HW04_Chenxin

2024-04-08

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.5.0     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(igraph)
```

```
##
## Attaching package: 'igraph'
##
## The following objects are masked from 'package:lubridate':
##
##     %--%, union
##
## The following objects are masked from 'package:dplyr':
##
##     as_data_frame, groups, union
##
## The following objects are masked from 'package:purrr':
##
##     compose, simplify
##
## The following object is masked from 'package:tidyr':
##
##     crossing
##
## The following object is masked from 'package:tibble':
##
##     as_data_frame
##
## The following objects are masked from 'package:stats':
##
##     decompose, spectrum
##
## The following object is masked from 'package:base':
##
##     union
```

```r
library(arules)  # has a big ecosystem of packages built around it
```

```
## Loading required package: Matrix
##
## Attaching package: 'Matrix'
##
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
##
##
## Attaching package: 'arules'
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
## The following objects are masked from 'package:base':
##
##     abbreviate, write
```

```r
library(arulesViz)
# (1) Load and Process the Data:
library(arules)
groceries = read.transactions('/Users/vita/Desktop/HW04/groceries.txt')
```

```
## Warning in asMethod(object): removing duplicated items in transactions
```

```r
groceries_list = readLines('/Users/vita/Desktop/HW04/groceries.txt')
groceries_list = strsplit(groceries_list, ",")

# Remove duplicates ("de-dupe")
groceries_list_unique = lapply(groceries_list, unique)

# (2) Convert list to transactions
groceries_transactions <- as(groceries_list_unique, "transactions")

# Convert the cleaned list to transactions
# Cast this variable as a special arules "transactions" class.
```

```r
# (3) Apply the Apriori Algorithm
rules = apriori(groceries_transactions,
                parameter = list(support = 0.005, confidence = 0.1, minlen = 4))
```

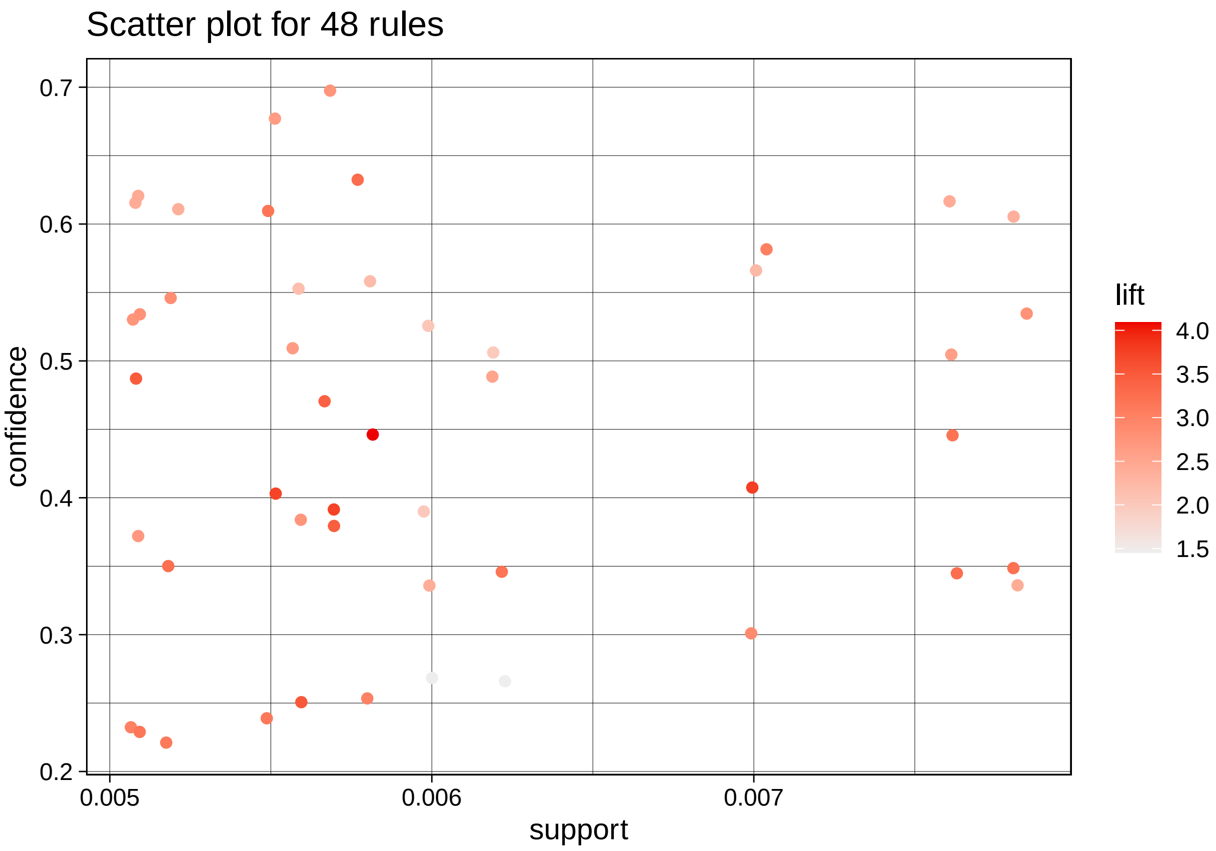
```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##         0.1    0.1    1 none FALSE            TRUE       5   0.005      4
##  maxlen target  ext
##      10  rules TRUE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
```

```
##
## Absolute minimum support count: 49
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[169 item(s), 9835 transaction(s)] done [0.00s].
## sorting and recoding items ... [120 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 3 4 done [0.00s].
## writing ... [48 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```

```
# Look at rules with support > .005 & confidence >.1 & length (#) <= 4

# Thresholds for lift and confidence: A support threshold of 0.005 means we're interested in itemsets t

# A confidence threshold of 0.2 is chosen to ensure that at least 20% of the time, the items on the lef

# Setting minlen to 2 ensures that the rules consist of at least two items. This is the smallest possib
```

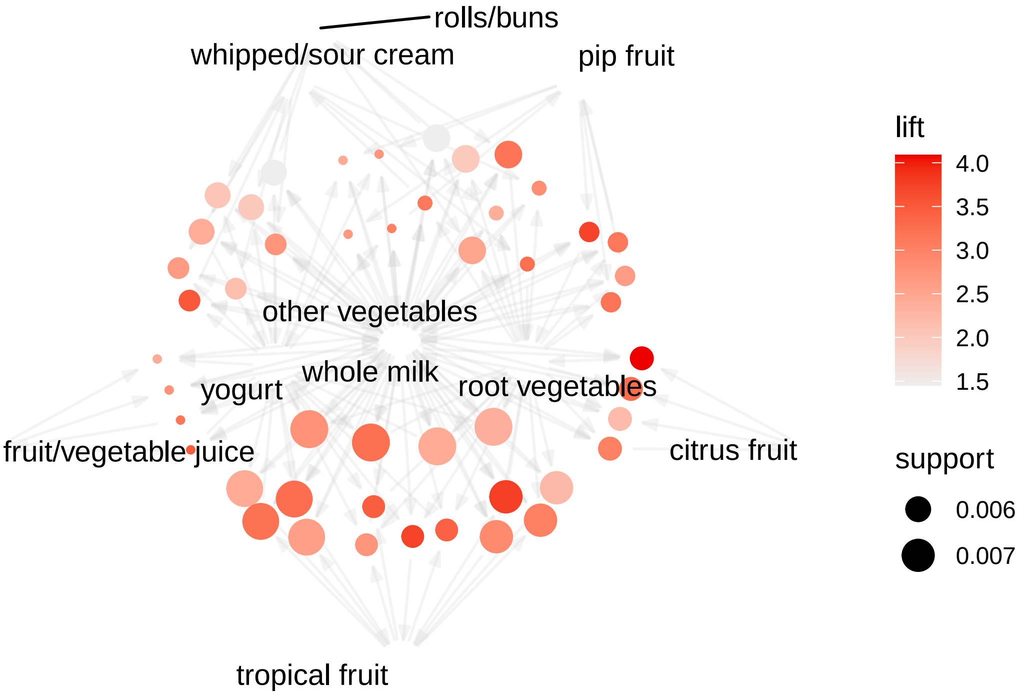
```
# Analyze and Visualize the Results
# Basic plot of rules
plot(rules)
```

```
## To reduce overplotting, jitter is added! Use jitter = 0 to prevent jitter.
```



```
plot(rules, method = "graph", control = list(type = "items"))
```

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



```r
# For more detailed exploration, inspect rules
inspect(head(sort(rules, by = "lift"), 10))
```

```
##      lhs                            rhs                     support confidence    coverage     lift cou
## [1]  {citrus fruit,
##       other vegetables,
##       whole milk}                => {root vegetables}    0.005795628  0.4453125 0.013014743 4.085493
## [2]  {other vegetables,
##       tropical fruit,
##       whole milk}                => {root vegetables}    0.007015760  0.4107143 0.017081851 3.768074
## [3]  {root vegetables,
##       whole milk,
##       yogurt}                    => {tropical fruit}     0.005693950  0.3916084 0.014539908 3.732043
## [4]  {other vegetables,
##       pip fruit,
##       whole milk}                => {root vegetables}    0.005490595  0.4060150 0.013523132 3.724961
## [5]  {other vegetables,
##       whole milk,
##       yogurt}                    => {whipped/sour cream} 0.005592272  0.2511416 0.022267412 3.503514
## [6]  {fruit/vegetable juice,
##       other vegetables,
##       whole milk}             => {yogurt}            0.005083884  0.4854369 0.010472801 3.479790
## [7]  {tropical fruit,
##       whole milk,
##       yogurt}                 => {root vegetables}    0.005693950  0.3758389 0.015149975 3.448112
## [8]  {root vegetables,
##       tropical fruit,
##       whole milk}             => {yogurt}            0.005693950  0.4745763 0.011997966 3.401937
## [9]  {citrus fruit,
##       root vegetables,
##       whole milk}             => {other vegetables}   0.005795628  0.6333333 0.009150991 3.273165
## [10] {other vegetables,
##       whole milk,
##       yogurt}                 => {tropical fruit}     0.007625826  0.3424658 0.022267412 3.263712
```

```
# Support and Confidence Levels:
# Graph 01: The scatter plot indicates that most rules have a support between 0.005 and 0.0075. This is
# Graph 02: The graph visualization clusters items like 'whole milk', 'yogurt', 'other vegetables', 'ro
```