Lab 6 - Association Rule Mining Assignments

2020/12/06

```
library(arules)
```

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

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

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

Các câu hỏi dưới đây liên quan đến file `"browsing.txt"`. Đây là dữ liệu về các trang mà khách hàng truy cập trong mỗi phiên giao dịch của một cửa hàng bán lẻ online. Để đơn giản hóa cho bài tập này, dữ liệu đã được tiền xử lý để mỗi dòng là một phiên giao dịch gồm id của các trang (cách nhau bởi khoảng trắng).

Câu 1.

Load file `"browsing.txt"` lưu vào biến `browsing`. Cho biết các thông tin tổng quan về dữ liệu: dữ liệu có bao nhiêu phiên giao dịch (dòng), bao nhiêu trang (cột)? Liệt kê 5 trang được truy cập nhiều nhất. Bao nhiêu phiên giao dịch truy cập ít trang nhất, số lượng ít nhất là bao nhiêu? Bao nhiêu phiên giao dịch truy cập nhiều trang nhất, số lượng nhiều nhất là bao nhiêu? Trung bình có bao nhiêu trang được truy cập trên mỗi phiên giao dịch?

```
browsing <- read.transactions("/home/giangvdq/workspaces/documentation-for-noobs/DataMining-QuachDinhHo
data(browsing)
```

```
## Warning in data(browsing): data set 'browsing' not found
```

```
summary(browsing)
```

```
## transactions as itemMatrix in sparse format with
##  31101 rows (elements/itemsets/transactions) and
##  30697 columns (items) and a density of 3.257647e-05
##
## most frequent items:
##                                                DAI62779 SNA53220 FRO19221 SNA93860
##                                                                                 72
##                                                ELE59028 DAI62779 FRO85978 SNA93860
##                                                                                 14
## ELE66810 ELE65859 GRO54782 DAI34002 FRO92511 DAI54444 ELE30911 ELE88583 SNA24799
##                                                                                 12
##                                       ELE66810 GRO43063 DAI86157 SNA55617 SNA24799
##                                                                                 12
##                                       DAI62779 SNA53220 FRO19221 ELE69552 SNA93860
##                                                                                 11
##                                                                           (Other)
##                                                                             30980
##
```

```
## element (itemset/transaction) length distribution:
## sizes
##     1
## 31101
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       1       1       1       1       1       1
##
## includes extended item information - examples:
##
## 1                                                                                                  DAI11223
## 2 DAI11238 SNA82274 SNA96466 GRO88324 SNA43409 FRO35729 GRO83463 GRO30912 ELE34234 ELE26753 ELE45560
## 3 DAI11290 DAI37288 ELE55848 ELE32164 DAI43747 GRO17794 DAI43223 ELE20196 SNA26019 ELE62598 SNA42528
```

```
inspect(head(browsing))
```

```
##     items
## [1] {FRO11987 ELE17451 ELE89019 SNA90258 GRO99222}
## [2] {GRO99222 GRO12298 FRO12685 ELE91550 SNA11465 ELE26917 ELE52966 FRO90334 SNA30755 ELE17451 FRO841
## [3] {ELE17451 GRO73461 DAI22896 SNA99873 FRO86643}
## [4] {ELE17451 ELE37798 FRO86643 GRO56989 ELE23393 SNA11465}
## [5] {ELE17451 SNA69641 FRO86643 FRO78087 SNA11465 GRO39357 ELE28573 ELE11375 DAI54444}
## [6] {ELE17451 GRO73461 DAI22896 SNA99873 FRO18919 DAI50921 SNA80192 GRO75578}
```

```
itemFrequency(browsing[, 1:5])
```

```
##                                                                                                DAI11223 EL
##
## DAI11238 SNA82274 SNA96466 GRO88324 SNA43409 FRO35729 GRO83463 GRO30912 ELE34234 ELE26753 ELE45560 EL
##
## DAI11290 DAI37288 ELE55848 ELE32164 DAI43747 GRO17794 DAI43223 ELE20196 SNA26019 ELE62598 SNA42528 DA
##
##                                                                                    DAI11290 DAI55148 DA
##
##                                                                                                      DA
##
```

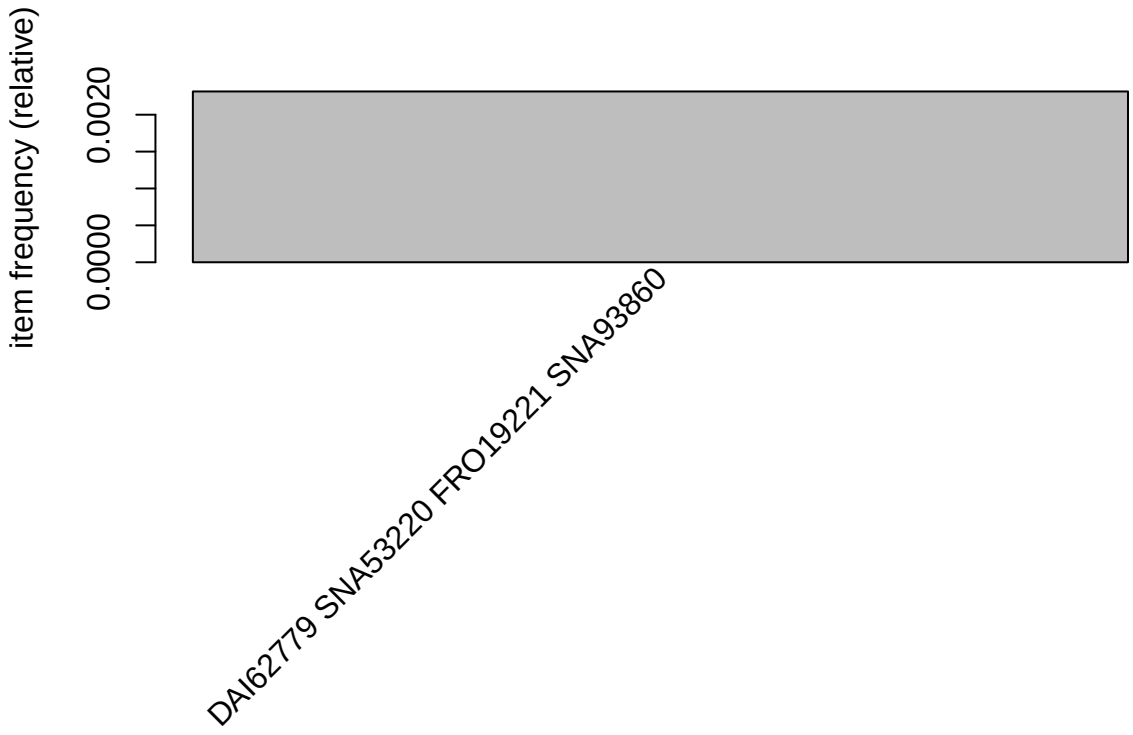
Câu 2.

Chọn một giá trị support mà bạn cho là phù hợp với dữ liệu. Giải thích ngắn gọn lý do bạn chọn giá trị đó. Vẽ biểu đồ tần số của các trang trong các phiên giao dịch với support bạn chọn. Vẽ biểu đồ tần số của top 20 trang được truy cập nhiều nhất trong các phiên giao dịch.

```
itemFrequency(browsing[, 1:5])
```

```
##                                                                                                DAI11223 EL
##
## DAI11238 SNA82274 SNA96466 GRO88324 SNA43409 FRO35729 GRO83463 GRO30912 ELE34234 ELE26753 ELE45560 EL
##
## DAI11290 DAI37288 ELE55848 ELE32164 DAI43747 GRO17794 DAI43223 ELE20196 SNA26019 ELE62598 SNA42528 DA
##
##                                                                                    DAI11290 DAI55148 DA
##
##                                                                                                      DA
##
```

```r
itemFrequencyPlot(browsing, topN = 1)
```



Câu 3.

Có bao nhiêu frequent itemset? Bao nhiêu frequent itemset có số item (trang) ít nhất? Bao nhiêu frequent itemset có số item là nhiều nhất? Bao nhiêu frequent itemset có ít nhất k item trở lên (thử với k = 2, 3, ...)? Xem top 10 itemset sắp xếp theo support. Xem top 10 k-itemset sắp xếp theo support (thử với k = 1, 2, 3, ...). Vẽ biểu đồ tần số theo bậc (số lượng item) của các frequent itemset.

```r
is.freq <- apriori(browsing, parameter = list(target = "frequent itemsets"))
```

```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##          NA    0.1    1 none FALSE            TRUE       5     0.1      1
##  maxlen            target  ext
##      10 frequent itemsets TRUE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 3110
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[30697 item(s), 31101 transaction(s)] done [0.03s].
## sorting and recoding items ... [0 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 done [0.00s].
## sorting transactions ... done [0.00s].
```

```
## writing ... [0 set(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```

```
is.freq
```

```
## set of 0 itemsets
```

```
is.freq <- apriori(browsing, parameter = list(target = "frequent itemsets", support = 0.05))
```

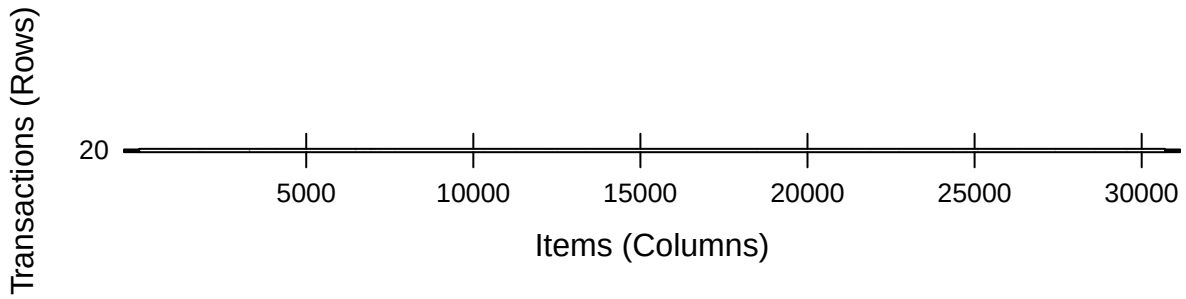
```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##          NA    0.1    1 none FALSE            TRUE       5    0.05      1
##  maxlen            target  ext
##      10 frequent itemsets TRUE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 1555
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[30697 item(s), 31101 transaction(s)] done [0.03s].
## sorting and recoding items ... [0 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 done [0.00s].
## sorting transactions ... done [0.00s].
## writing ... [0 set(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```

```
is.freq
```

```
## set of 0 itemsets
```

```
is.freq <- sort(is.freq, by = "support")
inspect(head(is.freq, n = 10))
```

```
image(browsing[1:100])
```




```
inspect(is.freq)
```

Câu 4.

Có bao nhiêu closed intemset? Bao nhiêu closed itemset có số item là ít nhất? Bao nhiêu closed itemset có số item là nhiều nhất? Bao nhiêu closed itemset có ít nhất k item trở lên (thử với k = 2, 3, ...)? Xem top 10

closed k-itemset sắp xếp theo support (thử với k = 1, 2, 3, …). Vẽ biểu đồ tần số theo bậc (số lượng item) của các closed itemset.

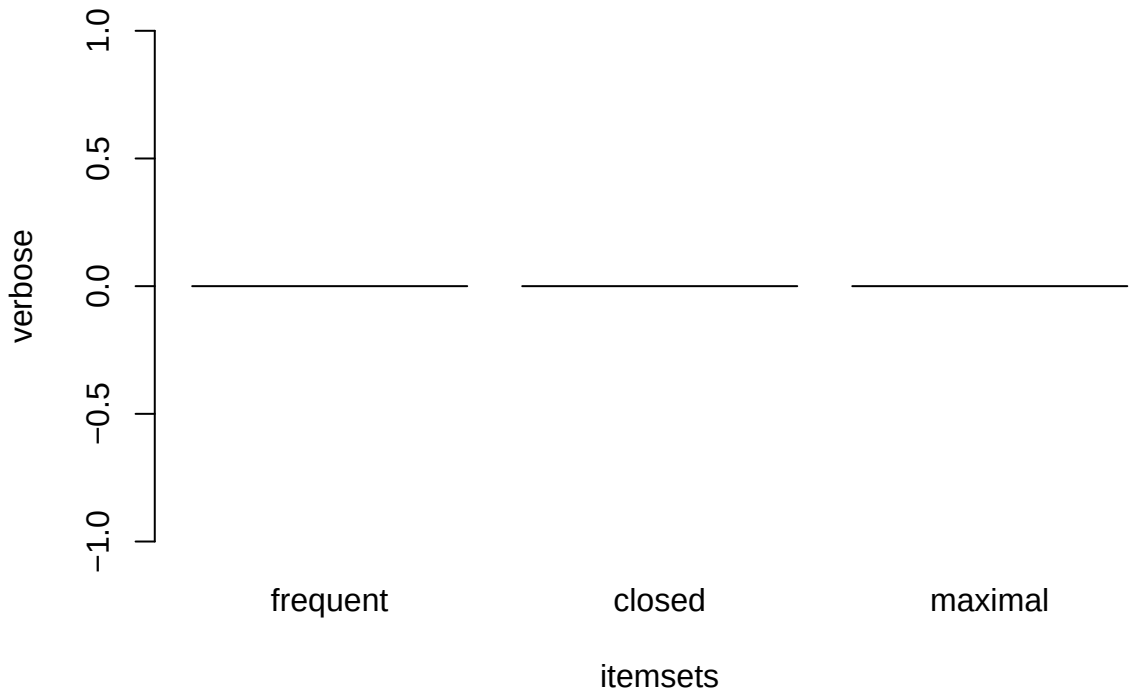
```
is.closed <- is.freq[is.closed(is.freq)]
inspect(head(sort(is.closed, by = "support"), n = 10))
```

Câu 5.

Có bao nhiêu maximal intemset? Bao nhiêu maximal itemset có số item ít nhất? Bao nhiêu maximal itemset có số item là nhiều nhất? Bao nhiêu maximal itemset có ít nhất k item trở lên (thử với k = 2, 3)? Xem top 10 maximal k-itemset sắp xếp theo support (thử với k = 1, 2, 3, …). Vẽ biểu đồ tần số theo bậc (số lượng item) của các maximal itemset.

```
is.max <- is.freq[is.maximal(is.freq)]
inspect(head(sort(is.max, by = "support"), n = 10))
```

```
barplot(c(frequent = length(is.freq),
          closed = length(is.closed),
          maximal = length(is.max)),
        ylab="verbose", xlab="itemsets")
```



Câu 6.

Chọn một cặp giá trị support và confidence mà bạn cho là phù hợp với dữ liệu. Tìm tất cả các rule có tối thiểu 2 item ứng với cặp giá trị support và confidence mà bạn chọn. Cho biết thông tin tổng quan về các rule tìm được: Có bao nhiêu rule? Bậc thấp nhất và cao nhất của các rule là bao nhiêu? Support cao (thấp) nhất của các rule? Cho biết top 10 rule sắp xếp theo độ đo lift.

```
rules <- apriori(browsing)
```

```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
```

```
##         0.8    0.1    1 none FALSE            TRUE       5     0.1      1
##  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: 3110
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[30697 item(s), 31101 transaction(s)] done [0.02s].
## sorting and recoding items ... [0 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 done [0.00s].
## writing ... [0 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```

```
rules
```

```
## set of 0 rules
```

```
90/nrow(browsing)
```

```
## [1] 0.002893798
```

```
rules <- apriori(browsing,
                 parameter = list(support = 0.009,
                                  confidence = 0.25,
                                  minlen = 2))
```

```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##        0.25    0.1    1 none FALSE            TRUE       5   0.009      2
##  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: 279
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[30697 item(s), 31101 transaction(s)] done [0.02s].
## sorting and recoding items ... [0 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 done [0.00s].
## writing ... [0 rule(s)] done [0.00s].
## creating S4 object  ... done [0.01s].
```

```
rules
```

```
## set of 0 rules
```

```r
#summary(rules)
```

```r
rules <- sort(rules, by = "lift")
inspect(head(rules, n = 10))
```

Câu 7.

Cho biết có bao nhiêu rule mà vế trái có ít nhất k items (thử với k = 2, 3, …)? Vẽ biểu đồ các rule dựa trên số bậc. Vẽ biểu đồ dạng graph của top 50 rule theo độ đo lift.

```r
#Vì không có rules nên không thể vẽ biểu đồ được
library(arulesViz)
```

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