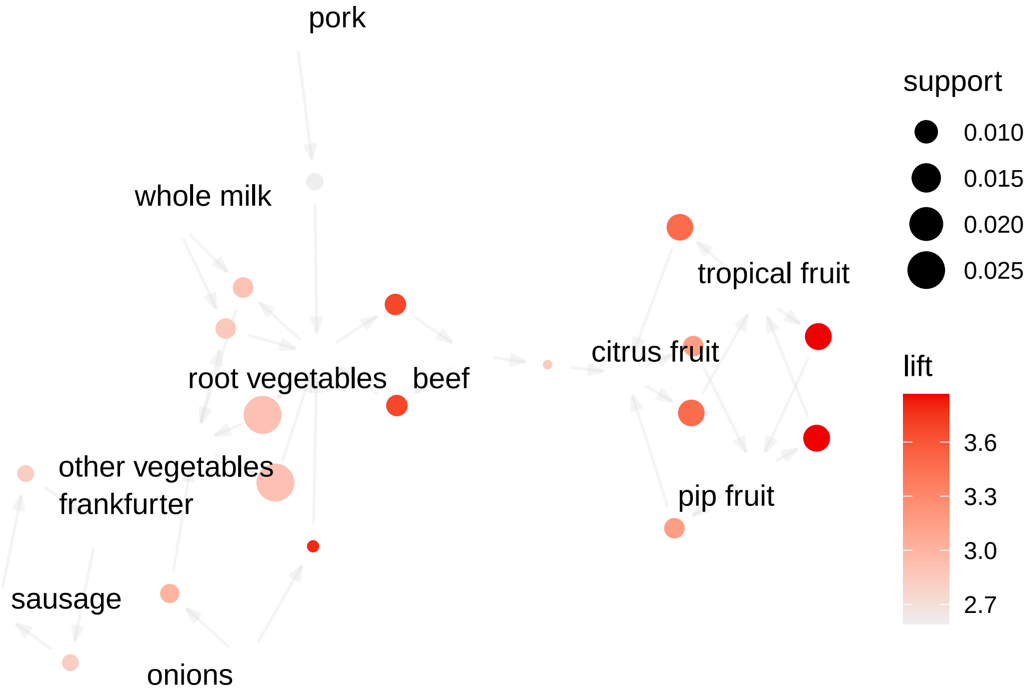
These high lift association rules provide a glimpse into intriguing purchasing patterns. Connection between onions and root vegetables suggests that customers who buy onions are likely to purchase other root vegetables and other vegetables as well. Beef buyers also show an interest in citrus fruit and root vegetables.

Picking Thresholds: I set a minimum confidence threshold of 10% and a minimum lift threshold of 2.5. Confidence Threshold: The confidence threshold of 10% was chosen to identify rules where the consequent (rhs) appears in at least 10% of transactions containing the antecedent (lhs). Lift Threshold: A lift value of 2.5 implies that the items are 2.5 times more likely to be bought together than expected by chance.

```
# Load necessary packages for visualization
library(arulesViz)

# Visualize the high lift rules, including weaker associations
plot(high_lift_rules, method = "graph", control = list(type = "items", shading = "lift"))
```

```
## Warning: Unknown control parameters: type, shading
```

```
## Available control parameters (with default values):
## layout    =  stress
## circular  =  FALSE
## ggraphdots    =  NULL
## edges     =  <environment>
## nodes     =  <environment>
## nodetext  =  <environment>
## colors    =  c("#EE0000FF", "#EEEEEEFF")
## engine    =  ggplot2
## max   =  100
## verbose   =  FALSE
```



In the graph, each item is represented as a node (circle). These nodes are interconnected by edges (lines), illustrating the linkages between items based on the high lift association rules.

The direction of the arrows on the edges indicates the direction of association between items. For instance, customers who purchase whole milk are inclined to also buy vegetables. On the right-hand side, distinct clusters of fruits emerge, categorized as tropical, citrus, and pip. This alignment is logical, as these fruits tend to be grouped together and are often selected by shoppers as complementary choices. Sausages, onions, and various vegetables are in proximity, indicating a likelihood of concurrent purchase. This proximity implies a shared association, since it is common to include these items in dinner preparations.