Association Rule Mining

Nimish Amlathe

15 August 2016

Association rule mining

Read in the grocery data.

- First creating a list of baskets: vectors of items by consumer.
- Analagous to bags of words.
- Apriori algorithm expects a list of baskets in a special format.
- Removing duplicates and then Casting this variable as a arules "transactions" class.

```
library(arules)
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'arules'
```

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

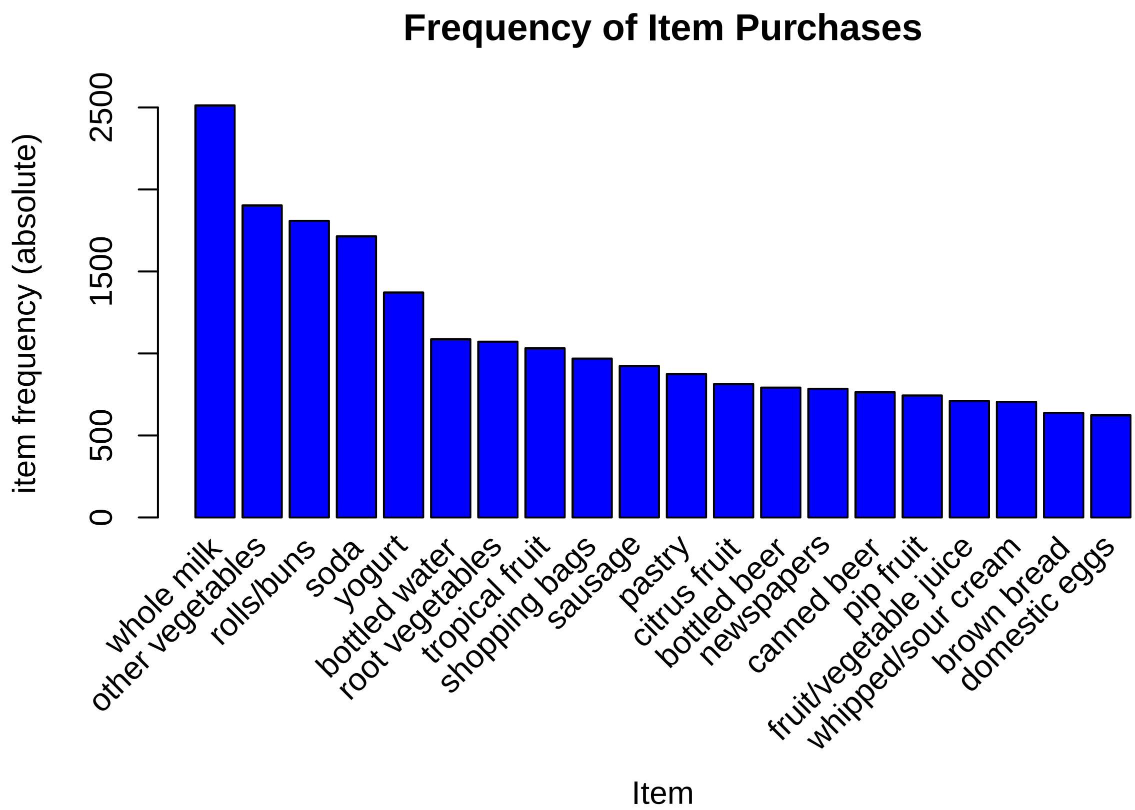
```
library(arulesViz)
```

```
## Loading required package: grid
```

```
groceries = read.transactions(file = "files/groceries.txt", rm.duplicates = TRUE, format = "basket", sep
```

Let's plot an Item-frequency chart to guage how much more often certain items are present in the dataset.

```
# Plotting top 20 items by frequency
itemFrequencyPlot(groceries,topN=20,type = "absolute", col = 'blue', xlab = 'Item', main = 'Frequency of
```



Running the 'apriori' algorithm.

Looking at rules with support > .01 & confidence >.5 & length #items <= 6

```
grocrules <- apriori(groceries, parameter=list(support=.01, confidence=.5, maxlen=6))
```

```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport support minlen maxlen
##         0.5    0.1    1 none FALSE            TRUE    0.01      1      6
##  target   ext
##   rules FALSE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 98
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[169 item(s), 9835 transaction(s)] done [0.00s].
## sorting and recoding items ... [88 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 3 4 done [0.00s].
## writing ... [15 rule(s)] done [0.00s].
```

```
## creating S4 object  ... done [0.07s].
```

```r
inspect(grocrules)
```

```
##    lhs                        rhs                    support confidence     lift
## 1  {curd,
##     yogurt}               => {whole milk}       0.01006609  0.5823529 2.279125
## 2  {butter,
##     other vegetables}     => {whole milk}       0.01148958  0.5736041 2.244885
## 3  {domestic eggs,
##     other vegetables}     => {whole milk}       0.01230300  0.5525114 2.162336
## 4  {whipped/sour cream,
##     yogurt}               => {whole milk}       0.01087951  0.5245098 2.052747
## 5  {other vegetables,
##     whipped/sour cream} => {whole milk}       0.01464159  0.5070423 1.984385
## 6  {other vegetables,
##     pip fruit}            => {whole milk}       0.01352313  0.5175097 2.025351
## 7  {citrus fruit,
##     root vegetables}      => {other vegetables} 0.01037112  0.5862069 3.029608
## 8  {root vegetables,
##     tropical fruit}       => {other vegetables} 0.01230300  0.5845411 3.020999
## 9  {root vegetables,
##     tropical fruit}       => {whole milk}       0.01199797  0.5700483 2.230969
## 10 {tropical fruit,
##     yogurt}               => {whole milk}       0.01514997  0.5173611 2.024770
## 11 {root vegetables,
##     yogurt}               => {other vegetables} 0.01291307  0.5000000 2.584078
## 12 {root vegetables,
##     yogurt}               => {whole milk}       0.01453991  0.5629921 2.203354
## 13 {rolls/buns,
##     root vegetables}      => {other vegetables} 0.01220132  0.5020921 2.594890
## 14 {rolls/buns,
##     root vegetables}      => {whole milk}       0.01270971  0.5230126 2.046888
## 15 {other vegetables,
##     yogurt}               => {whole milk}       0.02226741  0.5128806 2.007235
```

Trying various subsets

```r
## Trying various subsets
inspect(subset(grocrules, subset = lift > 3))
```

```
##   lhs                   rhs                    support confidence     lift
## 1 {citrus fruit,
##    root vegetables} => {other vegetables} 0.01037112  0.5862069 3.029608
## 2 {root vegetables,
##    tropical fruit}  => {other vegetables} 0.01230300  0.5845411 3.020999
```

```r
inspect(subset(grocrules, subset = confidence > 0.5))
```

```
##    lhs                        rhs                    support confidence     lift
## 1  {curd,
##     yogurt}               => {whole milk}       0.01006609  0.5823529 2.279125
## 2  {butter,
##     other vegetables}     => {whole milk}       0.01148958  0.5736041 2.244885
## 3  {domestic eggs,
##     other vegetables}     => {whole milk}       0.01230300  0.5525114 2.162336
```

```
## 4  {whipped/sour cream,
##     yogurt}            => {whole milk}       0.01087951  0.5245098 2.052747
## 5  {other vegetables,
##     whipped/sour cream} => {whole milk}      0.01464159  0.5070423 1.984385
## 6  {other vegetables,
##     pip fruit}         => {whole milk}       0.01352313  0.5175097 2.025351
## 7  {citrus fruit,
##     root vegetables}   => {other vegetables} 0.01037112  0.5862069 3.029608
## 8  {root vegetables,
##     tropical fruit}    => {other vegetables} 0.01230300  0.5845411 3.020999
## 9  {root vegetables,
##     tropical fruit}    => {whole milk}       0.01199797  0.5700483 2.230969
## 10 {tropical fruit,
##     yogurt}            => {whole milk}       0.01514997  0.5173611 2.024770
## 11 {root vegetables,
##     yogurt}            => {whole milk}       0.01453991  0.5629921 2.203354
## 12 {rolls/buns,
##     root vegetables}   => {other vegetables} 0.01220132  0.5020921 2.594890
## 13 {rolls/buns,
##     root vegetables}   => {whole milk}       0.01270971  0.5230126 2.046888
## 14 {other vegetables,
##     yogurt}            => {whole milk}       0.02226741  0.5128806 2.007235
```

```r
inspect(subset(grocrules, subset = support > .01 & confidence > 0.3))
```

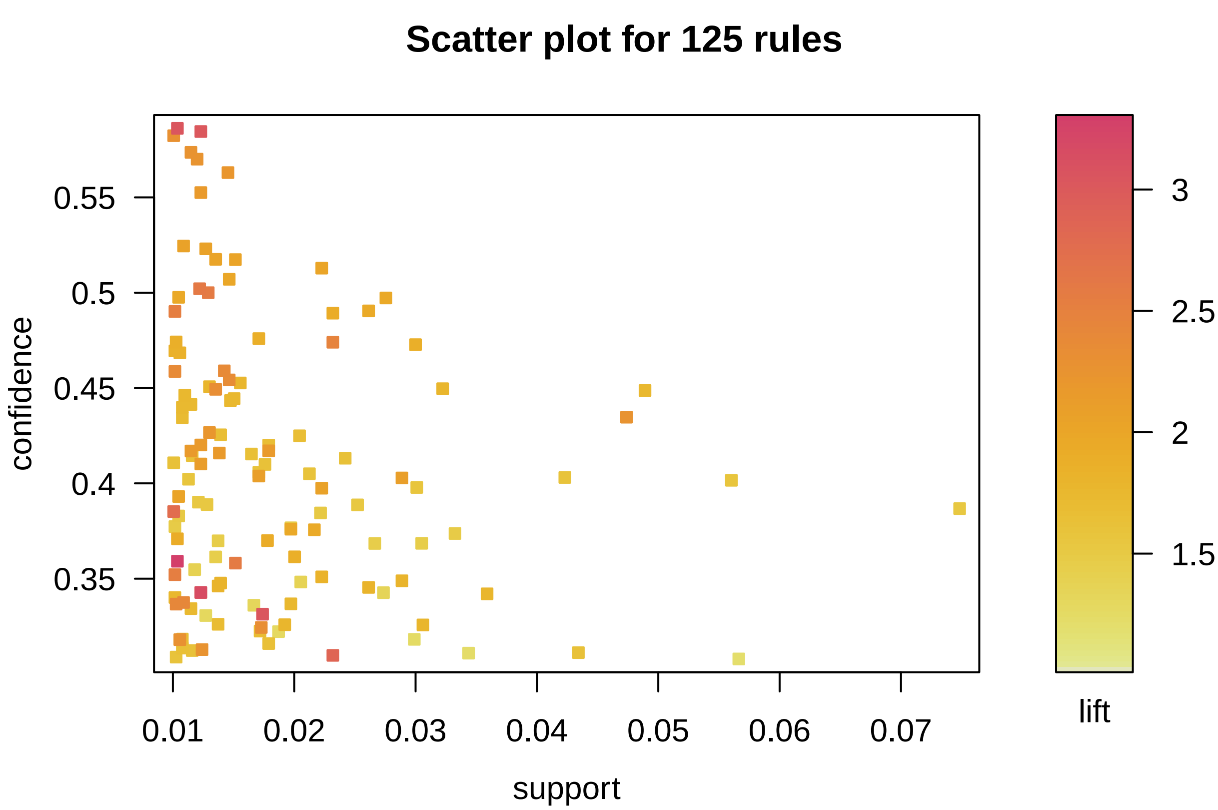
```
##    lhs                      rhs                   support confidence     lift
## 1  {curd,
##     yogurt}            => {whole milk}       0.01006609  0.5823529 2.279125
## 2  {butter,
##     other vegetables}  => {whole milk}       0.01148958  0.5736041 2.244885
## 3  {domestic eggs,
##     other vegetables}  => {whole milk}       0.01230300  0.5525114 2.162336
## 4  {whipped/sour cream,
##     yogurt}            => {whole milk}       0.01087951  0.5245098 2.052747
## 5  {other vegetables,
##     whipped/sour cream} => {whole milk}      0.01464159  0.5070423 1.984385
## 6  {other vegetables,
##     pip fruit}         => {whole milk}       0.01352313  0.5175097 2.025351
## 7  {citrus fruit,
##     root vegetables}   => {other vegetables} 0.01037112  0.5862069 3.029608
## 8  {root vegetables,
##     tropical fruit}    => {other vegetables} 0.01230300  0.5845411 3.020999
## 9  {root vegetables,
##     tropical fruit}    => {whole milk}       0.01199797  0.5700483 2.230969
## 10 {tropical fruit,
##     yogurt}            => {whole milk}       0.01514997  0.5173611 2.024770
## 11 {root vegetables,
##     yogurt}            => {other vegetables} 0.01291307  0.5000000 2.584078
## 12 {root vegetables,
##     yogurt}            => {whole milk}       0.01453991  0.5629921 2.203354
## 13 {rolls/buns,
##     root vegetables}   => {other vegetables} 0.01220132  0.5020921 2.594890
## 14 {rolls/buns,
##     root vegetables}   => {whole milk}       0.01270971  0.5230126 2.046888
## 15 {other vegetables,
```

```
##     yogurt}              => {whole milk}       0.02226741  0.5128806 2.007235
```

```r
rules = apriori(groceries, parameter = list(support=.01, confidence=.3, target='rules'))
```

```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport support minlen maxlen
##         0.3    0.1    1 none FALSE            TRUE    0.01      1     10
##  target   ext
##   rules FALSE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 98
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[169 item(s), 9835 transaction(s)] done [0.00s].
## sorting and recoding items ... [88 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 3 4 done [0.00s].
## writing ... [125 rule(s)] done [0.00s].
## creating S4 object  ... done [0.06s].
```

```r
plot(rules)
```



Conclusion

Whole milk, other vegetables and yogurt are some of the most likely to be purchased items based on various itemsets. These are also amongst the items with the highest support counts.

The various itemsets we have seen so far point to associations between people who buy a certain kind of items also buying some of the more frequently occuring items. For example, people who buy a lot of diary products tend to also buy milk, and people how buy a lot of fruits and vegetables also tend to biuy milk and other vegetables, etc. We did not see any rare patterns or patterns amongst itemsets with very low support (some niche products, etc.)